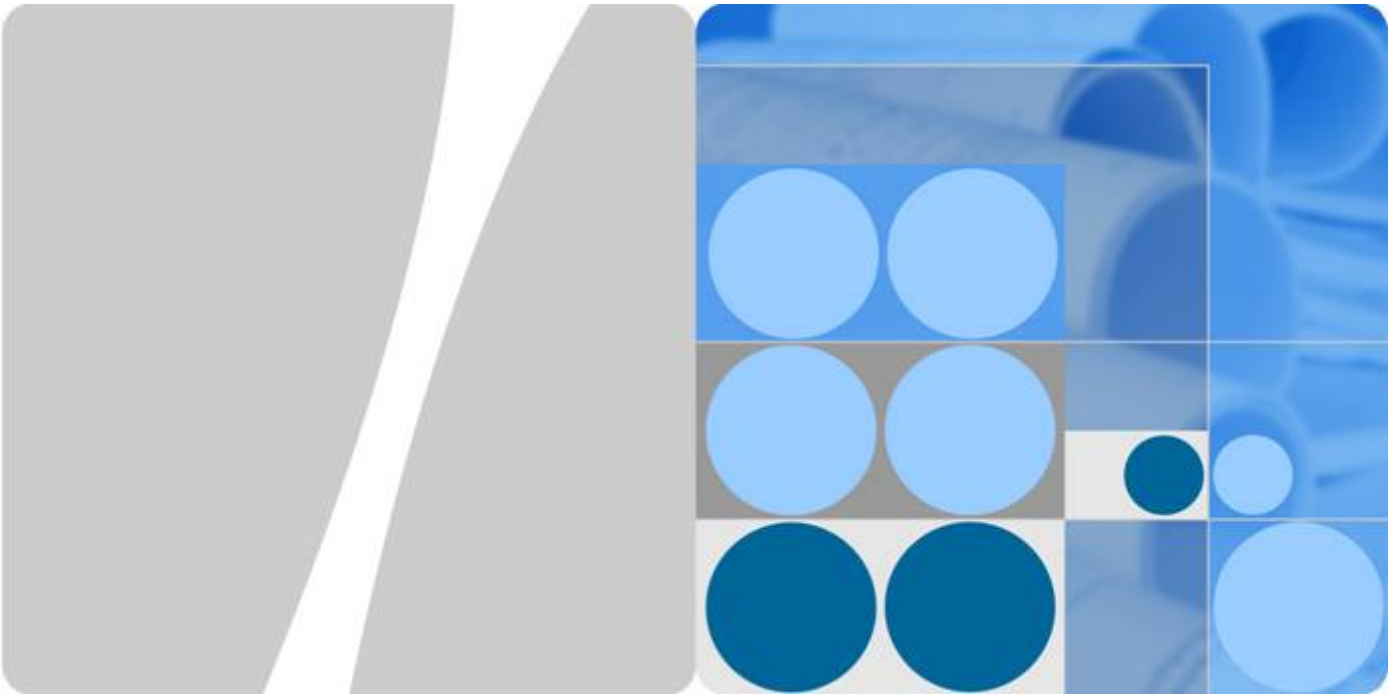


MapReduce 服务

用户指南

发布日期　2024-09-30

目 录

1 简介.....1
1.1 什么是 MRS.....1
1.2 MRS 与自建 Hadoop 对比优势.....3
1.3 应用场景.....7
1.4 组件介绍.....9
1.4.1 Alluxio.....9
1.4.2 CarbonData.....10
1.4.3 ClickHouse.....11
1.4.4 DBService.....15
1.4.4.1 DBService 基本原理.....15
1.4.4.2 DBService 与其他组件的关系.....16
1.4.5 Flink.....16
1.4.5.1 Flink 基本原理.....17
1.4.5.2 Flink HA 方案介绍.....22
1.4.5.3 Flink 与其他组件的关系.....24
1.4.5.4 Flink 开源增强特性.....25
1.4.5.4.1 窗口.....25
1.4.5.4.2 Job Pipeline.....27
1.4.5.4.3 配置表.....32
1.4.5.4.4 Stream SQL Join.....34
1.4.5.4.5 Flink CEP in SQL.....34
1.4.6 Flume.....36
1.4.6.1 Flume 基本原理.....36
1.4.6.2 Flume 与其他组件的关系.....39
1.4.6.3 Flume 开源增强特性.....40
1.4.7 HBase.....40
1.4.7.1 HBase 基本原理.....40
1.4.7.2 HBase HA 方案介绍.....45
1.4.7.3 HBase 与其他组件的关系.....46
1.4.7.4 HBase 开源增强特性.....47
1.4.8 HDFS.....54
1.4.8.1 HDFS 基本原理.....54
1.4.8.2 HDFS HA 方案介绍.....57

1.4.8.3 HDFS 与其他组件的关系 58
1.4.8.4 HDFS 开源增强特性 60
1.4.9 Hive 66
1.4.9.1 Hive 基本原理 66
1.4.9.2 Hive CBO 原理介绍 69
1.4.9.3 Hive 与其他组件的关系 73
1.4.9.4 Hive 开源增强特性 73
1.4.10 Hue 74
1.4.10.1 Hue 基本原理 75
1.4.10.2 Hue 与其他组件的关系 77
1.4.10.3 Hue 开源增强特性 78
1.4.11 Impala 78
1.4.12 Kafka 79
1.4.12.1 Kafka 基本原理 79
1.4.12.2 Kafka 与其他组件的关系 81
1.4.12.3 Kafka 开源增强特性 82
1.4.13 KafkaManager 82
1.4.14 KrbServer 及 LdapServer 83
1.4.14.1 KrbServer 及 LdapServer 基本原理 83
1.4.14.2 KrbServer 及 LdapServer 开源增强特性 86
1.4.15 Kudu 86
1.4.16 Loader 87
1.4.16.1 Loader 基本原理 87
1.4.16.2 Loader 与其他组件的关系 89
1.4.16.3 Loader 开源增强特性 89
1.4.17 Manager 90
1.4.17.1 Manager 基本原理 90
1.4.17.2 Manager 关键特性 93
1.4.18 MapReduce 94
1.4.18.1 MapReduce 基本原理 94
1.4.18.2 MapReduce 与其他组件的关系 95
1.4.18.3 MapReduce 开源增强特性 95
1.4.19 Oozie 98
1.4.19.1 Oozie 基本原理 98
1.4.19.2 Oozie 开源增强特性 100
1.4.20 OpenTSDB 100
1.4.21 Presto 101
1.4.22 Ranger 102
1.4.22.1 Ranger 基本原理 102
1.4.22.2 Ranger 与其他组件的关系 104
1.4.23 Spark 104
1.4.23.1 Spark 基本原理 104

1.4.23.2 Spark HA 方案介绍 119
1.4.23.3 Spark 与 HDFS 和 YARN 的关系 124
1.4.23.4 Spark 开源增强特性：跨源复杂数据的 SQL 查询优化 128
1.4.24 Spark2x 130
1.4.24.1 Spark2x 基本原理 130
1.4.24.2 Spark2x HA 方案介绍 143
1.4.24.2.1 Spark2x 多主实例 143
1.4.24.2.2 Spark2x 多租户 146
1.4.24.3 Spark2x 与组件的关系 149
1.4.24.4 Spark2x 开源新特性 153
1.4.24.5 Spark2x 开源增强特性 153
1.4.24.5.1 CarbonData 简介 153
1.4.24.5.2 跨源复杂数据的 SQL 查询优化 155
1.4.24.5.3 数据倾斜优化 157
1.4.25 Storm 158
1.4.25.1 Storm 基本原理 158
1.4.25.2 Storm 与其他组件的关系 162
1.4.25.3 Storm 开源增强特性 163
1.4.26 Tez 164
1.4.27 YARN 165
1.4.27.1 YARN 基本原理 165
1.4.27.2 YARN HA 方案介绍 169
1.4.27.3 Yarn 与其他组件的关系 170
1.4.27.4 YARN 开源增强特性 173
1.4.28 ZooKeeper 179
1.4.28.1 ZooKeeper 基本原理 179
1.4.28.2 ZooKeeper 与其他组件的关系 181
1.4.28.3 ZooKeeper 开源增强特性 184
1.5 产品功能 187
1.5.1 多租户 187
1.5.2 安全增强 188
1.5.3 组件 WebUI 便捷访问 189
1.5.4 可靠性增强 190
1.5.5 作业管理 191
1.5.6 自定义引导操作 191
1.5.7 企业项目管理 192
1.5.8 元数据 192
1.5.9 集群管理 192
1.5.9.1 集群生命周期管理 192
1.5.9.2 集群扩缩容 194
1.5.9.3 自动弹性伸缩 194
1.5.9.4 创建 Task 节点 195

1.5.9.5 升级 Master 节点规格 196
1.5.9.6 隔离主机 196
1.5.9.7 标签管理 196
1.5.10 集群运维 196
1.5.11 消息通知 197
1.6 约束与限制 198
1.7 技术支持 199
1.8 权限管理 199
1.9 与其他云服务的关系 203
1.10 常见概念 204
2 入门 208
2.1 如何使用 MRS 208
2.2 创建集群 208
2.3 上传示例数据和程序 210
2.4 添加作业 212
2.5 快速使用 Kerberos 认证集群 215
2.6 删除集群 220
3 准备用户 221
3.1 创建 MRS 操作用户 221
3.2 创建 MRS 自定义策略 225
3.3 IAM 用户同步 MRS 说明 230
4 配置集群 235
4.1 创建方式简介 235
4.2 快速创建集群 235
4.2.1 快速创建 Hadoop 分析集群 235
4.2.2 快速创建 HBase 查询集群 236
4.2.3 快速创建 Kafka 流式集群 238
4.2.4 快速创建 ClickHouse 集群 239
4.2.5 快速创建实时分析集群 240
4.3 创建自定义集群 241
4.4 创建自定义拓扑集群 251
4.5 添加集群标签 259
4.6 授权安全通信 261
4.7 配置弹性伸缩规则 263
4.8 管理数据连接 272
4.8.1 配置数据连接 272
4.8.2 配置 Ranger 数据连接 276
4.8.3 配置 Hive 数据连接 280
4.9 通过引导操作安装第三方软件 282
4.9.1 引导操作简介 282
4.9.2 准备引导操作脚本 282

4.9.3 查看执行记录 283
4.9.4 添加引导操作 283
4.10 查看失败的集群操作任务 285
4.11 查看历史集群信息 286
5 管理集群 288
5.1 登录集群 288
5.1.1 MRS 集群节点简介快速创建 Hadoop 分析集群 288
5.1.2 登录集群节点 289
5.1.3 如何确认 Manager 的主备管理节点 293
5.2 集群概览 294
5.2.1 集群列表简介 294
5.2.2 查看集群状态 295
5.2.3 查看集群基本信息 298
5.2.4 查看集群补丁信息 301
5.2.5 查看和定制集群监控指标 302
5.2.6 管理组件和主机监控 304
5.3 集群运维 308
5.3.1 导入导出数据 308
5.3.2 切换集群子网 311
5.3.3 配置消息通知 313
5.3.4 健康检查 315
5.3.4.1 使用前须知 315
5.3.4.2 执行健康检查 315
5.3.4.3 查看并导出检查报告 316
5.3.5 远程运维 317
5.3.5.1 运维授权 317
5.3.5.2 日志共享 317
5.3.6 查看 MRS 服务操作日志 318
5.3.7 删除集群 319
5.4 节点管理 319
5.4.1 扩容集群 319
5.4.2 缩容集群 322
5.4.3 管理主机（节点）操作 324
5.4.4 隔离主机 324
5.4.5 取消隔离主机 325
5.4.6 升级 Master 节点规格 326
5.5 作业管理 326
5.5.1 MRS 作业简介 326
5.5.2 运行 MapReduce 作业 331
5.5.3 运行 SparkSubmit 作业 334
5.5.4 运行 HiveSql 作业 337
5.5.5 运行 SparkSql 作业 340

5.5.6 运行 Flink 作业 344
5.5.7 运行 Kafka 作业 349
5.5.8 查看作业配置信息和日志 350
5.5.9 停止作业 351
5.5.10 删除作业 351
5.5.11 使用 OBS 加密数据运行作业 352
5.5.12 配置作业消息通知 358
5.6 组件管理 359
5.6.1 对象管理简介 359
5.6.2 查看配置 359
5.6.3 管理服务操作 360
5.6.4 配置服务参数 361
5.6.5 配置服务自定义参数 362
5.6.6 同步服务配置 363
5.6.7 管理角色实例操作 364
5.6.8 配置角色实例参数 364
5.6.9 同步角色实例配置 365
5.6.10 退服和入服角色实例 365
5.6.11 启动及停止集群 366
5.6.12 同步集群配置 367
5.6.13 导出集群的配置数据 367
5.6.14 支持滚动重启 368
5.7 告警管理 371
5.7.1 查看告警列表 371
5.7.2 查看事件列表 373
5.7.3 查看与手动清除告警 376
5.8 补丁管理 377
5.8.1 MRS 3.x 之前版本补丁操作指导 377
5.8.2 滚动补丁 378
5.8.3 修复隔离主机补丁 380
5.9 租户管理 381
5.9.1 使用前须知 381
5.9.2 租户简介 381
5.9.3 添加租户 382
5.9.4 添加子租户 384
5.9.5 删除租户 386
5.9.6 管理租户目录 387
5.9.7 恢复租户数据 389
5.9.8 添加资源池 389
5.9.9 修改资源池 390
5.9.10 删除资源池 391
5.9.11 配置队列 391

5.9.12 配置资源池的队列容量策略 393
5.9.13 清除队列配置 394
6 使用 MRS 客户端 396
6.1 安装客户端 396
6.1.1 安装客户端（3.x 及之后版本） 396
6.1.2 安装客户端（3.x 之前版本） 400
6.2 更新客户端 405
6.2.1 更新客户端（3.x 及之后版本） 405
6.2.2 更新客户端（3.x 之前版本） 406
6.3 各组件客户端使用实践 409
6.3.1 使用 ClickHouse 客户端 410
6.3.2 使用 Flink 客户端 412
6.3.3 使用 Flume 客户端 419
6.3.4 使用 HBase 客户端 425
6.3.5 使用 HDFS 客户端 426
6.3.6 使用 Hive 客户端 428
6.3.7 使用 Impala 客户端 432
6.3.8 使用 Kafka 客户端 434
6.3.9 使用 Kudu 客户端 436
6.3.10 使用 Oozie 客户端 437
6.3.11 使用 Storm 客户端 438
6.3.12 使用 Yarn 客户端 439
7 配置存算分离 441
7.1 存算分离简介 441
7.2 配置存算分离集群（委托方式） 441
7.3 配置存算分离集群（AKSK 方式） 448
7.4 使用存算分离集群 451
7.4.1 Flink 对接 OBS 文件系统 451
7.4.2 Flume 对接 OBS 文件系统 452
7.4.3 HDFS 客户端对接 OBS 文件系统 453
7.4.4 Hive 对接 OBS 文件系统 454
7.4.5 MapReduce 对接 OBS 文件系统 457
7.4.6 Spark2x 对接 OBS 文件系统 457
7.4.7 Sqoop 对接外部存储系统 460
8 访问 MRS 集群上托管的开源组件 Web 页面 464
8.1 开源组件 Web 站点 464
8.2 开源组件端口列表 467
8.3 通过专线访问 479
8.4 通过弹性公网 IP 访问 480
8.5 通过 Windows 弹性云服务器访问 481
8.6 创建连接 MRS 集群的 SSH 隧道并配置浏览器 483

9 访问集群 Manager 486
9.1 访问 FusionInsight Manager（MRS 3.x 及之后版本） 486
9.2 访问 MRS Manager（MRS 2.x 及之前版本） 488
10 FusionInsight Manager 操作指导（适用于 3.x） 492
10.1 从这里开始 492
10.1.1 FusionInsight Manager 入门指导 492
10.1.2 查询 FusionInsight Manager 版本号 493
10.1.3 登录管理系统 494
10.1.4 登录管理节点 494
10.2 主页 495
10.2.1 主页概述 495
10.2.2 管理监控指标数据报表 496
10.3 集群 498
10.3.1 管理集群 498
10.3.1.1 集群管理概述 498
10.3.1.2 滚动重启集群 499
10.3.1.3 管理配置过期 501
10.3.1.4 下载客户端 501
10.3.1.5 修改集群属性 502
10.3.1.6 管理集群配置 503
10.3.1.7 静态服务池 504
10.3.1.7.1 静态服务资源 504
10.3.1.7.2 配置集群静态资源 505
10.3.1.7.3 查看集群静态资源 506
10.3.1.8 客户端管理 507
10.3.1.8.1 管理客户端 507
10.3.1.8.2 批量升级客户端 508
10.3.1.8.3 批量刷新 hosts 文件 510
10.3.2 管理服务 510
10.3.2.1 服务管理概述 510
10.3.2.2 其他服务管理操作 514
10.3.2.2.1 服务详情概述 514
10.3.2.2.2 执行角色实例主备倒换 516
10.3.2.2.3 资源监控 517
10.3.2.2.4 采集堆栈信息 519
10.3.2.2.5 切换 Ranger 鉴权 520
10.3.2.3 服务配置 521
10.3.2.3.1 修改服务配置参数 521
10.3.2.3.2 修改服务自定义配置参数 523
10.3.3 管理实例 524
10.3.3.1 实例管理概述 524
10.3.3.2 入服与退服实例 526

10.3.3.3 管理实例配置 527
10.3.3.4 查看实例配置文件 528
10.3.3.5 实例组 529
10.3.3.5.1 管理实例组 529
10.3.3.5.2 查看实例组信息 531
10.3.3.5.3 配置实例组参数 531
10.4 主机 532
10.4.1 主机管理页面 532
10.4.1.1 查看主机列表 532
10.4.1.2 查看主机概览 533
10.4.1.3 查看主机进程及资源 533
10.4.2 主机维护操作 534
10.4.2.1 启动、停止主机上的所有实例 534
10.4.2.2 执行主机健康检查 534
10.4.2.3 分配机架 535
10.4.2.4 隔离主机 537
10.4.2.5 导出主机信息 538
10.4.3 资源概况 538
10.4.3.1 分布 538
10.4.3.2 趋势 541
10.4.3.3 集群 542
10.4.3.4 主机 542
10.5 运维 543
10.5.1 告警 543
10.5.1.1 告警与事件概述 543
10.5.1.2 配置阈值 546
10.5.1.3 配置告警屏蔽状态 557
10.5.2 日志 558
10.5.2.1 在线检索日志 559
10.5.2.2 下载日志 560
10.5.3 健康检查 561
10.5.3.1 查看健康检查任务 561
10.5.3.2 管理健康检查报告 562
10.5.3.3 修改健康检查配置 562
10.5.4 备份恢复设置 563
10.5.4.1 创建备份任务 563
10.5.4.2 创建恢复任务 564
10.5.4.3 其他任务管理说明 564
10.6 审计 565
10.6.1 审计管理页面概述 565
10.6.2 配置审计日志转储 566
10.7 租户资源 567

10.7.1 多租户介绍 567
10.7.1.1 简介 567
10.7.1.2 技术原理 568
10.7.1.2.1 多租户管理页面概述 568
10.7.1.2.2 相关模型 571
10.7.1.2.3 资源概述 574
10.7.1.2.4 动态资源 574
10.7.1.2.5 存储资源 576
10.7.1.3 多租户使用 577
10.7.1.3.1 使用说明 577
10.7.1.3.2 流程概述 578
10.7.2 使用 Superior 调度器的租户业务 579
10.7.2.1 创建租户 579
10.7.2.1.1 添加租户 579
10.7.2.1.2 添加子租户 582
10.7.2.1.3 添加用户并绑定租户的角色 585
10.7.2.2 管理租户 586
10.7.2.2.1 管理租户目录 586
10.7.2.2.2 恢复租户数据 588
10.7.2.2.3 删除租户 589
10.7.2.3 管理资源 589
10.7.2.3.1 添加资源池 590
10.7.2.3.2 修改资源池 590
10.7.2.3.3 删除资源池 591
10.7.2.3.4 配置队列 591
10.7.2.3.5 配置资源池的队列容量策略 593
10.7.2.3.6 清除队列容量配置 594
10.7.2.4 管理全局用户策略 594
10.7.3 使用 Capacity 调度器的租户业务 595
10.7.3.1 创建租户 595
10.7.3.1.1 添加租户 595
10.7.3.1.2 添加子租户 598
10.7.3.1.3 添加用户并绑定租户的角色 601
10.7.3.2 管理租户 602
10.7.3.2.1 管理租户目录 602
10.7.3.2.2 恢复租户数据 604
10.7.3.2.3 删除租户 605
10.7.3.2.4 Capacity Scheduler 模式下清除租户非关联队列 605
10.7.3.3 管理资源 606
10.7.3.3.1 添加资源池 606
10.7.3.3.2 修改资源池 607
10.7.3.3.3 删除资源池 608

10.7.3.3.4 配置队列 608
10.7.3.3.5 配置资源池的队列容量策略 609
10.7.3.3.6 清除队列容量配置 610
10.7.4 切换调度器 611
10.8 系统设置 612
10.8.1 权限设置 613
10.8.1.1 用户管理 613
10.8.1.1.1 创建用户 613
10.8.1.1.2 修改用户信息 614
10.8.1.1.3 导出用户信息 614
10.8.1.1.4 锁定用户 615
10.8.1.1.5 解锁用户 615
10.8.1.1.6 删除用户 616
10.8.1.1.7 修改用户密码 616
10.8.1.1.8 初始化用户密码 618
10.8.1.1.9 导出认证凭据文件 618
10.8.1.2 用户组管理 619
10.8.1.3 角色管理 620
10.8.1.4 安全策略 622
10.8.1.4.1 配置密码策略 622
10.8.1.4.2 配置私有属性 623
10.8.2 对接设置 624
10.8.2.1 配置 SNMP 北向参数 624
10.8.2.2 配置 Syslog 北向参数 626
10.8.2.3 配置监控指标数据转储 630
10.8.3 导入证书 632
10.8.4 OMS 管理 633
10.8.4.1 OMS 维护页面概述 633
10.8.4.2 修改 OMS 服务配置参数 634
10.8.5 部件管理 636
10.8.5.1 查看部件包 636
10.9 集群管理 636
10.9.1 配置客户端 636
10.9.1.1 安装客户端 636
10.9.1.2 使用客户端 641
10.9.1.3 更新已安装客户端的配置 641
10.9.2 集群互信管理 643
10.9.2.1 集群互信概述 643
10.9.2.2 修改 Manager 系统域名 643
10.9.2.3 配置跨 Manager 集群互信 646
10.9.2.4 配置跨集群互信后的用户权限 648
10.9.3 配置定时备份告警与审计信息 649

10.9.4 修改 FusionInsight Manager 添加的路由表 650
10.9.5 切换维护模式 652
10.9.6 例行维护 654
10.10 日志管理 656
10.10.1 关于日志 656
10.10.2 Manager 日志清单 671
10.10.3 配置日志级别与文件大小 679
10.10.4 配置审计日志本地备份数 681
10.10.5 查看角色实例日志 682
10.11 备份恢复管理 683
10.11.1 备份恢复简介 683
10.11.2 备份数据 687
10.11.2.1 备份 OMS 数据 688
10.11.2.2 备份 DBService 数据 691
10.11.2.3 备份 HBase 元数据 694
10.11.2.4 备份 HBase 业务数据 697
10.11.2.5 备份 NameNode 数据 701
10.11.2.6 备份 HDFS 业务数据 704
10.11.2.7 备份 Hive 业务数据 708
10.11.2.8 备份 Kafka 元数据 712
10.11.3 恢复数据 714
10.11.3.1 恢复 OMS 数据 715
10.11.3.2 恢复 DBService 数据 718
10.11.3.3 恢复 HBase 元数据 721
10.11.3.4 恢复 HBase 业务数据 724
10.11.3.5 恢复 NameNode 数据 727
10.11.3.6 恢复 HDFS 业务数据 731
10.11.3.7 恢复 Hive 业务数据 734
10.11.3.8 恢复 Kafka 元数据 738
10.11.4 启用集群间拷贝功能 740
10.11.5 管理本地快速恢复任务 741
10.11.6 修改备份任务 742
10.11.7 查看备份恢复任务 743
10.12 安全管理 744
10.12.1 安全概述 744
10.12.1.1 权限模型 744
10.12.1.2 权限机制 745
10.12.1.3 认证策略 746
10.12.1.4 鉴权策略 748
10.12.1.5 用户帐号一览表 750
10.12.1.6 默认权限信息一览 772
10.12.1.7 FusionInsight Manager 安全功能 775

10.12.2 帐户管理 775
10.12.2.1 帐户安全设置 775
10.12.2.1.1 解锁 LDAP 用户和管理帐户 775
10.12.2.1.2 解锁系统内部用户 776
10.12.2.1.3 修改集群组件鉴权配置开关 777
10.12.2.1.4 使用普通模式集群用户在非集群节点登录 779
10.12.2.2 修改系统用户密码 781
10.12.2.2.1 修改 admin 密码 781
10.12.2.2.2 修改操作系统用户密码 781
10.12.2.3 修改系统内部用户密码 782
10.12.2.3.1 修改 Kerberos 管理员密码 782
10.12.2.3.2 修改 OMS Kerberos 管理员密码 783
10.12.2.3.3 修改 LDAP 管理员和 LDAP 用户密码（含 OMS LDAP） 783
10.12.2.3.4 修改 LDAP 管理帐户密码 785
10.12.2.3.5 修改组件运行用户密码 786
10.12.2.4 修改默认数据库用户密码 787
10.12.2.4.1 修改 OMS 数据库管理员密码 787
10.12.2.4.2 修改 OMS 数据库访问用户密码 788
10.12.2.4.3 修改组件数据库用户密码 789
10.12.2.4.4 修改 DBService 数据库 omm 用户密码 789
10.12.3 安全加固 790
10.12.3.1 加固策略 790
10.12.3.2 配置受信任 IP 访问 LDAP 791
10.12.3.3 加密 HFile 和 WAL 内容 794
10.12.3.4 安全配置 798
10.12.3.5 配置 HBase 允许修改操作的 IP 地址白名单 800
10.12.3.6 更新集群密钥 801
10.12.3.7 加固 LDAP 802
10.12.3.8 配置 Kafka 数据传输加密 803
10.12.3.9 配置 HDFS 数据传输加密 804
10.12.3.10 配置 Controller 与 Agent 间通信加密 805
10.12.3.11 更新 omm 用户 ssh 密钥 806
10.12.4 安全维护 807
10.12.4.1 帐户维护建议 808
10.12.4.2 密码维护建议 808
10.12.4.3 日志维护建议 808
10.12.5 安全声明 808
10.13 告警参考（适用于 MRS 3.x 版本） 809
10.13.1 ALM-12001 审计日志转储失败 809
10.13.2 ALM-12004 OLdap 资源异常 811
10.13.3 ALM-12005 OKerberos 资源异常 813
10.13.4 ALM-12006 节点故障 814

10.13.5 ALM-12007 进程故障 817
10.13.6 ALM-12010 Manager 主备节点间心跳中断 819
10.13.7 ALM-12011 Manager 主备节点同步数据异常 821
10.13.8 ALM-12014 设备分区丢失 824
10.13.9 ALM-12015 设备分区文件系统只读 826
10.13.10 ALM-12016 CPU 使用率超过阈值 827
10.13.11 ALM-12017 磁盘容量不足 830
10.13.12 ALM-12018 内存使用率超过阈值 833
10.13.13 ALM-12027 主机 PID 使用率超过阈值 835
10.13.14 ALM-12028 主机 D 状态进程数超过阈值 836
10.13.15 ALM-12033 慢盘故障 838
10.13.16 ALM-12034 周期备份任务失败 843
10.13.17 ALM-12035 恢复任务失败后数据状态未知 845
10.13.18 ALM-12038 监控指标转储失败 846
10.13.19 ALM-12039 OMS 数据库主备不同步 849
10.13.20 ALM-12040 系统熵值不足 851
10.13.21 ALM-12041 关键文件权限异常 853
10.13.22 ALM-12042 关键文件配置异常 855
10.13.23 ALM-12045 网络读包丢包率超过阈值 857
10.13.24 ALM-12046 网络写包丢包率超过阈值 862
10.13.25 ALM-12047 网络读包错误率超过阈值 864
10.13.26 ALM-12048 网络写包错误率超过阈值 867
10.13.27 ALM-12049 网络读吞吐率超过阈值 869
10.13.28 ALM-12050 网络写吞吐率超过阈值 872
10.13.29 ALM-12051 磁盘 Inode 使用率超过阈值 874
10.13.30 ALM-12052 TCP 临时端口使用率超过阈值 876
10.13.31 ALM-12053 主机文件句柄使用率超过阈值 879
10.13.32 ALM-12054 证书文件失效 881
10.13.33 ALM-12055 证书文件即将过期 883
10.13.34 ALM-12057 元数据未配置周期备份到第三方服务器的任务 886
10.13.35 ALM-12061 进程使用率超过阈值 887
10.13.36 ALM-12062 OMS 参数配置同集群规模不匹配 890
10.13.37 ALM-12063 磁盘不可用 892
10.13.38 ALM-12064 主机随机端口范围配置与集群使用端口冲突 894
10.13.39 ALM-12066 节点间互信失效 896
10.13.40 ALM-12067 tomcat 资源异常 899
10.13.41 ALM-12068 acs 资源异常 900
10.13.42 ALM-12069 aos 资源异常 902
10.13.43 ALM-12070 controller 资源异常 903
10.13.44 ALM-12071 httpd 资源异常 905
10.13.45 ALM-12072 floatip 资源异常 907
10.13.46 ALM-12073 cep 资源异常 908

10.13.47 ALM-12074 fms 资源异常 910
10.13.48 ALM-12075 pms 资源异常 912
10.13.49 ALM-12076 gaussDB 资源异常 913
10.13.50 ALM-12077 omm 用户过期 916
10.13.51 ALM-12078 omm 密码过期 917
10.13.52 ALM-12079 omm 用户即将过期 919
10.13.53 ALM-12080 omm 密码即将过期 920
10.13.54 ALM-12081 ommdba 用户过期 922
10.13.55 ALM-12082 ommdba 用户即将过期 923
10.13.56 ALM-12083 ommdba 密码即将过期 925
10.13.57 ALM-12084 ommdba 密码过期 926
10.13.58 ALM-12085 服务审计日志转储失败 928
10.13.59 ALM-12087 系统处于升级观察期 930
10.13.60 ALM-12089 节点间网络互通异常 931
10.13.61 ALM-12101 AZ 不健康 933
10.13.62 ALM-12102 AZ 高可用组件未按容灾需求部署 935
10.13.63 ALM-12110 获取 ECS 临时 ak/sk 失败 936
10.13.64 ALM-13000 ZooKeeper 服务不可用 937
10.13.65 ALM-13001 ZooKeeper 可用连接数不足 940
10.13.66 ALM-13002 ZooKeeper 直接内存使用率超过阈值 943
10.13.67 ALM-13003 ZooKeeper 进程垃圾回收（GC）时间超过阈值 945
10.13.68 ALM-13004 ZooKeeper 堆内存使用率超过阈值 946
10.13.69 ALM-13005 ZooKeeper 中组件顶层目录的配额设置失败 948
10.13.70 ALM-13006 Znode 数量或容量超过阈值 950
10.13.71 ALM-13007 ZooKeeper 客户端可用连接数不足 952
10.13.72 ALM-13008 ZooKeeper Znode 数量使用率超出阈值 953
10.13.73 ALM-13009 ZooKeeper Znode 容量使用率超出阈值 955
10.13.74 ALM-13010 配置 quota 的目录 Znode 使用率超出阈值 957
10.13.75 ALM-14000 HDFS 服务不可用 958
10.13.76 ALM-14001 HDFS 磁盘空间使用率超过阈值 960
10.13.77 ALM-14002 DataNode 磁盘空间使用率超过阈值 962
10.13.78 ALM-14003 丢失的 HDFS 块数量超过阈值 965
10.13.79 ALM-14006 HDFS 文件数超过阈值 967
10.13.80 ALM-14007 NameNode 堆内存使用率超过阈值 970
10.13.81 ALM-14008 DataNode 堆内存使用率超过阈值 972
10.13.82 ALM-14009 Dead DataNode 数量超过阈值 975
10.13.83 ALM-14010 NameService 服务异常 978
10.13.84 ALM-14011 DataNode 数据目录配置不合理 981
10.13.85 ALM-14012 Journalnode 数据不同步 984
10.13.86 ALM-14013 NameNode FsImage 文件更新失败 986
10.13.87 ALM-14014 NameNode 进程垃圾回收（GC）时间超过阈值 990
10.13.88 ALM-14015 DataNode 进程垃圾回收（GC）时间超过阈值 992

10.13.89 ALM-14016 DataNode 直接内存使用率超过阈值 994
10.13.90 ALM-14017 NameNode 直接内存使用率超过阈值 995
10.13.91 ALM-14018 NameNode 非堆内存使用率超过阈值 997
10.13.92 ALM-14019 DataNode 非堆内存使用率超过阈值 1000
10.13.93 ALM-14020 HDFS 目录条目数量超过阈值 1002
10.13.94 ALM-14021 NameNode RPC 处理平均时间超过阈值 1004
10.13.95 ALM-14022 NameNode RPC 队列平均时间超过阈值 1007
10.13.96 ALM-14023 总副本预留磁盘空间所占比率超过阈值 1010
10.13.97 ALM-14024 租户空间使用率超过阈值 1013
10.13.98 ALM-14025 租户文件对象使用率超过阈值 1015
10.13.99 ALM-14026 DataNode 块数超过阈值 1017
10.13.100 ALM-14027 DataNode 磁盘故障 1019
10.13.101 ALM-14028 待补齐的块数超过阈值 1021
10.13.102 ALM-14029 单副本的块数超过阈值 1024
10.13.103 ALM-16000 连接到 HiveServer 的 session 数占最大允许数的百分比超过阈值 1026
10.13.104 ALM-16001 Hive 数据仓库空间使用率超过阈值 1027
10.13.105 ALM-16002 Hive SQL 执行成功率低于阈值 1029
10.13.106 ALM-16003 Background 线程使用率超过阈值 1032
10.13.107 ALM-16004 Hive 服务不可用 1034
10.13.108 ALM-16005 Hive 服务进程堆内存使用超出阈值 1037
10.13.109 ALM-16006 Hive 服务进程直接内存使用超出阈值 1039
10.13.110 ALM-16007 Hive GC 时间超出阈值 1042
10.13.111 ALM-16008 Hive 服务进程非堆内存使用超出阈值 1044
10.13.112 ALM-16009 Map 数超过阈值 1046
10.13.113 ALM-16045 Hive 数据仓库被删除 1047
10.13.114 ALM-16046 Hive 数据仓库权限被修改 1049
10.13.115 ALM-16047 HiveServer 已从 Zookeeper 注销 1050
10.13.116 ALM-16048 Tez 或者 Spark 库路径不存在 1051
10.13.117 ALM-17003 Oozie 服务不可用 1053
10.13.118 ALM-17004 Oozie 堆内存使用率超过阈值 1056
10.13.119 ALM-17005 Oozie 非堆内存使用率超过阈值 1058
10.13.120 ALM-17006 Oozie 直接内存使用率超过阈值 1060
10.13.121 ALM-17007 Oozie 进程垃圾回收（GC）时间超过阈值 1061
10.13.122 ALM-18000 Yarn 服务不可用 1063
10.13.123 ALM-18002 NodeManager 心跳丢失 1065
10.13.124 ALM-18003 NodeManager 不健康 1068
10.13.125 ALM-18008 ResourceManager 堆内存使用率超过阈值 1070
10.13.126 ALM-18009 JobHistoryServer 堆内存使用率超过阈值 1072
10.13.127 ALM-18010 ResourceManager 进程垃圾回收（GC）时间超过阈值 1074
10.13.128 ALM-18011 NodeManager 进程垃圾回收（GC）时间超过阈值 1076
10.13.129 ALM-18012 JobHistoryServer 进程垃圾回收（GC）时间超过阈值 1078
10.13.130 ALM-18013 ResourceManager 直接内存使用率超过阈值 1080

10.13.131 ALM-18014 NodeManager 直接内存使用率超过阈值 1082
10.13.132 ALM-18015 JobHistoryServer 直接内存使用率超过阈值 1084
10.13.133 ALM-18016 ResourceManager 非堆内存使用率超过阈值 1085
10.13.134 ALM-18017 NodeManager 非堆内存使用率超过阈值 1088
10.13.135 ALM-18018 NodeManager 堆内存使用率超过阈值 1090
10.13.136 ALM-18019 JobHistoryServer 非堆内存使用率超过阈值 1091
10.13.137 ALM-18020 Yarn 任务执行超时 1093
10.13.138 ALM-18021 Mapreduce 服务不可用 1096
10.13.139 ALM-18022 Yarn 队列资源不足 1098
10.13.140 ALM-18023 Yarn 任务挂起数超过阈值 1101
10.13.141 ALM-18024 Yarn 任务挂起内存量超阈值 1102
10.13.142 ALM-18025 Yarn 被终止的任务数超过阈值 1104
10.13.143 ALM-18026 Yarn 上运行失败的任务数超过阈值 1106
10.13.144 ALM-19000 HBase 服务不可用 1107
10.13.145 ALM-19006 HBase 容灾同步失败 1112
10.13.146 ALM-19007 HBase GC 时间超出阈值 1115
10.13.147 ALM-19008 HBase 服务进程堆内存使用率超出阈值 1117
10.13.148 ALM-19009 HBase 服务进程直接内存使用率超出阈值 1120
10.13.149 ALM-19011 RegionServer 的 Region 数量超出阈值 1122
10.13.150 ALM-19012 HBase 系统表目录或文件丢失 1125
10.13.151 ALM-19013 region 处在 RIT 状态的时长超过阈值 1127
10.13.152 ALM-19014 在 ZooKeeper 上的容量配额使用率严重超过阈值 1129
10.13.153 ALM-19015 在 ZooKeeper 上的数量配额使用率超过阈值 1132
10.13.154 ALM-19016 在 ZooKeeper 上的数量配额使用率严重超过阈值 1134
10.13.155 ALM-19017 在 ZooKeeper 上的容量配额使用率超过阈值 1136
10.13.156 ALM-19018 HBase 合并队列超出阈值 1138
10.13.157 ALM-19019 HBase 容灾等待同步的 HFile 文件数量超过阈值 1140
10.13.158 ALM-19020 HBase 容灾等待同步的 wal 文件数量超过阈值 1143
10.13.159 ALM-20002 Hue 服务不可用 1145
10.13.160 ALM-24000 Flume 服务不可用 1148
10.13.161 ALM-24001 Flume Agent 异常 1149
10.13.162 ALM-24003 Flume Client 连接中断 1152
10.13.163 ALM-24004 Flume 读取数据异常 1154
10.13.164 ALM-24005 Flume 传输数据异常 1156
10.13.165 ALM-24006 Flume Server 堆内存使用率超过阈值 1159
10.13.166 ALM-24007 Flume Server 直接内存使用率超过阈值 1161
10.13.167 ALM-24008 Flume Server 非堆内存使用率超过阈值 1162
10.13.168 ALM-24009 Flume Server 垃圾回收(GC)时间超过阈值 1164
10.13.169 ALM-24010 Flume 证书文件非法或已损坏 1166
10.13.170 ALM-24011 Flume 证书文件即将过期 1168
10.13.171 ALM-24012 Flume 证书文件已过期 1170
10.13.172 ALM-24013 Flume MonitorServer 证书文件非法或已损坏 1172

10.13.173 ALM-24014 Flume MonitorServer 证书文件即将过期.....1174
10.13.174 ALM-24015 Flume MonitorServer 证书文件已过期.....1176
10.13.175 ALM-25000 LdapServer 服务不可用.....1178
10.13.176 ALM-25004 LdapServer 数据同步异常.....1180
10.13.177 ALM-25005 Nscd 服务异常.....1182
10.13.178 ALM-25006 Sssd 服务异常.....1185
10.13.179 ALM-25500 KrbServer 服务不可用.....1188
10.13.180 ALM-26051 Storm 服务不可用.....1190
10.13.181 ALM-26052 Storm 服务可用 Supervisor 数量小于阈值.....1192
10.13.182 ALM-26053 Storm Slot 使用率超过阈值.....1194
10.13.183 ALM-26054 Nimbus 堆内存使用率超过阈值.....1196
10.13.184 ALM-27001 DBService 服务不可用.....1198
10.13.185 ALM-27003 DBService 主备节点间心跳中断.....1200
10.13.186 ALM-27004 DBService 主备数据不同步.....1202
10.13.187 ALM-27005 数据库连接数使用率超过阈值.....1204
10.13.188 ALM-27006 数据目录磁盘空间使用率超过阈值.....1208
10.13.189 ALM-27007 数据库进入只读模式.....1210
10.13.190 ALM-29000 Impala 服务不可用.....1213
10.13.191 ALM-29004 Impalad 进程内存占用率超过阈值.....1214
10.13.192 ALM-29005 Impalad JDBC 连接数超过阈值.....1216
10.13.193 ALM-29006 Impalad ODBC 连接数超过阈值.....1218
10.13.194 ALM-29100 Kudu 服务不可用.....1219
10.13.195 ALM-29104 Tserver 进程内存占用率超过阈值.....1221
10.13.196 ALM-29106 Tserver 进程 CPU 占用率过高.....1222
10.13.197 ALM-29107 Tserver 进程内存使用百分比超过阈值.....1224
10.13.198 ALM-38000 Kafka 服务不可用.....1225
10.13.199 ALM-38001 Kafka 磁盘容量不足.....1227
10.13.200 ALM-38002 Kafka 堆内存使用率超过阈值.....1232
10.13.201 ALM-38004 Kafka 直接内存使用率超过阈值.....1233
10.13.202 ALM-38005 Broker 进程垃圾回收（GC）时间超过阈值.....1235
10.13.203 ALM-38006 Kafka 未完全同步的 Partition 百分比超过阈值.....1237
10.13.204 ALM-38007 Kafka 默认用户状态异常.....1239
10.13.205 ALM-38008 Kafka 数据目录状态异常.....1241
10.13.206 ALM-38009 Broker 磁盘 IO 繁忙.....1242
10.13.207 ALM-38010 存在单副本的 Topic.....1245
10.13.208 ALM-43001 Spark2x 服务不可用.....1246
10.13.209 ALM-43006 JobHistory2x 进程堆内存使用超出阈值.....1248
10.13.210 ALM-43007 JobHistory2x 进程非堆内存使用超出阈值.....1250
10.13.211 ALM-43008 JobHistory2x 进程直接内存使用超出阈值.....1252
10.13.212 ALM-43009 JobHistory2x 进程 GC 时间超出阈值.....1254
10.13.213 ALM-43010 JDBCServer2x 进程堆内存使用超出阈值.....1256
10.13.214 ALM-43011 JDBCServer2x 进程非堆内存使用超出阈值.....1258

10.13.215 ALM-43012 JDBCServer2x 进程直接内存使用超出阈值 1259
10.13.216 ALM-43013 JDBCServer2x 进程 GC 时间超出阈值 1261
10.13.217 ALM-43017 JDBCServer2x 进程 Full GC 次数超出阈值 1263
10.13.218 ALM-43018 JobHistory2x 进程 Full GC 次数超出阈值 1265
10.13.219 ALM-43019 IndexServer2x 进程堆内存使用超出阈值 1266
10.13.220 ALM-43020 IndexServer2x 进程非堆内存使用超出阈值 1268
10.13.221 ALM-43021 IndexServer2x 进程直接内存使用超出阈值 1270
10.13.222 ALM-43022 IndexServer2x 进程 GC 时间超出阈值 1272
10.13.223 ALM-43023 IndexServer2x 进程 Full GC 次数超出阈值 1274
10.13.224 ALM-44004 Presto Coordinator 资源组排队任务超过阈值 1275
10.13.225 ALM-44005 Presto Coordinator 进程垃圾收集时间超出阈值 1277
10.13.226 ALM-44006 Presto Worker 进程垃圾收集时间超出阈值 1278
10.13.227 ALM-45175 OBS 元数据接口调用平均时间超过阈值 1279
10.13.228 ALM-45176 OBS 元数据接口调用成功率低于阈值 1281
10.13.229 ALM-45177 OBS 数据读操作接口调用成功率低于阈值 1283
10.13.230 ALM-45178 OBS 数据写操作接口调用成功率低于阈值 1284
10.13.231 ALM-45275 Ranger 服务不可用 1286
10.13.232 ALM-45276 RangerAdmin 状态异常 1287
10.13.233 ALM-45277 RangerAdmin 堆内存使用率超过阈值 1289
10.13.234 ALM-45278 RangerAdmin 直接内存使用率超过阈值 1291
10.13.235 ALM-45279 RangerAdmin 非堆内存使用率超过阈值 1292
10.13.236 ALM-45280 RangerAdmin 垃圾回收(GC)时间超过阈值 1294
10.13.237 ALM-45281 UserSync 堆内存使用率超过阈值 1296
10.13.238 ALM-45282 UserSync 直接内存使用率超过阈值 1297
10.13.239 ALM-45283 UserSync 非堆内存使用率超过阈值 1299
10.13.240 ALM-45284 UserSync 垃圾回收(GC)时间超过阈值 1301
10.13.241 ALM-45285 TagSync 堆内存使用率超过阈值 1302
10.13.242 ALM-45286 TagSync 直接内存使用率超过阈值 1304
10.13.243 ALM-45287 TagSync 非堆内存使用率超过阈值 1306
10.13.244 ALM-45288 TagSync 垃圾回收(GC)时间超过阈值 1307
10.13.245 ALM-45425 ClickHouse 服务不可用 1309
10.13.246 ALM-45426 ClickHouse 服务在 ZooKeeper 的数量配额使用率超过阈值 1311
10.13.247 ALM-45427 ClickHouse 服务在 ZooKeeper 的容量配额使用率超过阈值 1313
10.13.248 ALM-45736 Guardian 服务不可用 1315

11 MRS Manager 操作指导（适用于 2.x 及之前） 1318

11.1 MRS Manager 简介 1318
11.2 查看集群运行任务 1320
11.3 监控管理 1321
11.3.1 系统概览 1321
11.3.2 管理服务和主机监控 1322
11.3.3 管理资源分布 1326
11.3.4 配置监控指标转储 1327

11.4 告警管理 1328
11.4.1 查看与手动清除告警 1328
11.4.2 配置监控与告警阈值 1329
11.4.3 配置 Syslog 北向参数 1331
11.4.4 配置 SNMP 北向参数 1333
11.5 对象管理 1335
11.5.1 对象管理简介 1335
11.5.2 查看配置 1336
11.5.3 管理服务操作 1336
11.5.4 配置服务参数 1337
11.5.5 配置服务自定义参数 1338
11.5.6 同步服务配置 1339
11.5.7 管理角色实例操作 1340
11.5.8 配置角色实例参数 1340
11.5.9 同步角色实例配置 1341
11.5.10 退服和入服务角色实例 1342
11.5.11 管理主机操作 1342
11.5.12 隔离主机 1343
11.5.13 取消隔离主机 1343
11.5.14 启动及停止集群 1344
11.5.15 同步集群配置 1344
11.5.16 导出集群的配置数据 1345
11.6 日志管理 1345
11.6.1 关于日志 1345
11.6.2 Manager 日志清单 1356
11.6.3 查看及导出审计日志 1363
11.6.4 导出服务日志 1365
11.6.5 配置审计日志导出参数 1366
11.7 健康检查管理 1367
11.7.1 执行健康检查 1367
11.7.2 查看并导出检查报告 1368
11.7.3 配置健康检查报告保存数 1368
11.7.4 管理健康检查报告 1369
11.7.5 DBService 健康检查指标项说明 1369
11.7.6 Flume 健康检查指标项说明 1370
11.7.7 HBase 健康检查指标项说明 1370
11.7.8 Host 健康检查指标项说明 1370
11.7.9 HDFS 健康检查指标项说明 1377
11.7.10 Hive 健康检查指标项说明 1377
11.7.11 Kafka 健康检查指标项说明 1378
11.7.12 KrbServer 健康检查指标项说明 1378
11.7.13 LdapServer 健康检查指标项说明 1379

11.7.14 Loader 健康检查指标项说明 1380
11.7.15 MapReduce 健康检查指标项说明 1381
11.7.16 OMS 健康检查指标项说明 1381
11.7.17 Spark 健康检查指标项说明 1385
11.7.18 Storm 健康检查指标项说明 1385
11.7.19 Yarn 健康检查指标项说明 1386
11.7.20 ZooKeeper 健康检查指标项说明 1386
11.8 静态服务池管理 1387
11.8.1 查看静态服务池状态 1387
11.8.2 配置静态服务池 1389
11.9 租户管理 1391
11.9.1 租户简介 1391
11.9.2 添加租户 1392
11.9.3 添加子租户 1394
11.9.4 删除租户 1396
11.9.5 管理租户目录 1397
11.9.6 恢复租户数据 1398
11.9.7 添加资源池 1399
11.9.8 修改资源池 1400
11.9.9 删除资源池 1400
11.9.10 配置队列 1401
11.9.11 配置资源池的队列容量策略 1402
11.9.12 清除队列配置 1402
11.10 备份与恢复 1403
11.10.1 备份与恢复简介 1403
11.10.2 备份元数据 1405
11.10.3 恢复元数据 1406
11.10.4 修改备份任务 1408
11.10.5 查看备份恢复任务 1409
11.11 安全管理 1410
11.11.1 未开启 Kerberos 认证集群中的默认用户清单 1410
11.11.2 开启 Kerberos 认证集群中的默认用户清单 1413
11.11.3 修改操作系统用户密码 1418
11.11.4 修改 admin 密码 1418
11.11.5 修改 Kerberos 管理员密码 1420
11.11.6 修改 LDAP 管理员和 LDAP 用户密码 1421
11.11.7 修改组件运行用户密码 1422
11.11.8 修改 OMS 数据库管理员密码 1423
11.11.9 修改 OMS 数据库数据访问用户密码 1423
11.11.10 修改组件数据库用户密码 1424
11.11.11 更新集群密钥 1425
11.12 权限管理 1426

11.12.1 创建角色 1426
11.12.2 创建用户组 1431
11.12.3 创建用户 1432
11.12.4 修改用户信息 1434
11.12.5 锁定用户 1434
11.12.6 解锁用户 1435
11.12.7 删除用户 1435
11.12.8 修改操作用户密码 1435
11.12.9 初始化系统用户密码 1436
11.12.10 下载用户认证文件 1437
11.12.11 修改密码策略 1438
11.13 MRS 多用户权限管理 1439
11.13.1 MRS 集群中的用户与权限 1439
11.13.2 开启 Kerberos 认证集群中的默认用户清单 1443
11.13.3 创建角色 1448
11.13.4 创建用户组 1453
11.13.5 创建用户 1455
11.13.6 修改用户信息 1456
11.13.7 锁定用户 1457
11.13.8 解锁用户 1457
11.13.9 删除用户 1458
11.13.10 修改操作用户密码 1458
11.13.11 初始化系统用户密码 1459
11.13.12 下载用户认证文件 1461
11.13.13 修改密码策略 1461
11.13.14 配置跨集群互信 1463
11.13.15 配置并使用互信集群的用户 1466
11.13.16 配置 MRS 多用户访问 OBS 细粒度权限 1467
11.14 补丁操作指导 1471
11.14.1 补丁操作指导 1471
11.14.2 支持滚动补丁 1472
11.15 修复隔离主机补丁 1475
11.16 支持滚动重启 1475
12 MRS 集群组件操作指导 1479
12.1 使用 Alluxio 1479
12.1.1 配置底层存储系统 1479
12.1.2 通过数据应用访问 Alluxio 1480
12.1.3 Alluxio 常用操作 1483
12.2 使用 CarbonData（MRS 3.x 之前版本） 1485
12.2.1 从零开始使用 CarbonData 1485
12.2.2 CarbonData 表简介 1487
12.2.3 创建 CarbonData 表 1488

12.2.4 删除 CarbonData 表 1489
12.3 使用 CarbonData（MRS 3.x 及之后版本） 1490
12.3.1 概述 1490
12.3.1.1 CarbonData 简介 1490
12.3.1.2 CarbonData 主要规格 1492
12.3.2 配置参考 1494
12.3.3 CarbonData 操作指导 1505
12.3.3.1 CarbonData 快速入门 1505
12.3.3.2 管理 CarbonData Table 1508
12.3.3.2.1 CarbonData Table 简介 1508
12.3.3.2.2 新建 CarbonData Table 1509
12.3.3.2.3 删除 CarbonData Table 1511
12.3.3.2.4 修改 CarbonData Table 1512
12.3.3.3 管理 CarbonData Table 数据 1512
12.3.3.3.1 加载数据 1512
12.3.3.3.2 删除 Segments 1513
12.3.3.3.3 合并 Segments 1514
12.3.3.4 迁移 CarbonData 数据 1517
12.3.3.5 迁移 Spark1.5 的 Carbondata 数据到 Spark2x 的 Carbondata 中 1518
12.3.4 CarbonData 性能调优 1520
12.3.4.1 调优指导 1520
12.3.4.2 创建 CarbonData Table 的建议 1522
12.3.4.3 性能调优的相关配置 1524
12.3.5 CarbonData 访问控制 1526
12.3.6 CarbonData 语法参考 1528
12.3.6.1 DDL 1528
12.3.6.1.1 CREATE TABLE 1528
12.3.6.1.2 CREATE TABLE As SELECT 1531
12.3.6.1.3 DROP TABLE 1531
12.3.6.1.4 SHOW TABLES 1532
12.3.6.1.5 ALTER TABLE COMPACTION 1533
12.3.6.1.6 TABLE RENAME 1534
12.3.6.1.7 ADD COLUMNS 1535
12.3.6.1.8 DROP COLUMNS 1536
12.3.6.1.9 CHANGE DATA TYPE 1537
12.3.6.1.10 REFRESH TABLE 1538
12.3.6.1.11 REGISTER INDEX TABLE 1539
12.3.6.2 DML 1540
12.3.6.2.1 LOAD DATA 1540
12.3.6.2.2 UPDATE CARBON TABLE 1544
12.3.6.2.3 DELETE RECORDS from CARBON TABLE 1546
12.3.6.2.4 INSERT INTO CARBON TABLE 1547

12.3.6.2.5 DELETE SEGMENT by ID 1548
12.3.6.2.6 DELETE SEGMENT by DATE 1548
12.3.6.2.7 SHOW SEGMENTS 1549
12.3.6.2.8 CREATE SECONDARY INDEX 1550
12.3.6.2.9 SHOW SECONDARY INDEXES 1551
12.3.6.2.10 DROP SECONDARY INDEX 1552
12.3.6.2.11 CLEAN FILES 1553
12.3.6.2.12 SET/RESET 1554
12.3.6.3 操作并发 1557
12.3.6.4 API 1560
12.3.6.5 空间索引 1562
12.3.7 CarbonData 故障处理 1574
12.3.7.1 当在 Filter 中使用 Big Double 类型数值时，过滤结果与 Hive 不一致 1574
12.3.7.2 查询性能下降 1575
12.3.8 CarbonData FAQ 1575
12.3.8.1 为什么对 decimal 数据类型进行带过滤条件的查询时会出现异常输出? 1575
12.3.8.2 如何避免对历史数据进行 minor compaction? 1576
12.3.8.3 如何在 CarbonData 数据加载时修改默认的组名? 1577
12.3.8.4 为什么 INSERT INTO CARBON TABLE 失败? 1577
12.3.8.5 为什么含转义字符的输入数据记录到 Bad Records 中的值与原始数据不同? 1578
12.3.8.6 为什么 Bad Records 导致数据加载性能降低? 1578
12.3.8.7 当初始 Executor 为 0 时，为什么 INSERT INTO/LOAD DATA 任务分配不正确，打开的 task 少于可用的 Executor? 1578
12.3.8.8 为什么并行度大于待处理的 block 数目时，CarbonData 仍需要额外的 executor? 1579
12.3.8.9 为什么在 off heap 时数据加载失败? 1579
12.3.8.10 为什么创建 Hive 表失败? 1579
12.3.8.11 为什么在 V100R002C50RC1 版本中创建的 CarbonData 表不具有 Hive 特权为非所有者提供的特权? 1580
12.3.8.12 如何在不同的 namespaces 上逻辑地分割数据 1581
12.3.8.13 为什么 drop 数据库抛出 Missing Privileges 异常? 1582
12.3.8.14 为什么在 Spark Shell 中不能执行更新命令? 1582
12.3.8.15 如何在 CarbonData 中配置非安全内存? 1582
12.3.8.16 设置了 HDFS 存储目录的磁盘空间配额，CarbonData 为什么会发生异常? 1583
12.3.8.17 为什么数据查询/加载失败，且抛出“org.apache.carbondata.core.memory.MemoryException: Not enough memory”异常? 1583
12.3.8.18 开启防误删下，为什么 Carbon 表没有执行 drop table 命令，回收站中也会存在该表的文件? 1584
12.4 使用 ClickHouse 1584
12.4.1 从零开始使用 ClickHouse 1584
12.4.2 ClickHouse 表引擎介绍 1587
12.4.3 ClickHouse 表创建 1593
12.4.4 ClickHouse 常用 SQL 语法 1598
12.4.4.1 CREATE DATABASE 创建数据库 1598
12.4.4.2 CREATE TABLE 创建表 1598

12.4.4.3 INSERT INTO 插入表数据 1599
12.4.4.4 SELECT 查询表数据 1600
12.4.4.5 ALTER TABLE 修改表结构 1601
12.4.4.6 DESC 查询表结构 1602
12.4.4.7 DROP 删除表 1602
12.4.4.8 SHOW 显示数据库和表信息 1602
12.4.5 ClickHouse 数据迁移 1603
12.4.5.1 ClickHouse 数据导入导出 1603
12.4.5.2 将 Kafka 数据同步至 ClickHouse 1604
12.4.5.3 使用 ClickHouse 数据迁移工具 1608
12.4.6 用户管理及认证 1611
12.4.6.1 ClickHouse 用户及权限管理 1611
12.4.6.2 ClickHouse 使用 OpenLDAP 认证 1616
12.4.7 通过数据文件备份恢复 ClickHouse 数据 1619
12.4.8 ClickHouse 日志介绍 1621
12.5 使用 DBService 1623
12.5.1 DBService 日志介绍 1623
12.6 使用 Flink 1626
12.6.1 从零开始使用 Flink 1626
12.6.2 查看 Flink 作业信息 1633
12.6.3 配置管理 Flink 1634
12.6.3.1 配置参数路径 1634
12.6.3.2 JobManager & TaskManager 1634
12.6.3.3 Blob 1639
12.6.3.4 Distributed Coordination (via Akka) 1640
12.6.3.5 SSL 1644
12.6.3.6 Network communication (via Netty) 1646
12.6.3.7 JobManager Web Frontend 1647
12.6.3.8 File Systems 1650
12.6.3.9 State Backend 1650
12.6.3.10 Kerberos-based Security 1652
12.6.3.11 HA 1653
12.6.3.12 Environment 1655
12.6.3.13 Yarn 1656
12.6.3.14 Pipeline 1657
12.6.4 安全配置 1658
12.6.4.1 安全特性描述 1658
12.6.4.2 配置对接 Kafka 1658
12.6.4.3 配置 Pipeline 1660
12.6.5 安全加固 1660
12.6.5.1 认证和加密 1661
12.6.5.2 ACL 控制 1667

12.6.5.3 web 安全 1667
12.6.6 安全声明 1670
12.6.7 使用 Flink WebUI 1670
12.6.7.1 概述 1670
12.6.7.1.1 Flink WebUI 应用简介 1670
12.6.7.1.2 Flink WebUI 应用流程 1671
12.6.7.2 FlinkServer 权限管理 1673
12.6.7.2.1 概述 1673
12.6.7.2.2 基于用户和角色的鉴权 1673
12.6.7.3 访问 Flink WebUI 1674
12.6.7.4 在 Flink WebUI 创建应用 1675
12.6.7.5 在 Flink WebUI 创建集群连接 1675
12.6.7.6 在 Flink WebUI 创建数据连接 1677
12.6.7.7 使用 Flink WebUI 的流表管理 1679
12.6.7.8 使用 Flink WebUI 的作业管理 1681
12.6.8 Flink 日志介绍 1686
12.6.9 Flink 性能调优 1687
12.6.9.1 DataStream 调优 1687
12.6.9.1.1 配置内存 1688
12.6.9.1.2 设置并行度 1688
12.6.9.1.3 配置进程参数 1689
12.6.9.1.4 设计分区方法 1690
12.6.9.1.5 配置 netty 网络通信 1691
12.6.9.1.6 经验总结 1691
12.6.10 Flink 常见 Shell 命令 1692
12.6.11 参考 1697
12.6.11.1 签发证书样例 1697
12.7 使用 Flume 1701
12.7.1 从零开始使用 Flume 1701
12.7.2 使用简介 1707
12.7.3 安装 Flume 客户端 1709
12.7.3.1 安装 MRS 3.x 之前版本 Flume 客户端 1710
12.7.3.2 安装 MRS 3.x 及之后版本 Flume 客户端 1712
12.7.4 查看 Flume 客户端日志 1714
12.7.5 停止或卸载 Flume 客户端 1715
12.7.6 使用 Flume 客户端加密工具 1716
12.7.7 Flume 业务配置指南 1716
12.7.8 Flume 配置参数说明 1739
12.7.9 在配置文件 properties.properties 中使用环境变量 1751
12.7.10 非加密传输 1752
12.7.10.1 配置非加密传输 1752
12.7.10.2 典型场景：从本地采集静态日志保存到 Kafka 1754

12.7.10.3 典型场景：从本地采集静态日志保存到 HDFS.....1759
12.7.10.4 典型场景：从本地采集动态日志保存到 HDFS.....1764
12.7.10.5 典型场景：从 Kafka 采集日志保存到 HDFS.....1769
12.7.10.6 典型场景：从 Kafka 客户端采集日志经 Flume 客户端保存到 HDFS.....1774
12.7.10.7 典型场景：从本地采集静态日志保存到 HBase.....1777
12.7.11 加密传输.....1782
12.7.11.1 配置加密传输.....1782
12.7.11.2 典型场景：从本地采集静态日志保存到 HDFS.....1790
12.7.12 查看 Flume 客户端监控信息.....1800
12.7.13 Flume 对接安全 Kafka 指导.....1801
12.7.14 Flume 对接安全 Hive 指导.....1801
12.7.15 Flume 业务模型配置指导.....1804
12.7.15.1 概述.....1804
12.7.15.2 业务模型配置指导.....1804
12.7.16 Flume 日志介绍.....1809
12.7.17 Flume 客户端 Cgroup 使用指导.....1811
12.7.18 Flume 第三方插件二次开发指导.....1812
12.7.19 Flume 常见问题.....1813
12.8 使用 HBase.....1814
12.8.1 从零开始使用 HBase.....1814
12.8.2 使用 HBase 客户端.....1818
12.8.3 创建 HBase 角色.....1820
12.8.4 配置 HBase 备份.....1822
12.8.5 配置 HBase 参数.....1830
12.8.6 启用集群间拷贝功能.....1831
12.8.7 使用 ReplicationSyncUp 工具.....1832
12.8.8 使用 HIndex.....1833
12.8.8.1 HIndex 介绍.....1833
12.8.8.2 批量加载索引数据.....1842
12.8.8.3 使用索引生成工具.....1845
12.8.8.4 索引数据迁移.....1847
12.8.9 配置 HBase 容灾.....1849
12.8.10 配置 HBase 数据压缩和编码.....1856
12.8.11 HBase 容灾业务切换.....1858
12.8.12 HBase 容灾主备集群倒换.....1859
12.8.13 社区 BulkLoad Tool.....1861
12.8.14 配置 MOB.....1861
12.8.15 配置安全的 HBase Replication.....1862
12.8.16 配置 Region Transition 恢复线程.....1863
12.8.17 使用二级索引.....1864
12.8.18 HBase 日志介绍.....1865
12.8.19 HBase 性能调优.....1868

12.8.19.1 提升 BulkLoad 效率 1868
12.8.19.2 提升连续 put 场景性能 1869
12.8.19.3 Put 和 Scan 性能综合调优 1870
12.8.19.4 提升实时写数据效率 1872
12.8.19.5 提升实时读数据效率 1880
12.8.19.6 JVM 参数优化 1887
12.8.20 HBase 常见问题 1887
12.8.20.1 客户端连接服务端时，长时间无法连接成功 1887
12.8.20.2 结束 BulkLoad 客户端程序，导致作业执行失败 1889
12.8.20.3 在 HBase 连续对同一个表名做删除创建操作时，可能出现创建表异常 1889
12.8.20.4 HBase 占用网络端口，连接数过大会导致其他服务不稳定 1890
12.8.20.5 HBase bulkload 任务（单个表有 26T 数据）有 210000 个 map 和 10000 个 reduce，任务失败 1890
12.8.20.6 如何修复长时间处于 RIT 状态的 Region 1891
12.8.20.7 HMaster 等待 namespace 表上线时超时退出 1891
12.8.20.8 客户端查询 HBase 出现 SocketTimeoutException 异常 1892
12.8.20.9 使用 scan 命令仍然可以查询到已修改和已删除的数据 1893
12.8.20.10 在启动 HBase shell 时，为什么会抛出“java.lang.UnsatisfiedLinkError: Permission denied”异常 1894
12.8.20.11 在 HMaster Web UI 中显示处于“Dead Region Servers”状态的 RegionServer 什么时候会被清除掉 1894
12.8.20.12 使用 HBase bulkload 导入数据成功，执行相同的查询时却可能返回不同的结果 1895
12.8.20.13 如何处理由于 Region 处于 FAILED_OPEN 状态而造成的建表失败异常 1895
12.8.20.14 如何清理由于建表失败残留在 ZooKeeper 中/hbase/table-lock 目录下的表名 1896
12.8.20.15 为什么给 HDFS 上的 HBase 使用的目录设置 quota 会造成 HBase 故障 1896
12.8.20.16 为什么在使用 OfflineMetaRepair 工具重新构建元数据后，HMaster 启动的时候会等待 namespace 表分配超时，最后启动失败 1897
12.8.20.17 为什么 splitWAL 期间 HMaster 日志中频繁打印出 FileNotFoundException 及 no lease 信息 1898
12.8.20.18 当使用与 Region Server 相同的 Linux 用户但不同的 kerberos 用户时，为什么 ImportTsv 工具执行失败报“Permission denied”的异常 1899
12.8.20.19 租户访问 Phoenix 提示权限不足 1900
12.8.20.20 如何解决 HBase 恢复数据任务失败后错误详情中提示：Rollback recovery failed 的回滚失败问题 1900
12.8.20.21 如何修复 Region Overlap 1901
12.8.20.22 HBase RegionServer GC 参数 Xms，Xmx 配置 31G，导致 RegionServer 启动失败 1902
12.8.20.23 使用集群内节点执行批量导入，为什么 LoadIncrementalHFiles 工具执行失败报“Permission denied”的异常 1902
12.8.20.24 Phoenix sqlline 脚本使用，报 import argparse 错误 1903
12.8.20.25 Phoenix BulkLoad Tool 限制 1904
12.8.20.26 CTBase 对接 Ranger 权限插件，提示权限不足 1905
12.9 使用 HDFS 1905
12.9.1 从零开始使用 Hadoop 1905
12.9.2 配置内存管理 1907
12.9.3 创建 HDFS 角色 1908
12.9.4 使用 HDFS 客户端 1910

12.9.5 使用 distcp 命令 1912
12.9.6 HDFS 文件系统目录简介 1916
12.9.7 更改 DataNode 的存储目录 1922
12.9.8 配置 HDFS 目录权限 1925
12.9.9 配置 NFS 1926
12.9.10 规划 HDFS 容量 1927
12.9.11 设置 HBase 和 HDFS 的 ulimit 1930
12.9.12 配置 DataNode 容量均衡 1931
12.9.13 配置 DataNode 节点间容量异构时的副本放置策略 1935
12.9.14 配置 HDFS 单目录文件数量 1936
12.9.15 配置回收站机制 1936
12.9.16 配置文件和目录的权限 1937
12.9.17 配置 token 的最大存活时间和时间间隔 1938
12.9.18 配置磁盘坏卷 1938
12.9.19 使用安全加密通道 1939
12.9.20 在网络不稳定的情况下，降低客户端运行异常概率 1940
12.9.21 配置 NameNode blacklist 1941
12.9.22 优化 HDFS NameNode RPC 的服务质量 1943
12.9.23 优化 HDFS DataNode RPC 的服务质量 1945
12.9.24 配置 DataNode 预留磁盘百分比 1945
12.9.25 配置 HDFS NodeLabel 1946
12.9.26 配置 HDFS Mover 1951
12.9.27 使用 HDFS AZ Mover 1952
12.9.28 配置 HDFS DiskBalancer 1954
12.9.29 配置从 NameNode 支持读 1956
12.9.30 使用 HDFS 文件并发操作命令 1957
12.9.31 HDFS 日志介绍 1959
12.9.32 HDFS 性能调优 1963
12.9.32.1 提升写性能 1963
12.9.32.2 使用客户端元数据缓存提高读取性能 1964
12.9.32.3 使用当前活动缓存提升客户端与 NameNode 的连接性能 1965
12.9.33 HDFS 常见问题 1966
12.9.33.1 NameNode 启动慢 1966
12.9.33.2 DataNode 状态正常，但无法正常上报数据块 1967
12.9.33.3 HDFS Web UI 无法正常刷新损坏数据的信息 1968
12.9.33.4 distcp 命令在安全集群上失败并抛出异常 1968
12.9.33.5 当 dfs.datanode.data.dir 中定义的磁盘数量等于 dfs.datanode.failed.volumes.tolerated 的值时，DataNode 启动失败 1969
12.9.33.6 当多个 data.dir 被配置在一个磁盘分区内，DataNode 的容量计算将会出错 1969
12.9.33.7 当 Standby NameNode 存储元数据（命名空间）时，出现断电的情况，Standby NameNode 启动失败 1970
12.9.33.8 在存储小文件过程中，系统断电，缓存中的数据丢失 1971
12.9.33.9 FileInputFormat split 的时候出现数组越界 1971

12.9.33.10 当分级存储策略为 LAZY_PERSIST 时，为什么文件的副本的存储类型都是 DISK.....1972
12.9.33.11 NameNode 节点长时间满负载，HDFS 客户端无响应.....1972
12.9.33.12 DataNode 禁止手动删除或修改数据存储目录.....1973
12.9.33.13 成功回滚后，为什么 NameNode UI 上显示有一些块缺失.....1973
12.9.33.14 为什么在往 HDFS 写数据时报"java.net.SocketException: No buffer space available"异常.....1974
12.9.33.15 为什么主 NameNode 重启后系统出现双备现象.....1976
12.9.33.16 HDFS 执行 Balance 时被异常停止，再次执行 Balance 会失败.....1977
12.9.33.17 IE 浏览器访问 HDFS 原生 UI 界面失败，显示无法显示此页.....1978
12.9.33.18 EditLog 不连续导致 NameNode 启动失败.....1978
12.10 使用 Hive.....1979
12.10.1 从零开始使用 Hive.....1979
12.10.2 配置 Hive 常用参数.....1983
12.10.3 Hive SQL.....1984
12.10.4 权限管理.....1987
12.10.4.1 Hive 权限介绍.....1987
12.10.4.2 创建 Hive 角色.....1990
12.10.4.3 配置 Hive 表、列或数据库的权限.....1994
12.10.4.4 配置 Hive 业务使用其他组件的权限.....1997
12.10.5 使用 Hive 客户端.....2001
12.10.6 使用 HDFS Colocation 存储 Hive 表.....2004
12.10.7 使用 Hive 列加密功能.....2005
12.10.8 自定义行分隔符.....2006
12.10.9 配置跨集群互信下 Hive on HBase.....2007
12.10.10 删除 Hive on HBase 表中的单行记录.....2008
12.10.11 配置基于 HTTPS/HTTP 协议的 REST 接口.....2008
12.10.12 配置是否禁用 Transform 功能.....2009
12.10.13 Hive 支持创建单表动态视图授权访问控制.....2009
12.10.14 配置创建临时函数是否需要 ADMIN 权限.....2010
12.10.15 使用 Hive 读取关系型数据库数据.....2011
12.10.16 Hive 支持的传统关系型数据库语法.....2012
12.10.17 创建 Hive 用户自定义函数.....2014
12.10.18 beeline 可靠性增强特性介绍.....2016
12.10.19 具备表 select 权限可用 show create table 查看表结构.....2017
12.10.20 Hive 写目录旧数据进回收站.....2018
12.10.21 Hive 能给一个不存在的目录插入数据.....2018
12.10.22 限定仅 admin 用户能创建库和在 default 库建表.....2019
12.10.23 限定创建 Hive 内部表不能指定 location.....2020
12.10.24 允许在只读权限的目录建外表.....2021
12.10.25 Hive 支持授权超过 32 个角色.....2021
12.10.26 Hive 任务支持限定最大 map 数.....2022
12.10.27 HiveServer 租约隔离使用.....2023
12.10.28 Hive 支持事务.....2024

12.10.29 切换 Hive 执行引擎为 Tez 2028
12.10.30 Hive 物化视图 2030
12.10.31 Hive 日志介绍 2032
12.10.32 Hive 性能调优 2035
12.10.32.1 建立表分区 2035
12.10.32.2 Join 优化 2036
12.10.32.3 Group By 优化 2038
12.10.32.4 数据存储优化 2039
12.10.32.5 SQL 优化 2039
12.10.32.6 使用 Hive CBO 优化查询 2041
12.10.33 Hive 常见问题 2042
12.10.33.1 如何在多个 HiveServer 之间同步删除 UDF 2042
12.10.33.2 已备份的 Hive 表无法执行 drop 操作 2043
12.10.33.3 如何在 Hive 自定义函数中操作本地文件 2044
12.10.33.4 如何强制停止 Hive 执行的 MapReduce 任务 2044
12.10.33.5 如何对 Hive 表大小数据进行监控 2045
12.10.33.6 如何对重点目录进行保护，防止“insert overwrite”语句误操作导致数据丢失 2045
12.10.33.7 未安装 HBase 时 Hive on Spark 任务卡顿处理 2046
12.10.33.8 FusionInsight Hive 使用 WHERE 条件查询超过 3.2 万分区的表报错 2046
12.10.33.9 使用 IBM 的 jdk 访问 Beeline 客户端出现连接 hiveserver 失败 2047
12.10.33.10 关于 Hive 表的 location 支持跨 OBS 和 HDFS 路径的说明 2047
12.10.33.11 通过 Tez 引擎执行 union 相关语句写入的数据，切换 MR 引擎后查询不出来。 2048
12.10.33.12 Hive 不支持对同一张表或分区进行并发写数据 2048
12.10.33.13 Hive 不支持向量化查询 2048
12.10.33.14 Hive 表 HDFS 数据目录被误删，但是元数据仍然存在，导致执行任务报错处理 2048
12.10.33.15 如何关闭 Hive 客户端日志 2049
12.10.33.16 Hive 快删目录配置类问题 2050
12.10.33.17 Hive 配置类问题 2050
12.11 使用 Hue（MRS 3.x 之前版本） 2051
12.11.1 从零开始使用 Hue 2051
12.11.2 访问 Hue 的 WebUI 2052
12.11.3 Hue 常用参数 2053
12.11.4 在 Hue WebUI 使用 HiveQL 编辑器 2053
12.11.5 在 Hue WebUI 使用元数据浏览器 2056
12.11.6 在 Hue WebUI 使用文件浏览器 2059
12.11.7 在 Hue WebUI 使用作业浏览器 2062
12.12 使用 Hue（MRS 3.x 及之后版本） 2063
12.12.1 从零开始使用 Hue 2063
12.12.2 访问 Hue 的 WebUI 2064
12.12.3 Hue 常用参数 2065
12.12.4 在 Hue WebUI 使用 HiveQL 编辑器 2066
12.12.5 在 Hue WebUI 使用 SparkSql 编辑器 2068

12.12.6 在 Hue WebUI 使用元数据浏览器 2070
12.12.7 在 Hue WebUI 使用文件浏览器 2070
12.12.8 在 Hue WebUI 使用作业浏览器 2073
12.12.9 在 Hue WebUI 使用 HBase 2074
12.12.10 典型场景 2075
12.12.10.1 HDFS on Hue 2075
12.12.10.2 配置 HDFS 冷热数据迁移 2078
12.12.10.3 Hive on Hue 2085
12.12.10.4 Oozie on Hue 2087
12.12.11 Hue 日志介绍 2088
12.12.12 Hue 常见问题 2090
12.12.12.1 如何解决使用 IE 浏览器在 Hue 中执行 HQL 失败的问题 2090
12.12.12.2 在使用 Hive 时，输入 use database 语句失效了 2090
12.12.12.3 如何处理使用 Hue WebUI 访问 HDFS 文件失败的问题 2091
12.12.12.4 Hue 页面上传大文件失败如何处理 2091
12.12.12.5 集群未安装 Hive 服务时 Hue 原生页面无法正常显示 2092
12.13 使用 Impala 2092
12.13.1 从零开始使用 Impala 2092
12.13.2 访问 Impala 的 WebUI 2095
12.13.3 使用 Impala 操作 Kudu 2096
12.13.4 Impala 对接外部 LDAP 2097
12.14 使用 Kafka 2098
12.14.1 从零开始使用 Kafka 2098
12.14.2 管理 Kafka 主题 2100
12.14.3 查看 Kafka 主题 2104
12.14.4 管理 Kafka 用户权限 2104
12.14.5 管理 Kafka 主题中的消息 2107
12.14.6 基于 binlog 的 MySQL 数据同步到 MRS 集群中 2109
12.14.7 创建 Kafka 角色 2114
12.14.8 Kafka 常用参数 2115
12.14.9 Kafka 安全使用说明 2118
12.14.10 Kafka 业务规格说明 2121
12.14.11 使用 Kafka 客户端 2122
12.14.12 配置 Kafka 高可用和高可靠参数 2123
12.14.13 更改 Broker 的存储目录 2126
12.14.14 查看 Consumer Group 消费情况 2127
12.14.15 Kafka 均衡工具使用说明 2129
12.14.16 Kafka 扩容节点后数据均衡 2131
12.14.17 Kafka Token 认证机制工具使用说明 2134
12.14.18 Kafka 日志介绍 2135
12.14.19 性能调优 2137
12.14.19.1 Kafka 性能调优 2137

12.14.20 Kafka 特性说明 2138
12.14.21 Kafka 节点内数据迁移 2140
12.14.22 Kafka 常见问题 2142
12.14.22.1 如何解决 Kafka topic 无法删除的问题 2142
12.15 使用 KafkaManager 2142
12.15.1 KafkaManager 介绍 2143
12.15.2 访问 KafkaManager 的 WebUI 2143
12.15.3 管理 Kafka 集群 2144
12.15.4 Kafka 集群监控管理 2145
12.16 使用 Kudu 2152
12.16.1 从零开始使用 Kudu 2152
12.16.2 访问 Kudu 的 WebUI 2153
12.17 使用 Loader 2154
12.17.1 从零开始使用 Loader 2154
12.17.2 Loader 使用简介 2155
12.17.3 Loader 连接配置说明 2156
12.17.4 管理 Loader 连接（MRS 3.x 之前版本） 2158
12.17.5 Loader 作业源连接配置说明 2160
12.17.6 Loader 作业目的连接配置说明 2162
12.17.7 管理 Loader 作业 2165
12.17.8 准备 MySQL 数据库连接的驱动 2168
12.17.9 Loader 日志介绍 2169
12.17.10 样例：通过 Loader 将数据从 OBS 导入 HDFS 2172
12.17.11 Loader 常见问题 2173
12.17.11.1 IE 10&IE 11 浏览器无法保存数据 2173
12.17.11.2 将 Oracle 数据库中的数据导入 HDFS 时各连接器的区别 2174
12.18 使用 Mapreduce 2175
12.18.1 配置日志归档和清理机制 2175
12.18.2 降低客户端应用的失败率 2176
12.18.3 将 MR 任务从 Windows 上提交到 Linux 上运行 2177
12.18.4 配置使用分布式缓存 2177
12.18.5 配置 MapReduce shuffle address 2179
12.18.6 配置集群管理员列表 2180
12.18.7 MapReduce 日志介绍 2180
12.18.8 MapReduce 性能调优 2183
12.18.8.1 多 CPU 内核下的调优配置 2183
12.18.8.2 确定 Job 基线 2186
12.18.8.3 Shuffle 调优 2188
12.18.8.4 大任务的 AM 调优 2192
12.18.8.5 推测执行 2192
12.18.8.6 通过“Slow Start”调优 2193
12.18.8.7 MR job commit 阶段优化 2193

12.18.9 MapReduce 常见问题 2194
12.18.9.1 ResourceManager 进行主备切换后，任务中断后运行时间过长 2194
12.18.9.2 MapReduce 任务长时间无进展 2194
12.18.9.3 运行任务时，客户端不可用 2195
12.18.9.4 在缓存中找不到 HDFS_DELEGATION_TOKEN 2195
12.18.9.5 如何在提交 MapReduce 任务时设置任务优先级 2196
12.18.9.6 MapReduce 任务运行失败，ApplicationMaster 出现物理内存溢出异常 2196
12.18.9.7 MapReduce JobHistoryServer 服务地址变更后，为什么运行完的 MapReduce 作业信息无法通过 ResourceManager Web UI 页面的 Tracking URL 打开 2197
12.18.9.8 多个 NameService 环境下，运行 MapReduce 任务失败 2198
12.18.9.9 基于分区的任务黑名单 2198
12.19 使用 Oozie 2199
12.19.1 从零开始使用 Oozie 2199
12.19.2 使用 Oozie 客户端 2200
12.19.3 使用 Oozie 客户端提交作业 2201
12.19.3.1 提交 Hive 任务 2201
12.19.3.2 提交 Spark2x 任务 2203
12.19.3.3 提交 Loader 任务 2205
12.19.3.4 提交 DistCp 任务 2207
12.19.3.5 提交其它任务 2209
12.19.4 使用 Hue 提交 Oozie 作业 2211
12.19.4.1 创建工作流 2211
12.19.4.2 提交 Workflow 工作流作业 2212
12.19.4.2.1 提交 Hive2 作业 2213
12.19.4.2.2 提交 Spark2x 作业 2214
12.19.4.2.3 提交 Java 作业 2215
12.19.4.2.4 提交 Loader 作业 2216
12.19.4.2.5 提交 Mapreduce 作业 2217
12.19.4.2.6 提交 Sub workflow 作业 2218
12.19.4.2.7 提交 Shell 作业 2218
12.19.4.2.8 提交 HDFS 作业 2219
12.19.4.2.9 提交 Streaming 作业 2220
12.19.4.2.10 提交 Distcp 作业 2221
12.19.4.2.11 互信操作示例 2222
12.19.4.2.12 提交 SSH 作业 2223
12.19.4.2.13 提交 Hive 脚本 2224
12.19.4.3 提交 Coordinator 定时调度作业 2224
12.19.4.4 提交 Bundle 批处理作业 2225
12.19.4.5 作业结果查询 2226
12.19.5 Oozie 日志介绍 2227
12.19.6 Oozie 常见问题 2229
12.19.6.1 Oozie 定时任务没有准时运行 2229
12.19.6.2 HDFS 上更新了 oozie 的 share lib 目录但没有生效 2229

12.19.6.3 Oozie 常用排查手段 2229
12.20 使用 Presto 2230
12.20.1 访问 Presto 的 WebUI 2230
12.20.2 使用客户端执行查询语句 2232
12.21 使用 Ranger（MRS 3.x） 2233
12.21.1 登录 Ranger 管理界面 2233
12.21.2 启用 Ranger 鉴权 2235
12.21.3 配置组件权限策略 2236
12.21.4 查看 Ranger 审计信息 2238
12.21.5 配置 Ranger 安全区 2239
12.21.6 普通集群修改 Ranger 数据源为 Ldap 2241
12.21.7 查看 Ranger 权限信息 2242
12.21.8 添加 HDFS 的 Ranger 访问权限策略 2244
12.21.9 添加 HBase 的 Ranger 访问权限策略 2247
12.21.10 添加 Hive 的 Ranger 访问权限策略 2251
12.21.11 添加 Yarn 的 Ranger 访问权限策略 2259
12.21.12 添加 Spark2x 的 Ranger 访问权限策略 2262
12.21.13 添加 Kafka 的 Ranger 访问权限策略 2269
12.21.14 添加 Storm 的 Ranger 访问权限策略 2277
12.21.15 Ranger 日志介绍 2279
12.21.16 Ranger 常见问题 2281
12.21.16.1 安装集群过程中，Ranger 启动失败 2281
12.21.16.2 如何判断某个服务是否使用了 Ranger 鉴权 2281
12.21.16.3 新创建用户修改完密码后无法登录 Ranger 2282
12.21.16.4 Ranger 界面添加或者修改 HBase 策略时，无法使用通配符搜索已存在的 HBase 表 2282
12.22 使用 Spark 2283
12.22.1 使用前须知 2283
12.22.2 从零开始使用 Spark 2283
12.22.3 从零开始使用 Spark SQL 2285
12.22.4 使用 Spark 客户端 2287
12.22.5 访问 Spark Web UI 界面 2288
12.22.6 Spark 对接 OpenTSDB 2289
12.22.6.1 创建表关联 OpenTSDB 2289
12.22.6.2 插入数据至 OpenTSDB 表 2290
12.22.6.3 查询 OpenTSDB 表 2291
12.22.6.4 默认配置修改 2292
12.23 使用 Spark2x 2292
12.23.1 使用前须知 2292
12.23.2 基本操作 2292
12.23.2.1 快速入门 2292
12.23.2.2 快速配置参数 2295
12.23.2.3 常用参数 2303

12.23.2.4 SparkOnHBase 概述及基本应用 2320
12.23.2.5 SparkOnHBasev2 概述及基本应用 2322
12.23.2.6 SparkSQL 权限管理（安全模式） 2324
12.23.2.6.1 SparkSQL 权限介绍 2324
12.23.2.6.2 创建 SparkSQL 角色 2328
12.23.2.6.3 配置表、列和数据库的权限 2331
12.23.2.6.4 配置 SparkSQL 业务使用其他组件的权限 2333
12.23.2.6.5 客户端和服务端配置 2335
12.23.2.7 场景化参数 2337
12.23.2.7.1 配置多主实例模式 2337
12.23.2.7.2 配置多租户模式 2337
12.23.2.7.3 配置多主实例与多租户模式切换 2339
12.23.2.7.4 配置事件队列的大小 2340
12.23.2.7.5 配置 executor 堆外内存大小 2341
12.23.2.7.6 增强有限内存下的稳定性 2341
12.23.2.7.7 配置 WebUI 上查看聚合后的 container 日志 2343
12.23.2.7.8 配置 YARN-Client 和 YARN-Cluster 不同模式下的环境变量 2344
12.23.2.7.9 配置 SparkSQL 的分块个数 2345
12.23.2.7.10 配置 parquet 表的压缩格式 2346
12.23.2.7.11 配置 WebUI 上显示的 Lost Executor 信息的个数 2347
12.23.2.7.12 动态设置日志级别 2347
12.23.2.7.13 配置 Spark 是否获取 HBase Token 2348
12.23.2.7.14 配置 Kafka 后进先出 2349
12.23.2.7.15 配置对接 Kafka 可靠性 2351
12.23.2.7.16 配置流式读取 driver 执行结果 2352
12.23.2.7.17 配置过滤掉分区表中路径不存在的分区 2353
12.23.2.7.18 配置 Spark2x Web UI ACL 2354
12.23.2.7.19 配置矢量化读取 ORC 数据 2355
12.23.2.7.20 Hive 分区修剪的谓词下推增强 2356
12.23.2.7.21 支持 Hive 动态分区覆盖语义 2357
12.23.2.7.22 配置列统计值直方图 Histogram 用以增强 CBO 准确度 2357
12.23.2.7.23 配置 JobHistory 本地磁盘缓存 2359
12.23.2.7.24 配置 Spark SQL 开启 Adaptive Execution 特性 2360
12.23.2.7.25 配置 eventlog 日志回滚 2362
12.23.2.8 使用 Ranger 时适配第三方 JDK 2363
12.23.3 Spark2x 日志介绍 2364
12.23.4 获取运行中 Spark 应用的 Container 日志 2367
12.23.5 小文件合并工具 2367
12.23.6 CarbonData 首查优化工具 2369
12.23.7 Spark2x 性能调优 2371
12.23.7.1 Spark Core 调优 2371
12.23.7.1.1 数据序列化 2371

12.23.7.1.2 配置内存 2372
12.23.7.1.3 设置并行度 2372
12.23.7.1.4 使用广播变量 2373
12.23.7.1.5 使用 External Shuffle Service 提升性能 2373
12.23.7.1.6 Yarn 模式下动态资源调度 2374
12.23.7.1.7 配置进程参数 2375
12.23.7.1.8 设计 DAG 2376
12.23.7.1.9 经验总结 2378
12.23.7.2 SQL 和 DataFrame 调优 2380
12.23.7.2.1 Spark SQL join 优化 2380
12.23.7.2.2 优化数据倾斜场景下的 Spark SQL 性能 2382
12.23.7.2.3 优化小文件场景下的 Spark SQL 性能 2383
12.23.7.2.4 INSERT...SELECT 操作调优 2384
12.23.7.2.5 多并发 JDBC 客户端连接 JDBCServer 2385
12.23.7.2.6 动态分区插入场景内存优化 2385
12.23.7.2.7 小文件优化 2386
12.23.7.2.8 聚合算法优化 2387
12.23.7.2.9 Datasource 表优化 2387
12.23.7.2.10 合并 CBO 优化 2388
12.23.7.2.11 跨源复杂数据的 SQL 查询优化 2389
12.23.7.2.12 多级嵌套子查询以及混合 Join 的 SQL 调优 2392
12.23.7.3 Spark Streaming 调优 2394
12.23.8 Spark2x 常见问题 2395
12.23.8.1 Spark Core 2395
12.23.8.1.1 日志聚合下，如何查看 Spark 已完成应用日志 2395
12.23.8.1.2 Driver 返回码和 RM WebUI 上应用状态显示不一致 2396
12.23.8.1.3 为什么 Driver 进程不能退出 2396
12.23.8.1.4 网络连接超时导致 FetchFailedException 2396
12.23.8.1.5 当事件队列溢出时如何配置事件队列的大小 2398
12.23.8.1.6 Spark 应用执行过程中，日志中一直打印 getApplicationReport 异常且应用较长时间不退出 2398
12.23.8.1.7 Spark 执行应用时上报“Connection to ip:port has been quiet for xxx ms while there are outstanding requests”并导致应用结束 2399
12.23.8.1.8 NodeManager 关闭导致 Executor(s)未移除 2401
12.23.8.1.9 Password cannot be null if SASL is enabled 异常 2401
12.23.8.1.10 向动态分区表中插入数据时，在重试的 task 中出现"Failed to CREATE_FILE"异常 2401
12.23.8.1.11 使用 Hash shuffle 出现任务失败 2402
12.23.8.1.12 访问 Spark 应用的聚合日志页面报“DNS 查找失败”错误 2402
12.23.8.1.13 由于 Timeout waiting for task 异常导致 Shuffle FetchFailed 2403
12.23.8.1.14 Executor 进程 Crash 导致 Stage 重试 2404
12.23.8.1.15 执行大数据量的 shuffle 过程时 Executor 注册 shuffle service 失败 2404
12.23.8.1.16 在 Spark 应用执行过程中 NodeManager 出现 OOM 异常 2405
12.23.8.1.17 安全集群使用 HiBench 工具运行 sparkbench 获取不到 realm 2406
12.23.8.2 SQL 和 DataFrame 2407

12.23.8.2.1 Spark SQL ROLLUP 和 CUBE 使用的注意事项 2407
12.23.8.2.2 Spark SQL 在不同 DB 都可以显示临时表 2408
12.23.8.2.3 如何在 Spark 命令中指定参数值 2409
12.23.8.2.4 SparkSQL 建表时的目录权限 2409
12.23.8.2.5 为什么不同服务之间互相删除 UDF 失败 2410
12.23.8.2.6 Spark SQL 无法查询到 Parquet 类型的 Hive 表的新插入数据 2410
12.23.8.2.7 cache table 使用指导 2411
12.23.8.2.8 Repartition 时有部分 Partition 没数据 2411
12.23.8.2.9 16T 的文本数据转成 4T Parquet 数据失败 2412
12.23.8.2.10 当表名为 table 时，执行相关操作时出现异常 2413
12.23.8.2.11 执行 analyze table 语句，因资源不足出现任务卡住 2413
12.23.8.2.12 为什么有时访问没有权限的 parquet 表时，在上报“Missing Privileges”错误提示之前，会运行一个 Job? 2414
12.23.8.2.13 执行 Hive 命令修改元数据时失败或不生效 2415
12.23.8.2.14 spark-sql 退出时打印 RejectedExecutionException 异常栈 2415
12.23.8.2.15 健康检查时，误将 JDBCServer Kill 2415
12.23.8.2.16 日期类型的字段作为过滤条件时匹配'2016-6-30'时没有查询结果 2416
12.23.8.2.17 为什么在启动 spark-beeline 的命令中指定“--hivevar”选项无效 2416
12.23.8.2.18 在 spark-beeline 中创建临时表/视图时，报 HDFS 目录无权限操作的错误 2417
12.23.8.2.19 执行复杂 SQL 语句时报“Code of method ... grows beyond 64 KB”的错误 2417
12.23.8.2.20 在 Beeline/JDBCServer 模式下连续运行 10T 的 TPCDS 测试套会出现内存不足的现象 2418
12.23.8.2.21 连上不同的 JDBCServer，function 不能正常使用 2418
12.23.8.2.22 Spark2x 无法访问 Spark1.5 创建的 DataSource 表 2420
12.23.8.2.23 为什么 spark-beeline 运行失败报“Failed to create ThriftService instance”的错误 2420
12.23.8.2.24 Spark SQL 无法查询到 ORC 类型的 Hive 表的新插入数据 2421
12.23.8.3 Spark Streaming 2422
12.23.8.3.1 Spark Streaming 任务一直阻塞 2422
12.23.8.3.2 运行 Spark Streaming 任务参数调优的注意事项 2423
12.23.8.3.3 为什么提交 Spark Streaming 应用超过 token 有效期，应用失败 2423
12.23.8.3.4 为什么 Spark Streaming 应用创建输入流，但该输入流无输出逻辑时，应用从 checkpoint 恢复启动失败 2424
12.23.8.3.5 Spark Streaming 应用运行过程中重启 Kafka，Web UI 界面部分 batch time 对应 Input Size 为 0 records 2426
12.23.8.4 访问 Spark 应用获取的 restful 接口信息有误 2426
12.23.8.5 为什么从 Yarn Web UI 页面无法跳转到 Spark Web UI 界面 2427
12.23.8.6 HistoryServer 缓存的应用被回收，导致此类应用页面访问时出错 2428
12.23.8.7 加载空的 part 文件时，app 无法显示在 JobHistory 的页面上 2429
12.23.8.8 Spark2x 导出带有相同字段名的表，结果导出失败 2429
12.23.8.9 为什么多次运行 Spark 应用程序会引发致命 JRE 错误 2430
12.23.8.10 IE 浏览器访问 Spark2x 原生 UI 界面失败，无法显示此页或者页面显示错误 2430
12.23.8.11 Spark2x 如何访问外部集群组件 2430
12.23.8.12 对同一目录创建多个外表，可能导致外表查询失败 2432
12.23.8.13 访问 Spark2x JobHistory 中某个应用的原生页面时页面显示错误 2433

12.23.8.14 对接 OBS 场景中，spark-beeline 登录后指定 loaction 到 OBS 建表失败 2433
12.23.8.15 Spark shuffle 异常处理 2434
12.24 使用 Sqoop 2434
12.24.1 从零开始使用 Sqoop 2434
12.24.2 Sqoop1.4.7 适配 MRS 3.x 集群 2438
12.24.3 Sqoop 常用命令及参数介绍 2440
12.24.4 Sqoop 常见问题 2443
12.24.4.1 报错找不到 QueryProvider 类 2443
12.24.4.2 连接 postgresql 或者 gaussdb 时报错 2444
12.24.4.3 使用 hive-table 方式同步数据到 obs 上的 hive 表报错 2444
12.24.4.4 使用 hive-table 方式同步数据到 orc 表或者 parquet 表失败 2445
12.24.4.5 使用 hive-table 方式同步数据报错 2445
12.24.4.6 使用 hcatalog 方式同步 hive parquet 表报错 2446
12.24.4.7 使用 Hcatalog 方式同步 Hive 和 MySQL 之间的数据，timestamp 和 data 类型字段会报错 2446
12.25 使用 Storm 2446
12.25.1 从零开始使用 Storm 2447
12.25.2 使用 Storm 客户端 2447
12.25.3 使用客户端提交 Storm 拓扑 2448
12.25.4 访问 Storm 的 WebUI 2450
12.25.5 管理 Storm 拓扑 2451
12.25.6 查看 Storm 拓扑日志 2451
12.25.7 Storm 常用参数 2452
12.25.8 配置 Storm 业务用户密码策略 2453
12.25.9 迁移 Storm 业务至 Flink 2455
12.25.9.1 概述 2455
12.25.9.2 完整迁移 Storm 业务 2456
12.25.9.3 嵌入式迁移 Storm 业务 2457
12.25.9.4 迁移 Storm 对接的外部安全组件业务 2458
12.25.10 Storm 日志介绍 2458
12.25.11 性能调优 2463
12.25.11.1 Storm 性能调优 2463
12.26 使用 Tez 2464
12.26.1 使用前须知 2464
12.26.2 Tez 常用参数 2464
12.26.3 访问 TezUI 2465
12.26.4 日志介绍 2465
12.26.5 常见问题 2467
12.26.5.1 TezUI 无法展示 Tez 任务执行细节 2467
12.26.5.2 进入 Tez 原生界面显示异常 2467
12.26.5.3 TezUI 界面无法查看 yarn 日志 2468
12.26.5.4 TezUI HiveQueries 界面表格数据为空 2469
12.27 使用 Yarn 2469

12.27.1 Yarn 常用参数 2469
12.27.2 创建 Yarn 角色 2472
12.27.3 使用 Yarn 客户端 2474
12.27.4 配置 NodeManager 角色实例使用的资源 2475
12.27.5 更改 NodeManager 的存储目录 2477
12.27.6 配置 YARN 严格权限控制 2480
12.27.7 配置 Container 日志聚合功能 2481
12.27.8 启用 CGroups 功能 2485
12.27.9 配置 AM 失败重试次数 2487
12.27.10 配置 AM 自动调整分配内存 2488
12.27.11 配置访问通道协议 2489
12.27.12 检测内存使用情况 2490
12.27.13 配置自定义调度器的 WebUI 2491
12.27.14 配置 YARN Restart 特性 2491
12.27.15 配置 AM 作业保留 2493
12.27.16 配置本地化日志级别 2494
12.27.17 配置运行任务的用户 2495
12.27.18 Yarn 日志介绍 2496
12.27.19 Yarn 性能调优 2499
12.27.19.1 抢占任务 2499
12.27.19.2 任务优先级 2500
12.27.19.3 节点配置调优 2501
12.27.20 Yarn 常见问题 2507
12.27.20.1 任务完成后 Container 挂载的文件目录未清除 2507
12.27.20.2 作业执行失败时会抛出 HDFS_DELEGATION_TOKEN 到期的异常 2507
12.27.20.3 重启 YARN，本地日志不被删除 2507
12.27.20.4 为什么执行任务时 AppAttempts 重试次数超过 2 次还没有运行失败 2508
12.27.20.5 为什么在 ResourceManager 重启后，应用程序会移回原来的队列 2508
12.27.20.6 为什么 YARN 资源池的所有节点都被加入黑名单，而 YARN 却没有释放黑名单，导致任务一直处于运行状态 2508
12.27.20.7 ResourceManager 持续主备倒换 2509
12.27.20.8 当一个 NodeManager 处于 unhealthy 的状态 10 分钟时，新应用程序失败 2509
12.27.20.9 Superior 通过 REST 接口查看已结束或不存在的 applicationID，返回的页面提示 Error Occurred 2510
12.27.20.10 Superior 调度模式下，单个 NodeManager 故障可能导致 MapReduce 任务失败 2510
12.27.20.11 当应用程序从 lost_and_found 队列移动到其他队列时，应用程序不能继续执行 2511
12.27.20.12 如何限制存储在 ZKstore 中的应用程序诊断消息的大小 2511
12.27.20.13 为什么将非 ViewFS 文件系统配置为 ViewFS 时 MapReduce 作业运行失败 2512
12.27.20.14 开启 Native Task 特性后，Reduce 任务在部分操作系统运行失败 2513
12.28 使用 ZooKeeper 2513
12.28.1 从零开始使用 Zookeeper 2513
12.28.2 ZooKeeper 常用参数 2517
12.28.3 使用 ZooKeeper 客户端 2518

12.28.4 ZooKeeper 权限设置指南 2518
12.28.5 ZooKeeper 日志介绍 2522
12.28.6 ZooKeeper 常见问题 2524
12.28.6.1 创建大量 znode 后，ZooKeeper Sever 启动失败 2524
12.28.6.2 为什么 ZooKeeper Server 出现 java.io.IOException: Len 的错误日志 2525
12.28.6.3 为什么在 Zookeeper 服务器上启用安全的 netty 配置时，四个字母的命令不能与 linux 的 netcat 命令一起使用 2527
12.28.6.4 如何查看哪个 ZooKeeper 实例是 leader 2527
12.28.6.5 使用 IBM JDK 时客户端无法连接 ZooKeeper 2528
12.28.6.6 ZooKeeper 客户端刷新 TGT 失败 2528
12.28.6.7 使用 deleteall 命令，删除大量 znode 时，偶现报错“Node does not exist”错误 2528
12.29 附录 2528
12.29.1 修改集群服务配置参数 2529
12.29.2 访问集群 Manager 2530
12.29.2.1 访问 MRS Manager（MRS 3.x 之前版本） 2530
12.29.2.2 访问 FusionInsight Manager（MRS 3.x 及之后版本） 2532
12.29.3 使用 MRS 客户端 2534
12.29.3.1 安装客户端（3.x 及之后版本） 2534
12.29.3.2 安装客户端（3.x 之前版本） 2537
12.29.3.3 更新客户端（3.x 及之后版本） 2541
12.29.3.4 更新客户端（3.x 之前版本） 2543
13 安全性说明 2547
13.1 集群（未启用 Kerberos 认证）安全配置建议 2547
13.2 安全认证原理和认证机制 2547
14 高危操作一览表 2551
15 常见问题 2573
15.1 产品咨询类 2573
15.1.1 MRS 可以做什么？ 2573
15.1.2 MRS 支持什么类型的分布式存储？ 2573
15.1.3 如何使用自定义安全组创建 MRS 集群？ 2574
15.1.4 如何使用 MRS？ 2575
15.1.5 如何保证数据和业务运行安全？ 2575
15.1.6 如何配置 Phoenix 连接池？ 2576
15.1.7 MRS 是否支持更换网段？ 2576
15.1.8 MRS 服务集群节点是否执行降配操作？ 2576
15.1.9 Hive 与其他组件有什么关系？ 2576
15.1.10 MRS 集群是否支持 Hive on Spark？ 2576
15.1.11 Hive 版本之间是否兼容？ 2576
15.1.12 MRS 集群哪个版本支持建立 Hive 连接且有用户同步功能？ 2577
15.1.13 数据存储在 OBS 和 HDFS 有什么区别？ 2577
15.1.14 Hadoop 压力测试工具如何获取？ 2577
15.1.15 Impala 与其他组件有什么关系？ 2577

15.1.16 关于 MRS 服务集成的开源第三方 SDK 中包含的公网 IP 地址声明.....2578
15.1.17 Kudu 和 HBase 间的关系？.....2578
15.1.18 MRS 是否支持 Hive on Kudu？.....2578
15.1.19 10 亿级数据量场景的解决方案.....2578
15.1.20 如何修改 DBService 的 IP？.....2578
15.1.21 MRS sudo log 能否清理？.....2578
15.1.22 MRS 2.1.0 集群版本对 Storm 日志也有 20G 的限制么.....2579
15.1.23 Spark ThriftServer 是什么.....2579
15.1.24 Kafka 目前支持的访问协议类型.....2579
15.1.25 zstd 的压缩比怎么样.....2579
15.1.26 创建 MRS 集群时，找不到 HDFS、Yarn、MapReduce 组件.....2579
15.1.27 创建 MRS 集群时，找不到 ZooKeeper 组件.....2579
15.1.28 MRS 3.1.0 集群版本，Spark 任务支持 python 哪些版本？.....2579
15.1.29 如何让不同的业务程序分别用不同的 Yarn 队列？.....2579
15.1.30 MRS 管理控制台和集群 Manager 页面区别与联系.....2582
15.1.31 MRS 如何解绑 EIP?.....2583
15.2 帐号密码类.....2583
15.2.1 登录 Manager 帐号的是什么？.....2583
15.2.2 帐号密码的过期时间如何查询和修改.....2583
15.3 帐号权限类.....2584
15.3.1 如果不开启 Kerberos 认证，MRS 集群能否支持访问权限细分？.....2585
15.3.2 如何给新建的帐号添加租户管理权限？.....2585
15.3.3 如何自定义配置 MRS 服务策略？.....2585
15.3.4 在 MRS Manager 页面“系统设置”中找不到用户管理，什么原因？.....2586
15.3.5 Hue 有没有配置帐号权限的功能？.....2586
15.4 客户端使用类.....2586
15.4.1 如何使用组件客户端？.....2586
15.4.2 怎么关闭 ZooKeeper SASL 认证.....2586
15.4.3 在 MRS 集群外客户端中执行 kinit 报错.....2586
15.5 Web 页面访问类.....2587
15.5.1 修改开源组件 Web 页面会话超时时间.....2587
15.5.2 MRS 租户管理中的动态资源计划页面无法刷新.....2589
15.5.3 Kafka Topic 监控页签在 Manager 页面不显示.....2589
15.5.4 访问 HDFS、Hue、Yarn、Flink 等组件的 WebUI 界面报错，或部分功能不可用.....2589
15.6 监控告警类.....2590
15.6.1 在 MRS 流式集群中，Kafka topic 监控是否支持发送告警？.....2590
15.6.2 产生告警“ALM-18022 Yarn 队列资源不足”时，在哪里可以看到在运行的资源队列.....2590
15.6.3 HBase 操作请求次数指标中的多级图表统计如何理解.....2590
15.7 性能优化类.....2592
15.7.1 MRS 集群是否支持重装系统？.....2592
15.7.2 MRS 集群是否支持切换操作系统？.....2592
15.7.3 如何提高集群 Core 节点的资源使用率？.....2592

15.7.4 如何关闭防火墙服务？ 2592
15.8 作业开发类 2592
15.8.1 如何准备 MRS 的数据源？ 2592
15.8.2 集群支持提交哪些形式的 Spark 作业？ 2593
15.8.3 MRS 集群的租户资源最小值改为 0 后，只能同时跑一个 Spark 任务吗？ 2593
15.8.4 Spark 作业 Client 模式和 Cluster 模式的区别 2593
15.8.5 如何查看 MRS 作业日志？ 2594
15.8.6 报错提示“当前用户在 MRS Manager 不存在，请先在 IAM 给予该用户足够的权限，再在概览页签进行 IAM 用户同步” 2594
15.8.7 LauncherJob 作业执行结果为 Failed. 报错信息为：jobPropertiesMap is null 2595
15.8.8 MRS Console 页面 Flink 作业状态与 Yarn 上的作业状态不一致 2595
15.8.9 提交长时作业 SparkStreaming，运行几十个小时后失败，报 OBS 访问 403 2595
15.8.10 ClickHouse 客户端执行 SQL 查询时报内存不足问题 2595
15.8.11 Spark 运行作业报错：java.io.IOException: Connection reset by peer 2596
15.8.12 Spark 作业访问 OBS 报错：requestId=4971883851071737250 2596
15.8.13 DataArts Studio 调度 spark 作业，偶现失败，重跑失败 2596
15.8.14 Flink 任务运行失败，报错：java.lang.NoSuchFieldError: SECURITY_SSL_ENCRYPT_ENABLED 2596
15.8.15 提交的 Yarn 作业在界面上查看不到 2597
15.8.16 如何修改现有集群的 HDFS NameSpace(fs.defaultFS) 2597
15.8.17 通过管控面提交 Flink 任务时 launcher-job 因 heap size 不够被 Yarn 结束 2597
15.8.18 Flink 作业提交时报错 slot request timeout 2597
15.8.19 DistCP 类型作业导入导出数据问题 2598
15.9 集群升级/补丁 2598
15.9.1 MRS 版本如何进行升级？ 2598
15.9.2 MRS 是否支持修改版本？ 2598
15.10 集群访问类 2598
15.10.1 MRS 登录集群节点的两种方式能够切换么？ 2598
15.10.2 如何获取 ZooKeeper 的 IP 地址和端口？ 2598
15.10.3 如何通过集群外的节点访问 MRS 集群？ 2599
15.11 大数据业务开发 2600
15.11.1 MRS 是否支持同时运行多个 Flume 任务？ 2600
15.11.2 如何修改 FlumeClient 的日志为标准输出日志？ 2600
15.11.3 Hadoop 组件 jar 包位置和环境变量的位置在哪里？ 2600
15.11.4 HBase 支持的压缩算法有哪些？ 2601
15.11.5 MRS 是否支持通过 Hive 的 HBase 外表将数据写入到 HBase？ 2601
15.11.6 如何查看 HBase 日志？ 2601
15.11.7 HBase 表如何设置和修改数据保留期？ 2601
15.11.8 HDFS 如何进行数据均衡？ 2601
15.11.9 如何修改 HDFS 的副本数？ 2601
15.11.10 如何使用 Python 远程连接 HDFS 的端口？ 2602
15.11.11 如何修改 HDFS 主备倒换类？ 2604
15.11.12 DynamoDB 的 number 在 Hive 表中用什么类型比较好？ 2605
15.11.13 Hive Driver 是否支持对接 dbcp2？ 2605

15.11.14 用户 A 如何查看用户 B 创建的 Hive 表? 2605
15.11.15 Hive 查询数据是否支持导出? 2606
15.11.16 Hive 使用 beeline -e 执行多条语句报错 2606
15.11.17 添加 Hive 服务后，提交 hivesql/hivescript 作业失败 2607
15.11.18 Hue 下载 excel 无法打开 2607
15.11.19 Hue 连接 hiveserver，不释放 session，报错 over max user connections 如何处理? 2608
15.11.20 如何重置 Kafka 数据? 2609
15.11.21 MRS Kafka 如何查看客户端版本信息? 2609
15.11.22 Kafka 目前支持的访问协议类型有哪些? 2609
15.11.23 消费 kafka topic，报错：Not Authorized to access group xxx 2609
15.11.24 Kudu 支持的压缩算法有哪些? 2609
15.11.25 如何查看 Kudu 日志? 2610
15.11.26 新建集群 Kudu 服务异常处理 2610
15.11.27 OpenTSDB 是否支持 python 的接口? 2611
15.11.28 Presto 如何配置其他数据源? 2611
15.11.29 MRS 如何连接 spark-shell 2612
15.11.30 MRS 如何连接 spark-beeline 2613
15.11.31 spark job 对应的执行日志保存在哪里? 2613
15.11.32 MRS 的 Storm 集群提交任务时如何指定日志路径? 2613
15.11.33 Yarn 的 ResourceManager 配置是否正常? 2613
15.11.34 如何修改 Clickhouse 服务的 allow_drop_detached 配置项? 2615
15.11.35 执行 Spark 任务报内存不足告警 2616
15.11.36 ClickHouse 占用大量 CPU，一直不下降 2616
15.11.37 ClickHouse 如何开启 Map 类型? 2616
15.11.38 SparkSQL 访问 hive 分区表大量调用 OBS 接口 2617
15.12 API 使用类 2617
15.12.1 使用调整集群节点接口时参数 node_id 如何配置? 2618
15.13 集群管理类 2618
15.13.1 如何查看所有集群? 2618
15.13.2 如何查看日志信息? 2618
15.13.3 如何查看集群配置信息? 2619
15.13.4 如何在 MRS 集群中安装 Kafka，Flume 组件? 2619
15.13.5 如何停止 MRS 集群? 2619
15.13.6 MRS 支持数据盘扩容吗? 2619
15.13.7 现有集群如何增加组件? 2619
15.13.8 MRS 集群中安装的组件能否删除? 2619
15.13.9 MRS 是否支持变更 MRS 集群节点? 2619
15.13.10 如何取消集群风险告警 2619
15.13.11 为什么 MRS 集群显示的资源池内存小于实际集群内存? 2620
15.13.12 如何配置 knox 内存? 2620
15.13.13 MRS 集群安装的 Python 版本是多少? 2620
15.13.14 如何查看各组件配置文件路径? 2620

15.13.15 MRS 节点时间不正确.....2621
15.13.16 如何查询 MRS 节点的启动时间.....2622
15.13.17 节点互信异常如何处理？.....2622
15.13.18 如何调整 manager-executor 进程内存？.....2623
15.14 Kerberos 使用.....2624
15.14.1 已创建的 MRS 集群如何修改 Kerberos 状态？.....2624
15.14.2 Kerberos 认证服务的端口有哪些？.....2624
15.14.3 如何在运行中的集群中部署 Kerberos 服务？.....2624
15.14.4 开启 Kerberos 认证的集群如何访问 Hive？.....2624
15.14.5 开启 Kerberos 认证的集群如何访问 Presto？.....2625
15.14.6 开启 Kerberos 认证的集群如何访问 Spark？.....2626
15.14.7 如何避免 Kerberos 认证过期？.....2627
15.15 元数据管理.....2627
15.15.1 Hive 元数据在哪里查看？.....2627
16 故障排除.....2629
16.1 Web 页面访问类.....2629
16.1.1 无法访问 MRS 集群管理页面（MRS Manager 界面）.....2629
16.1.2 升级 Python 后，无法登录 MRS Manager 页面.....2630
16.1.3 用户修改域名后无法登录 MRS Manager 页面.....2630
16.1.4 登录 Manager，页面空白不显示.....2632
16.1.5 用户名过长时下载认证凭据失败.....2632
16.2 集群管理类.....2633
16.2.1 缩容 Task 节点失败.....2633
16.2.2 MRS 集群添加新磁盘.....2634
16.2.3 MRS 集群更换磁盘（适用于 2.x 及之前）.....2637
16.2.4 MRS 集群更换磁盘（适用于 3.x）.....2640
16.2.5 MRS 备份失败.....2642
16.2.6 Core 节点出现 df 显示的容量和 du 显示的容量不一致.....2643
16.2.7 如何解除关联子网.....2644
16.2.8 修改 hostname，导致 MRS 状态异常.....2644
16.2.9 如何定位进程被 kill.....2645
16.2.10 MRS 集群使用 pip3 安装 python 包提示网络不可达.....2647
16.2.11 MRS 集群客户端无法下载.....2647
16.2.12 扩容失败.....2648
16.2.13 MRS 通过 beeline 执行插入命令的时候出错.....2649
16.2.14 MRS 集群如何进行 Euleros 系统漏洞升级？.....2650
16.2.15 使用 CDM 迁移数据至 HDFS.....2652
16.2.16 MRS 集群频繁产生告警.....2653
16.2.17 PMS 进程占用内存高问题处理.....2655
16.2.18 Knox 进程占用内存高.....2656
16.2.19 安全集群外节点安装客户端访问 HBase 很慢.....2657
16.2.20 作业无法提交如何定位？.....2658

16.2.21 HBase 日志文件过大导致 OS 盘空间不足.....2661
16.2.22 Manager 页面新建的租户删除失败.....2662
16.3 使用 Alluixo.....2662
16.3.1 Alluixo 在 HA 模式下出现 Does not contain a valid host:port authority 报错.....2662
16.4 使用 ClickHouse.....2663
16.4.1 ZooKeeper 上数据错乱导致 ClickHouse 启动失败问题.....2663
16.5 使用 DBservice.....2665
16.5.1 DBServer 实例状态异常.....2665
16.5.2 DBServer 实例一直处于 Restoring 状态.....2666
16.5.3 默认端口 20050 或 20051 被占用.....2667
16.5.4 /tmp 目录权限不对导致 DBserver 实例状态一直处于 Restoring.....2667
16.5.5 DBService 备份失败.....2669
16.5.6 DBService 状态正常，组件无法连接 DBService.....2669
16.5.7 DBServer 启动失败.....2670
16.5.8 浮动 IP 不通导致 DBService 备份失败.....2671
16.5.9 DBService 配置文件丢失导致启动失败.....2672
16.6 使用 Flink.....2674
16.6.1 安装客户端执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file :"security.kerberos.login.keytab".....2674
16.6.2 安装客户端修改配置后执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file.....2675
16.6.3 创建 Flink 集群时执行 yarn-session.sh 命令失败.....2676
16.6.4 使用不同用户，执行 yarn-session 创建集群失败.....2677
16.6.5 Flink 业务程序无法读取 NFS 盘上的文件.....2678
16.6.6 自定义 Flink log4j 日志输出级别.....2679
16.7 使用 Flume.....2679
16.7.1 Flume 向 Spark Streaming 提交作业，提交到集群后报类找不到.....2679
16.7.2 Flume 客户端安装失败.....2680
16.7.3 Flume 客户端无法连接服务端.....2681
16.7.4 Flume 数据写入组件失败.....2681
16.7.5 Flume 服务端进程故障.....2682
16.7.6 Flume 数据采集慢.....2682
16.7.7 Flume 启动失败.....2683
16.8 使用 HBase.....2684
16.8.1 连接到 HBase 响应慢.....2684
16.8.2 HBase 用户认证失败.....2685
16.8.3 端口被占用导致 RegionServer 启动失败.....2685
16.8.4 节点剩余内存不足导致 HBase 启动失败.....2686
16.8.5 HDFS 性能差导致 HBase 服务不可用告警.....2686
16.8.6 参数不合理导致 HBase 启动失败.....2687
16.8.7 残留进程导致 Regionsever 启动失败.....2688
16.8.8 HDFS 上设置配额导致 HBase 启动失败.....2688
16.8.9 HBase version 文件损坏导致启动失败.....2689

16.8.10 无业务情况下，RegionServer 占用 CPU 高.....2690
16.8.11 HBase 启动失败，RegionServer 日志中提示 FileNotFoundException 异常.....2691
16.8.12 HBase 启动后原生页面显示 RegionServer 个数多于实际个数.....2692
16.8.13 RegionServer 实例异常，处于 Restoring 状态.....2693
16.8.14 新安装的集群 HBase 启动失败.....2694
16.8.15 acl 表目录丢失导致 HBase 启动失败.....2694
16.8.16 集群上下电之后 HBase 启动失败.....2695
16.8.17 文件块过大导致 HBase 数据导入失败.....2697
16.8.18 使用 Phoenix 创建 HBase 表后，向索引表中加载数据报错.....2697
16.8.19 在 MRS 集群客户端无法执行 **hbase shell** 命令.....2699
16.8.20 HBase shell 客户端在使用中有 INFO 信息打印在控制台导致显示混乱.....2699
16.8.21 RegionServer 剩余内存不足导致 HBase 服务启动失败.....2700
16.9 使用 HDFS.....2701
16.9.1 修改集群 HDFS 服务的 NameNode RPC 端口后，NameNode 都变为备状态.....2701
16.9.2 通过公网 IP 连接主机，使用 HDFS 客户端报错.....2702
16.9.3 使用 Python 远程连接 HDFS 的端口失败.....2703
16.9.4 HDFS 容量使用达到 100%，导致上层服务 HBase、Spark 等上报服务不可用.....2703
16.9.5 启动 HDFS 和 Yarn 报错.....2704
16.9.6 HDFS 权限设置问题.....2705
16.9.7 HDFS 的 DataNode 一直显示退服中.....2706
16.9.8 内存不足导致 HDFS 启动失败.....2708
16.9.9 ntpdate 修改时间导致 HDFS 出现大量丢块.....2710
16.9.10 DataNode 概率性出现 CPU 占用接近 100%，导致节点丢失（ssh 连得很慢或者连不上）.....2712
16.9.11 单 NameNode 长期故障，如何使用客户端手动 checkpoint.....2713
16.9.12 文件读写常见故障.....2714
16.9.13 文件最大打开句柄数设置太小导致读写文件异常.....2715
16.9.14 客户端写文件 close 失败.....2716
16.9.15 文件错误导致上传文件到 HDFS 失败.....2718
16.9.16 界面配置 dfs.blocksize 后 put 数据，block 大小还是原来的大小.....2718
16.9.17 读取文件失败，FileNotFoundException.....2719
16.9.18 HDFS 写文件失败，item limit of / is exceeded.....2720
16.9.19 调整 shell 客户端日志级别.....2720
16.9.20 读文件失败 No common protection layer.....2720
16.9.21 HDFS 目录配额（quota）不足导致写文件失败.....2721
16.9.22 执行 balance 失败，Source and target differ in block-size.....2722
16.9.23 查询或者删除文件失败，父目录可以看见此文件（不可见字符）.....2723
16.9.24 非 HDFS 数据残留导致数据分布不均衡.....2724
16.9.25 客户端安装在数据节点导致数据分布不均衡.....2725
16.9.26 节点内 DataNode 磁盘使用率不均衡处理指导.....2725
16.9.27 执行 balance 常见问题定位方法.....2726
16.9.28 HDFS 显示磁盘空间不足，其实还有 10%磁盘空间.....2727
16.9.29 普通集群在 Core 节点安装 hdfs 客户端，使用时报错.....2727

16.9.30 集群外节点安装客户端使用 hdfs 上传文件失败 2728
16.9.31 HDFS 写并发较大时，报副本不足的问题 2729
16.9.32 HDFS 客户端无法删除超长目录 2729
16.9.33 集群外节点访问 MRS HDFS 报错 2731
16.10 使用 Hive 2732
16.10.1 Hive 各个日志里都存放了什么信息？ 2732
16.10.2 Hive 启动失败问题的原因有哪些？ 2733
16.10.3 安全集群执行 set 命令的时候报 Cannot modify xxx at runtime 2733
16.10.4 怎样在 Hive 提交任务的时候指定队列？ 2734
16.10.5 客户端怎么设置 Map/Reduce 内存？ 2735
16.10.6 如何在导入表时指定输出的文件压缩格式 2735
16.10.7 desc 描述表过长时，无法显示完整 2736
16.10.8 增加分区列后再 insert 数据显示为 NULL 2737
16.10.9 创建新用户，执行查询时报无权限 2737
16.10.10 执行 SQL 提交任务到指定队列报错 2739
16.10.11 执行 load data inpath 命令报错 2739
16.10.12 执行 load data local inpath 命令报错 2740
16.10.13 执行 create external table 报错 2741
16.10.14 在 beeline 客户端执行 dfs -put 命令报错 2741
16.10.15 执行 set role admin 报无权限 2742
16.10.16 通过 beeline 创建 UDF 时候报错 2743
16.10.17 Hive 服务健康状态和 Hive 实例健康状态的区别 2743
16.10.18 Hive 中的告警有哪些以及触发的场景 2744
16.10.19 Shell 客户端连接提示"authentication failed" 2745
16.10.20 客户端提示访问 ZooKeeper 失败 2745
16.10.21 使用 udf 函数提示"Invalid function" 2747
16.10.22 Hive 服务状态为 Unknown 总结 2747
16.10.23 Hiveserver 或者 Metastore 实例的健康状态为 unknown 2747
16.10.24 Hiveserver 或者 Metastore 实例的健康状态为 Concerning 2748
16.10.25 TEXTFILE 类型文件使用 ARC4 压缩时 select 结果乱码 2748
16.10.26 hive 任务运行过程中失败，重试成功 2749
16.10.27 执行 select 语句报错 2749
16.10.28 drop partition 操作，有大量分区时操作失败 2751
16.10.29 localtask 启动失败 2751
16.10.30 WebHCat 启动失败 2752
16.10.31 切域后 Hive 二次开发样例代码报错 2753
16.10.32 DBService 超过最大连接数，导致 metastore 异常 2754
16.10.33 beeline 报 Failed to execute session hooks: over max connections 错误 2755
16.10.34 beeline 报 OutOfMemoryError 错误 2756
16.10.35 输入文件数超出设置限制导致任务执行失败 2757
16.10.36 任务执行中报栈内存溢出导致任务执行失败 2758
16.10.37 对同一张表或分区并发写数据导致任务失败 2759

16.10.38 Hive 任务失败，报没有 HDFS 目录的权限 2759
16.10.39 Load 数据到 Hive 表失败 2760
16.10.40 HiveServer 和 HiveHCat 进程故障 2761
16.10.41 Hive 执行 insert into 语句报错，命令界面报错信息不明 2762
16.10.42 增加 Hive 表字段超时 2763
16.10.43 Hive 服务重启失败 2765
16.10.44 hive 执行删除表失败 2766
16.10.45 Hive 执行 msck repair table table_name 报错 2767
16.10.46 在 Hive 中 drop 表后，如何完全释放磁盘空间 2767
16.10.47 客户端执行 SQL 报错连接超时 2768
16.10.48 WebHCat 健康状态异常导致启动失败 2769
16.10.49 mapred-default.xml 文件解析异常导致 WebHCat 启动失败 2770
16.11 使用 Hue 2770
16.11.1 Hue 上有 job 在运行 2770
16.11.2 使用 IE 浏览器在 Hue 中执行 HQL 失败 2771
16.11.3 Hue（主）无法打开 web 网页 2771
16.11.4 Hue WebUI 访问失败 2772
16.11.5 Hue 界面无法加载 HBase 表 2772
16.12 使用 Impala 2773
16.12.1 用户连接 impala-shell 失败 2773
16.12.2 创建 Kudu 表报错 2774
16.12.3 Impala 客户端登录失败 2775
16.13 使用 Kafka 2777
16.13.1 运行 Kafka 获取 topic 报错 2777
16.13.2 Flume 可以正常连接 Kafka，但是发送消息失败。 2778
16.13.3 Producer 发送数据失败，抛出 NullPointerException 2779
16.13.4 Producer 发送数据失败，抛出 TOPIC_AUTHORIZATION_FAILED 2781
16.13.5 Producer 偶现发送数据失败，日志提示 Too many open files in system 2783
16.13.6 Consumer 初始化成功，但是无法从 Kafka 中获取指定 Topic 消息 2785
16.13.7 Consumer 消费数据失败，Consumer 一直处于等待状态 2789
16.13.8 SparkStreaming 消费 Kafka 消息失败，提示 Error getting partition metadata 2791
16.13.9 新建集群 Consumer 消费数据失败，提示 GROUP_COORDINATOR_NOT_AVAILABLE 2793
16.13.10 SparkStreaming 消费 Kafka 消息失败，提示 Couldn't find leader offsets 2794
16.13.11 Consumer 消费数据失败，提示 SchemaException: Error reading field 'brokers' 2796
16.13.12 Consumer 消费数据是否丢失排查 2797
16.13.13 帐号锁定导致启动组件失败 2797
16.13.14 Kafka Broker 上报进程异常，日志提示 IllegalArgumentException 2798
16.13.15 执行 Kafka Topic 删除操作，发现无法删除 2799
16.13.16 执行 Kafka Topic 删除操作，提示 AdminOperationException 2801
16.13.17 执行 Kafka Topic 创建操作，发现无法创建提示 NoAuthException 2802
16.13.18 执行 Kafka Topic 设置 ACL 操作失败，提示 NoAuthException 2804
16.13.19 执行 Kafka Topic 创建操作，发现无法创建提示 NoNode for /brokers/ids 2806

16.13.20 执行 Kakfa Topic 创建操作，发现无法创建提示 replication factor larger than available brokers2807
16.13.21 Consumer 消费数据存在重复消费现象 2808
16.13.22 执行 Kafka Topic 创建操作，发现 Partition 的 Leader 显示为 none 2810
16.13.23 Kafka 安全使用说明 2811
16.13.24 如何获取 Kafka Consumer Offset 信息 2815
16.13.25 如何针对 Topic 进行配置增加和删除 2817
16.13.26 如何读取“__consumer_offsets”内部 topic 的内容 2818
16.13.27 如何配置客户端 shell 命令的日志 2819
16.13.28 如何获取 Topic 的分布信息 2820
16.13.29 Kafka 高可靠使用说明 2821
16.13.30 Kafka 生产者写入单条记录过长问题 2824
16.13.31 Kakfa 消费者读取单条记录过长问题 2824
16.13.32 Kafka 集群节点内多磁盘数据量占用高处理办法 2825
16.14 使用 Oozie 2828
16.14.1 当并发提交大量 oozie 任务时，任务一直没有运行 2828
16.15 使用 Presto 2828
16.15.1 配置 sql-standard-with-group 创建 schema 失败报 Access Denied 2829
16.15.2 Presto 的 coordinator 无法正常启动 2830
16.15.3 Presto 查询 Kudu 表报错 2831
16.15.4 Presto 查询 Hive 表无数据 2832
16.16 使用 Spark 2833
16.16.1 Spark 应用下修改 split 值时报错 2833
16.16.2 使用 Spark 时报错 2834
16.16.3 引入 jar 包不正确，导致 Spark 任务无法运行 2834
16.16.4 Spark 任务由于内存不够，作业卡住 2835
16.16.5 运行 Spark 报错 2836
16.16.6 Driver 端提示 executor memory 超限 2837
16.16.7 Yarn-cluster 模式下，Can't get the Kerberos realm 异常 2838
16.16.8 JDK 版本不匹配启动 spark-sql，spark-shell 失败 2840
16.16.9 Yarn-client 模式提交 ApplicationMaster 尝试启动两次失败 2840
16.16.10 提交 Spark 任务时，连接 ResourceManager 异常 2841
16.16.11 DataArts Studio 调度 spark 作业失败 2842
16.16.12 Spark 作业 api 提交状态为 error 2843
16.16.13 集群反复出现 43006 告警 2844
16.16.14 在 spark-beeline 中创建/删除表失败 2844
16.16.15 集群外节点提交 Spark 作业到 Yarn 报错连不上 Driver 2846
16.16.16 运行 Spark 任务发现大量 shuffle 结果丢失 2847
16.16.17 JDBCServer 长时间运行导致磁盘空间不足 2848
16.16.18 spark-shell 执行 sql 跨文件系统 load 数据到 hive 表失败 2849
16.16.19 Spark 任务提交失败 2849
16.16.20 Spark 任务运行失败 2850
16.16.21 JDBCServer 连接失败 2851

16.16.22 查看 Spark 任务日志失败 2851
16.16.23 Spark 连接其他服务认证问题 2852
16.16.24 spark 连接 redis 报错 2852
16.16.25 spark-beeline 查询 Hive 视图报错 2854
16.17 使用 Sqoop 2855
16.17.1 Sqoop 如何连接 mysql 2855
16.17.2 Sqoop 读取 MySQL 中数据到 HBase 报 HBaseAdmin.<init>方法找不到异常 2856
16.17.3 HUE 界面的 Sqoop 任务 HBase 到 HDFS 报错 2857
16.17.4 Sqoop 从 hive 到 mysql8.0 报格式错误 2860
16.17.5 Sqoop import 从 pg 到 hive 报错 2862
16.17.6 Sqoop 读 mysql，写 parquet 文件到 OBS 失败 2862
16.18 使用 Storm 2863
16.18.1 Storm 组件的 Storm UI 页面中 events 超链接地址无效 2863
16.18.2 提交拓扑失败 2864
16.18.3 提交拓扑失败，提示 Failed to check principle for keytab 2866
16.18.4 提交拓扑后 Worker 日志为空 2867
16.18.5 提交拓扑后 Worker 运行异常，日志提示 Failed to bind to：host:ip 2869
16.18.6 使用 jstack 命令查看进程堆栈提示 well-known file is not secure 2870
16.18.7 使用 Storm-JDBC 插件开发 Oracle 写入 Bolt，发现数据无法写入 2873
16.18.8 业务拓扑配置 GC 参数不生效 2874
16.18.9 UI 查看信息显示 Internal Server Error 2876
16.19 使用 Ranger 2876
16.19.1 Hive 启用 Ranger 鉴权后，在 Hue 页面能查看到没有权限的表和库 2877
16.20 使用 Yarn 2878
16.20.1 启动 Yarn 后发现一堆 job 2878
16.20.2 通过客户端 hadoop jar 命令提交任务，客户端返回 GC overhead 2879
16.20.3 Yarn 汇聚日志过大导致磁盘被占满 2880
16.20.4 MR 任务异常临时文件不删除 2881
16.20.5 提交任务的 Yarn 的 ResourceManager 报错 connection refused，且配置的 Yarn 端口为 8032 2882
16.20.6 Yarn WebUI 作业查看日志提示“Could not access logs page!” 2883
16.20.7 Yarn 页面单击队列名称报错 2884
16.21 使用 ZooKeeper 2884
16.21.1 MRS 集群如何访问 ZooKeeper 2884
16.22 访问 OBS 2885
16.22.1 使用 MRS 多用户访问 OBS 功能时/tmp 目录没有权限 2885
16.22.2 Hadoop 客户端删除 OBS 上数据时.Trash 目录没有权限 2886
17 附录 2888
17.1 MRS 3.x 版本操作注意事项 2888

1 简介

1.1 什么是 MRS

大数据是人类进入互联网时代以来面临的一个巨大问题：社会生产生活产生的数据量越来越大，数据种类越来越多，数据产生的速度越来越快。传统的数据处理技术，比如说单机存储，关系数据库已经无法解决这些新的大数据问题。为解决以上大数据处理问题，Apache基金会推出了Hadoop大数据处理的开源解决方案。Hadoop是一个开源分布式计算平台，可以充分利用集群的计算和存储能力，完成海量数据的处理。企业自行部署Hadoop系统有成本高，周期长，难运维和不灵活等问题。

针对上述问题，云提供了大数据MapReduce服务（MRS），MRS是一个在云上部署和管理Hadoop系统的服务，一键即可部署Hadoop集群。MRS提供租户完全可控的一站式企业级大数据集群云服务，完全兼容开源接口，结合云计算、存储优势及大数据行业经验，为客户提供高性能、低成本、灵活易用的全栈大数据平台，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件，并具备在后续根据业务需要进行定制开发的能力，帮助企业快速构建海量数据信息处理系统，并通过对海量信息数据实时与非实时的分析挖掘，发现全新价值点和企业商机。

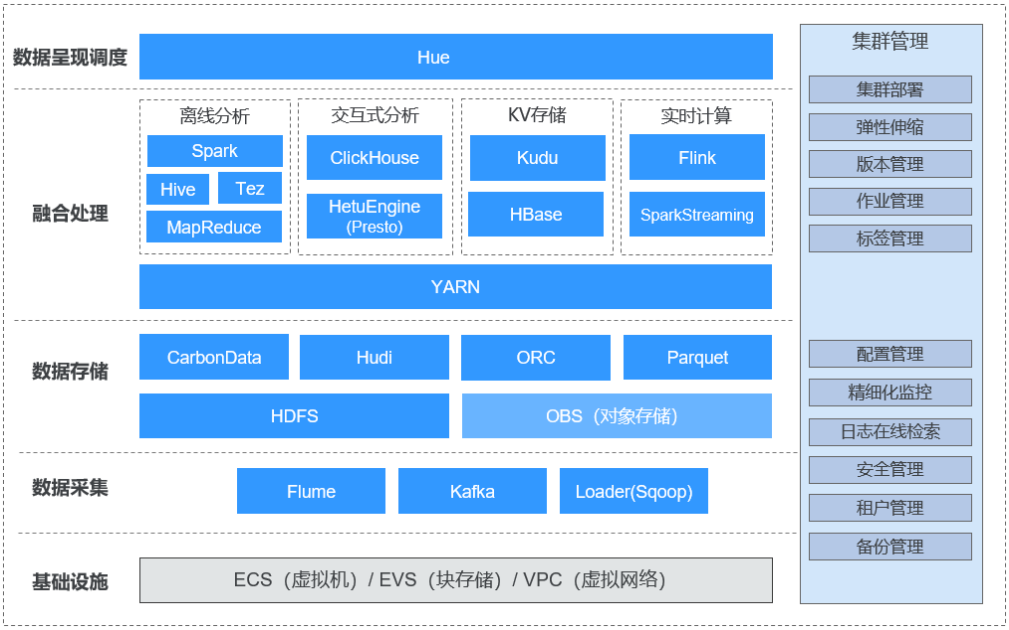
产品架构

MRS逻辑架构如图1-1所示。

说明

MRS 3.x及之后版本暂不支持在管理控制台执行补丁管理操作。

图 1-1 MRS 架构



MRS架构包括了基础设施和大数据处理流程各个阶段的能力。

- 基础设施

  MRS基于弹性云服务器ECS构建的大数据集群，充分利用了其虚拟化层的高可靠、高安全的能力。

  - 虚拟私有云（VPC）为每个租户提供的虚拟内部网络，默认与其他网络隔离。
  - 云硬盘（EVS）提供高可靠、高性能的存储。
  - 弹性云服务器（ECS）提供的弹性可扩展虚拟机，结合VPC、安全组、EVS数据多副本等能力打造一个高效、可靠、安全的计算环境。

- 数据集成

  数据集成层提供了数据接入到MRS集群的能力，包括Flume（数据采集）、Loader（关系型数据导入）、Kafka（高可靠消息队列），支持各种数据源导入数据到大数据集群中。

- 数据存储

  MRS支持结构化和非结构化数据在集群中的存储，并且支持多种高效的格式来满足不同计算引擎的要求。

  - HDFS是大数据上通用的分布式文件系统。
  - OBS是对象存储服务，具有高可用低成本的特点。
  - HBase支持带索引的数据存储，适合高性能基于索引查询的场景。

- 数据计算

  MRS提供多种主流计算引擎：MapReduce（批处理）、Tez（DAG模型）、Spark（内存计算）、SparkStreaming（微批流计算）、Storm（流计算）、Flink（流计算），满足多种大数据应用场景，将数据进行结构和逻辑的转换，转化成满足业务目标的数据模型。

- 数据分析

基于预设的数据模型，使用易用SQL的数据分析，用户可以选择Hive（数据仓库），SparkSQL以及Presto交互式查询引擎。

- 数据呈现调度

  用于数据分析结果的呈现，并与数据湖工厂（DLF）集成，提供一站式的大数据协同开发平台，帮助用户轻松完成数据建模、数据集成、脚本开发、作业调度、运维监控等多项任务，可以极大降低用户使用大数据的门槛，帮助用户快速构建大数据处理中心。

- 集群管理

  以Hadoop为基础的大数据生态的各种组件均是以分布式的方式进行部署，其部署、管理和运维复杂度较高。

  MRS集群管理提供了统一的运维管理平台，包括一键式部署集群能力，并提供多版本选择，支持运行过程中集群在无业务中断条件下，进行扩缩容、弹性伸缩。同时MRS集群管理还提供了作业管理、资源标签管理，以及对上述数据处理各层组件的运维，并提供监控、告警、配置、补丁升级等一站式运维能力。

产品优势

MRS服务拥有强大的Hadoop内核团队，基于FusionInsight大数据企业级平台构筑。历经行业数万节点部署量的考验，提供多级用户SLA保障。

MRS具有如下优势：

- 高性能

  MRS支持自研的CarbonData存储技术。CarbonData是一种高性能大数据存储方案，以一份数据同时支持多种应用场景，并通过多级索引、字典编码、预聚合、动态Partition、准实时数据查询等特性提升了IO扫描和计算性能，实现万亿数据分析秒级响应。同时MRS支持自研增强型调度器Superior，突破单集群规模瓶颈，单集群调度能力超10000节点。

- 低成本

  基于多样化的云基础设施，提供了丰富的计算、存储设施的选择，同时计算存储分离，提供了低成本海量数据存储方案。MRS可以按业务峰谷，自动弹性伸缩，帮助客户节省大数据平台闲时资源。MRS集群可以用时再创建、用时再扩容，用完就可以销毁、缩容，确保成本最优。

- 高安全

  MRS服务拥有企业级的大数据多租户权限管理能力，拥有企业级的大数据安全管理特性，支持按照表/按列控制访问权限，支持数据按照表/按列加密。

- 易运维

  MRS提供可视化大数据集群管理平台，提高运维效率。并支持滚动补丁升级，可视化补丁发布信息，一键式补丁安装，无需人工干预，不停业务，保障用户集群长期稳定。

- 高可靠

  MRS服务经过大规模的可靠性、长稳验证，满足企业级高可靠要求，同时支持数据跨AZ/跨Region自动备份的数据容灾能力，自动反亲和技术，虚拟机分布在不同物理机上。

1.2 MRS 与自建 Hadoop 对比优势

MapReduce服务（MRS）提供租户完全可控的企业级大数据集群云服务，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件，用户无需关注硬件的购买和

维护。MRS服务拥有强大的Hadoop内核团队，基于FusionInsight大数据企业级平台构筑，历经行业数万节点部署量的考验，提供多级用户SLA保障。与自建Hadoop集群相比，MRS还具有以下优势：

1. **MRS支持一键式创建、删除、扩缩容集群，并通过弹性公网IP便携访问MRS集群管理系统，让大数据集群更加易于使用**。
   - 用户自建大数据集群面临成本高、周期长、运维难和不灵活等问题。针对这些问题，MRS支持一键式创建、删除、扩容和缩容集群的能力，用户可以自定制集群的类型，组件范围，各类型的节点数、虚拟机规格、可用区、VPC网络、认证信息，MRS将为用户自动创建一个符合配置的集群，全程无需用户参与。同时支持用户快速创建多应用场景集群，比如快速创建Hadoop分析集群、HBase集群、Kafka集群。MRS支持部署异构集群，在集群中存在不同规格的虚拟机，允许在CPU类型，硬盘容量，硬盘类型，内存大小灵活组合。
   - MRS提供了基于弹性公网IP来便捷访问组件WebUI的安全通道，并且比用户自己绑定弹性公网IP更便捷，只需界面鼠标操作，即可简化原先用户需要自己登录虚拟私有云添加安全组规则，获取公网IP等步骤，减少了用户操作步骤。
   - MRS提供了自定义引导操作，用户可以以此为入口灵活配置自己的集群，通过引导操作用户可以自动化地完成安装MRS还没支持的第三方软件，修改集群运行环境等自定义操作。
   - MRS支持WrapperFS特性，提供OBS的翻译能力，兼容HDFS到OBS的平滑迁移，解决客户将HDFS中的数据迁移到OBS后，即可实现客户端无需修改自己的业务代码逻辑的情况下，访问存储到OBS的数据。
2. **MRS支持自动弹性伸缩，相对自建Hadoop集群的使用成本更低**。

   MRS可以按业务峰谷，自动弹性伸缩，在业务繁忙时申请额外资源，业务不繁忙时释放闲置资源，让用户按需使用，帮助用户节省大数据平台闲时资源，尽可能的帮助用户降低使用成本，聚焦核心业务。

   在大数据应用，尤其是周期性的数据分析处理场景中，需要根据业务数据的周期变化，动态调整集群计算资源以满足业务需要。MRS的弹性伸缩规则功能支持根据集群负载对集群进行弹性伸缩。此外，如果数据量为周期有规律的变化，并且希望在数据量变化前提前完成集群的扩缩容，可以使用MRS的资源计划特性。MRS服务支持规则和时间计划两种弹性伸缩的策略：
   - 弹性伸缩规则：根据集群实时负载对Task节点数量进行调整，数据量变化后触发扩缩容，有一定的延后性。
   - 资源计划：若数据量变化存在周期性规律，则可通过资源计划在数据量变化前提前完成集群的扩缩容，避免出现增加或减少资源的延后。

   弹性伸缩规则与资源计划均可触发弹性伸缩，两者即可同时配置也可单独配置。资源计划与基于负载的弹性伸缩规则叠加使用可以使得集群节点的弹性更好，足以应对偶尔超出预期的数据峰值出现。
3. **MRS支持存算分离，大幅提升大数据集群资源利用率**。

   针对传统存算一体大数据架构中扩容困难、资源利用率低等问题，MRS采用计算存储分离架构，存储基于公有云对象存储实现11个9的高可靠，无限容量，支撑企业数据量持续增长；计算资源支持0~N弹性扩缩，百节点快速发放。存算分离后，计算节点可实现真正的弹性伸缩；数据存储部分基于OBS的跨AZ等能力实现更高可靠性，无需担心地震、挖断光纤等突发事件。存储和计算资源可以灵活配置，根据业务需要各自独立进行弹性扩展，可使资源匹配精准化、合理化，让大数据集群资源利用率大幅提升，综合分析成本降低50%。

同时通过高性能的计算存储分离架构，打破存算一体架构并行计算的限制，最大化发挥对象存储的高带宽、高并发的特点，对数据访问效率和并行计算深度优化（元数据操作、写入算法优化等），实现性能提升。

4. **MRS支持自研CarbonData和自研超级调度器Superior Scheduler，性能更优。**
   - MRS支持自研的CarbonData存储技术。CarbonData是一种高性能大数据存储方案，以一份数据同时支持多种应用场景，并通过多级索引、字典编码、预聚合、动态Partition、准实时数据查询等特性提升了IO扫描和计算性能，实现万亿数据分析秒级响应。
   - MRS支持自研超级调度器Superior Scheduler，突破单集群规模瓶颈，单集群调度能力超10000节点。Superior Scheduler是一个专门为Hadoop YARN分布式资源管理系统设计的调度引擎，是针对企业客户融合资源池，多租户的业务诉求而设计的高性能企业级调度器。Superior Scheduler可实现开源调度器、Fair Scheduler以及Capacity Scheduler的所有功能。另外，相较于开源调度器，Superior Scheduler在企业级多租户调度策略、租户内多用户资源隔离和共享、调度性能、系统资源利用率和支持大集群扩展性方面都做了针对性的增强，让Superior Scheduler直接替代开源调度器。
5. **MRS基于鲲鹏处理器进行软硬件垂直优化，充分释放硬件算力，实现高性价比。**

   MRS支持自研鲲鹏服务器，充分利用鲲鹏多核高并发能力，提供芯片级的全栈自主优化能力，使用自研的操作系统EulerOS、JDK及数据加速层，充分释放硬件算力，为大数据计算提供高算力输出。在性能相当情况下，端到端的大数据解决方案成本下降30%。
6. **MRS支持多种隔离模式及企业级的大数据多租户权限管理能力，安全性更高。**
   - MRS服务支持资源专属区内部署，专属区内物理资源隔离，用户可以在专属区内灵活地组合计算存储资源，包括专属计算资源+共享存储资源、共享计算资源+专属存储资源、专属计算资源+专属存储资源。MRS集群内支持逻辑多租户，通过权限隔离，对集群的计算、存储、表格等资源按租户划分。
   - MRS支持Kerberos安全认证，实现了基于角色的安全控制及完善的审计功能。
   - MRS支持对接云审计服务（CTS），为用户提供MRS资源操作请求及请求结果的操作记录，供用户查询、审计和回溯使用。支持所有集群操作审计，所有用户行为可溯源。
   - MRS支持与主机安全服务对接，针对主机安全服务，做过兼容性测试，保证功能和性能不受影响的情况下，增强服务的安全能力。
   - MRS支持基于WebUI的统一的用户登录能力，Manager自带用户认证环节，用户只有通过Manager认证才能正常访问集群。
   - MRS支持数据存储加密，所有用户帐号密码加密存储，数据通道加密传输，服务模块跨信任区的数据访问支持双向证书认证等能力。
   - MRS大数据集群提供了完整的企业级大数据多租户解决方案。多租户是MRS大数据集群中的多个资源集合（每个资源集合是一个租户），具有分配和调度资源（资源包括计算资源和存储资源）的能力。多租户将大数据集群的资源隔离成一个个资源集合，彼此互不干扰，用户通过“租用”需要的资源集合，来运行应用和作业，并存放数据。在大数据集群上可以存在多个资源集合来支持多个用户的不同需求。
   - MRS支持细粒度权限管理，结合云IAM服务提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对MRS服务，管理员能够控制IAM用户仅能对集群进行指定的管理操作。如不允许某用户组删除集群，仅允许操作MRS集群基本操作，如创建集群、查询集群列表等。同时MRS支持多租户对OBS存储的细粒度权限管理，根据

多种用户角色来区分访问OBS桶及其内部的对象的权限，实现MRS用户对OBS桶下的目录权限控制。

- MRS支持企业项目管理。企业项目是一种云资源管理方式，企业管理（Enterprise Management）提供面向企业客户的云上资源管理、人员管理、权限管理、财务管理等综合管理服务。区别于管理控制台独立操控、配置云产品的方式，企业管理控制台以面向企业资源管理为出发点，帮助企业以公司、部门、项目等分级管理方式实现企业云上的人员、资源、权限、财务的管理。MRS支持已开通企业项目服务的用户在创建集群时为集群配置对应的项目，然后使用企业项目管理对MRS上的的资源进行分组管理。此特性适用于客户针对多个资源进行分组管理，并对相应的企业项目进行诸如权限控制、分项目费用查看等操作的场景。

7. **MRS管理节点均实现HA，支持完备的可靠性机制，让系统更加可靠。**

   MRS在基于Apache Hadoop开源软件的基础上，在主要业务部件的可靠性方面进行了优化和提升。

   - 管理节点均实现HA

     Hadoop开源版本的数据、计算节点已经是按照分布式系统进行设计的，单节点故障不影响系统整体运行；而以集中模式运作的管理节点可能出现的单点故障，就成为整个系统可靠性的短板。

     MRS对所有业务组件的管理节点都提供了类似的双机的机制，包括Manager、Presto、HDFS NameNode、Hive Server、HBase HMaster、YARN Resources Manager、Kerberos Server、Ldap Server等，全部采用主备或负荷分担配置，有效避免了单点故障场景对系统可靠性的影响。

   - 完备的可靠性机制

     通过可靠性分析方法，梳理软件、硬件异常场景下的处理措施，提升系统的可靠性。

     - 保障意外掉电时的数据可靠性，不论是单节点意外掉电，还是整个集群意外断电，恢复供电后系统能够正常恢复业务，除非硬盘介质损坏，否则关键数据不会丢失。
     - 硬盘亚健康检测和故障处理，对业务不造成实际影响。
     - 自动处理文件系统的故障，自动恢复受影响的业务。
     - 自动处理进程和节点的故障，自动恢复受影响的业务。
     - 自动处理网络故障，自动恢复受影响的业务。

8. **MRS提供统一的可视化大数据集群管理界面，让运维人员更加轻松。**

   - MRS提供统一的可视化大数据集群管理界面，包括服务启停、配置修改、健康检查等能力，并提供可视化、便捷的集群管理监控告警功能；支持一键式系统运行健康度巡检和审计，保障系统的正常运行，降低系统运维成本。
   - MRS联合消息通知服务(SMN)，在配置消息通知后，可以实时给用户发送MRS集群健康状态，用户可以通过手机短信或邮箱实时接收到MRS集群变更及组件告警信息，帮助用户轻松运维，实时监控，实时发送告警。
   - MRS支持滚动补丁升级，可视化补丁发布信息，一键式补丁安装，无需人工干预，不停业务，保障用户集群长期稳定。
   - MRS服务支持运维授权的功能，用户在使用MRS集群过程中，发生问题可以在MRS页面发起运维授权，由运维人员帮助客户快速定位问题，用户可以随时收回该授权。同时用户也可以在MRS页面发起日志共享，选择日志范围共享给运维人员，以便运维人员在不接触集群的情况下帮助定位问题。

- MRS支持将创建集群失败的日志转储到OBS，便于运维人员获取日志进行分析。

9. **MRS具有开放的生态，支持无缝对接周边服务，快速构建统一大数据平台。**
   - 以全栈大数据MRS服务为基础，企业可以一键式构筑数据接入、数据存储、数据分析和价值挖掘的统一大数据平台，并且与数据湖治理中心 DGC及数据可视化等服务对接，为客户轻松解决数据通道上云、大数据作业开发调度和数据展现的困难，使客户从复杂的大数据平台构建和专业大数据调优和维护中解脱出来，更加专注行业应用，使客户完成一份数据多业务场景使用的诉求。DGC是数据全生命周期一站式开发运营平台，提供数据集成、数据开发、数据治理、数据服务、数据可视化等功能。MRS数据支持连接DGC台，并基于可视化的图形开发界面、丰富的数据开发类型（脚本开发和作业开发）、全托管的作业调度和运维监控能力，内置行业数据处理pipeline，一键式开发，全流程可视化，支持多人在线协同开发，极大地降低了用户使用大数据的门槛，帮助用户快速构建大数据处理中心，对数据进行治理及开发调度，快速实现数据变现。
   - MRS服务100%兼容开源大数据生态，结合周边丰富的数据及应用迁移工具，能够帮助客户快速完成自建平台的平滑迁移，整个迁移过程可做到“代码0修改，业务0中断”。

1.3 应用场景

大数据在人们的生活中无处不在，在IoT、电子商务、金融、制造、医疗、能源和政府部门等行业均可以使用MRS服务进行大数据处理。

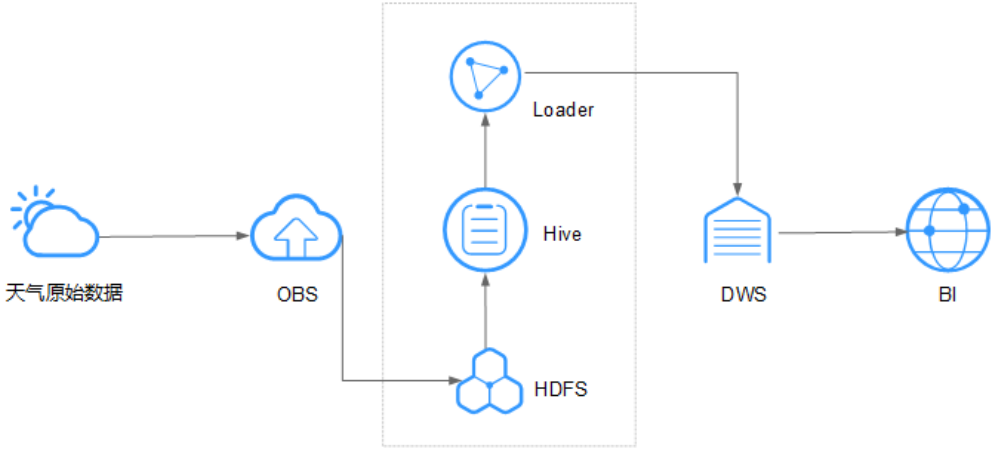
海量数据分析场景

海量数据分析是现代大数据系统中的主要场景。通常企业会包含多种数据源，接入后需要对数据进行ETL（Extract-Transform-Load）处理形成模型化数据，以便提供给各个业务模块进行分析梳理，这类业务通常有以下特点：

- 对执行实时性要求不高，作业执行时间在数十分钟到小时级别。
- 数据量巨大。
- 数据来源和格式多种多样。
- 数据处理通常由多个任务构成，对资源需要进行详细规划。

例如在环保行业中，可以将天气数据存储在OBS，定期转储到HDFS中进行批量分析，在1小时内MRS可以完成10TB的天气数据分析。

图 1-2 环保行业海量数据分析场景



该场景下MRS的优势如下所示。

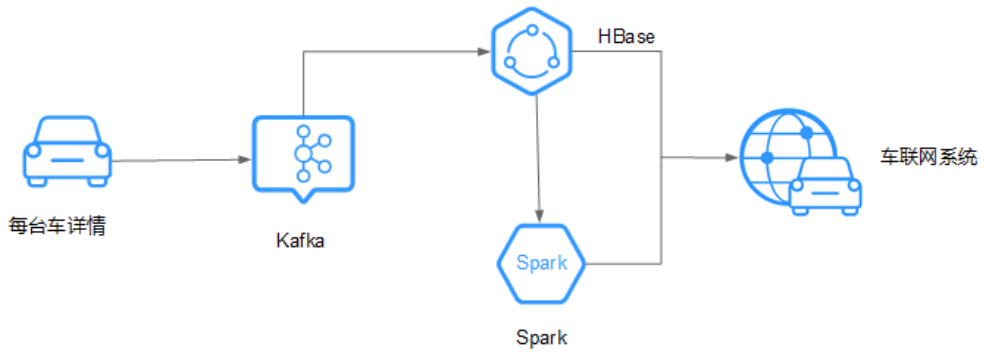
- 低成本：利用OBS实现低成本存储。
- 海量数据分析：利用Hive实现TB/PB级的数据分析。
- 可视化的导入导出工具：通过可视化导入导出工具Loader，将数据导出到DWS，完成BI分析。

海量数据存储场景

用户拥有大量结构化数据后，通常需要提供基于索引的准实时查询能力，如车联网场景下，根据汽车编号查询汽车维护信息，存储时，汽车信息会基于汽车编号进行索引，以实现该场景下的秒级响应。通常这类数据量比较庞大，用户可能保存1至3年的数据。

例如在车联网行业，某车企将数据储存在HBase中，以支持PB级别的数据存储和毫秒级的数据详单查询。

图 1-3 车联网行业海量数据存储场景



该场景下MRS的优势如下所示。

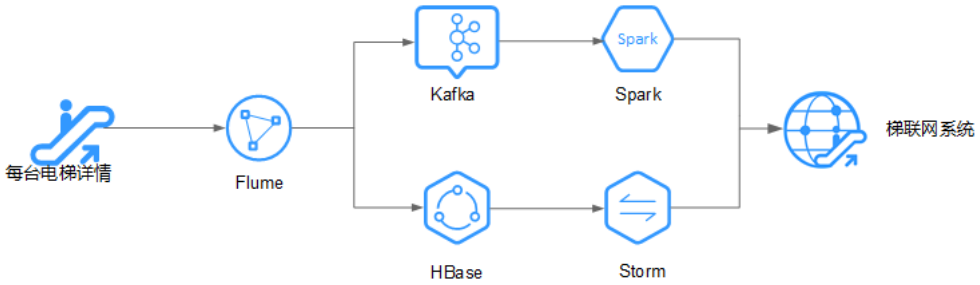
- 实时：利用Kafka实现海量汽车的消息实时接入。
- 海量数据存储：利用HBase实现海量数据存储，并实现毫秒级数据查询。
- 分布式数据查询：利用Spark实现海量数据的分析查询。

实时数据处理

实时数据处理通常用于异常检测、欺诈识别、基于规则告警、业务流程监控等场景，在数据输入系统的过程中，对数据进行处理。

例如在梯联网行业，智能电梯的数据，实时传入到MRS的流式集群中进行实时告警。

图 1-4 梯联网行业低时延流式处理场景



该场景下MRS的优势如下所示。

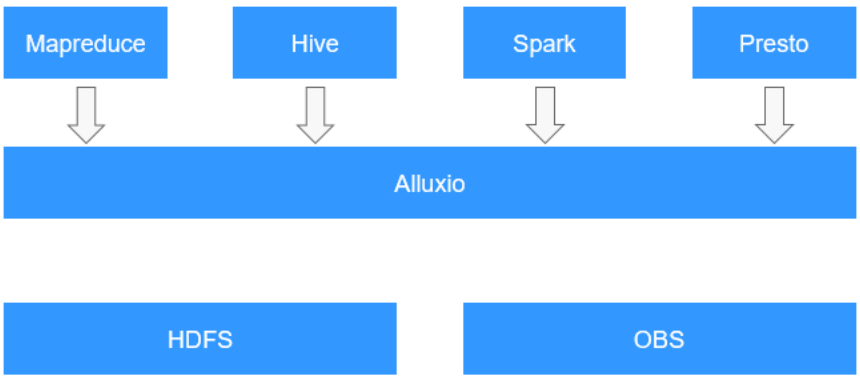
- 实时数据采集：利用Flume实现实时数据采集，并提供丰富的采集和存储连接方式。
- 海量的数据源接入：利用Kafka实现万级别的电梯数据的实时接入。

1.4 组件介绍

1.4.1 Alluxio

Alluxio是一个面向基于云的数据分析和人工智能的数据编排技术。在MRS的大数据生态系统中，Alluxio位于计算和存储之间，为包括Apache Spark、Presto、Mapreduce和Apache Hive的计算框架提供了数据抽象层，使上层的计算应用可以通过统一的客户端API和全局命名空间访问包括HDFS和OBS在内的持久化存储系统，从而实现了对计算和存储的分离。

图 1-5 Alluxio 架构



优势：

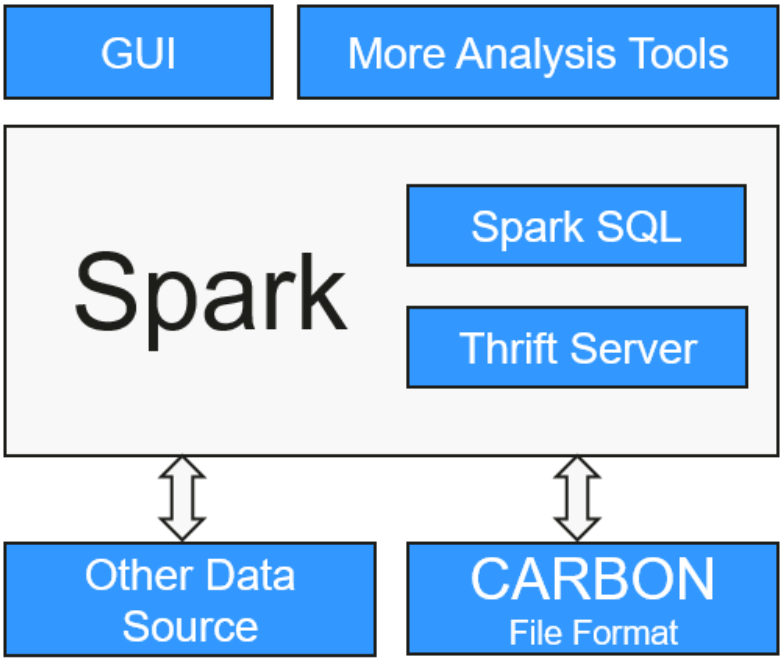
- 提供内存级 I/O 吞吐率，同时降低具有弹性扩张特性的数据驱动型应用的成本开销
- 简化云存储和对象存储接入
- 简化数据管理，提供对多数据源的单点访问
- 应用程序部署简易

有关Alluxio的详细信息，请参见：**https://docs.alluxio.io/os/user/stable/cn/Overview.html**。

1.4.2 CarbonData

CarbonData是一种新型的Apache Hadoop本地文件格式，使用先进的列式存储、索引、压缩和编码技术，以提高计算效率，有助于加速超过PB数量级的数据查询，可用于更快的交互查询。同时，CarbonData也是一种将数据源与Spark集成的高性能分析引擎。

图 1-6 CarbonData 基本架构



使用CarbonData的目的是对大数据即席查询提供超快速响应。CarbonData是一个OLAP引擎，采用类似于RDBMS中的表来存储数据。用户可将大量（10TB以上）的数据导入以CarbonData格式创建的表中，CarbonData将以压缩的多维索引列格式自动组织和存储数据。数据被加载到CarbonData后，就可以执行即席查询，CarbonData将对数据查询提供秒级响应。

CarbonData将数据源集成到Spark生态系统，用户可使用Spark SQL执行数据查询和分析，也可以使用Spark提供的第三方工具ThriftServer连接到Spark SQL。

CarbonData特性

- SQL功能：CarbonData与Spark SQL完全兼容，支持所有可以直接在Spark SQL上运行的SQL查询操作。

- 简单的Table数据集定义：CarbonData支持易于使用的DDL（数据定义语言）语句来定义和创建数据集。CarbonData DDL十分灵活、易于使用，并且足够强大，可以定义复杂类型的Table。
- 便捷的数据管理：CarbonData为数据加载和维护提供多种数据管理功能，支持加载历史数据以及增量加载新数据。CarbonData加载的数据可以基于加载时间进行删除，也可以撤销特定的数据加载操作。
- CarbonData文件格式是HDFS中的列式存储格式。该格式具有许多新型列存储文件的特性。例如，分割表，压缩模式等。

CarbonData独有的特点

- 伴随索引的数据存储：由于在查询中设置了过滤器，可以显著加快查询性能，减少I/O扫描次数和CPU资源占用。CarbonData索引由多个级别的索引组成，处理框架可以利用这个索引来减少需要安排和处理的任务，也可以通过在任务扫描中以更精细的单元（称为blocklet）进行skip扫描来代替对整个文件的扫描。
- 可选择的数据编码：通过支持高效的数据压缩和全局编码方案，可基于压缩/编码数据进行查询，在将结果返回给用户之前，才将编码转化为实际数据，这被称为“延迟物化”。
- 支持一种数据格式应用于多种用例场景：例如交互式OLAP-style查询，顺序访问（big scan），随机访问（narrow scan）。

CarbonData关键技术和优势

- 快速查询响应：高性能查询是CarbonData关键技术的优势之一。CarbonData查询速度大约是Spark SQL查询的10倍。CarbonData使用的专用数据格式围绕高性能查询进行设计，其中包括多种索引技术、全局字典编码和多次的Push down优化，从而对TB级数据查询进行最快响应。
- 高效率数据压缩：CarbonData使用轻量级压缩和重量级压缩的组合压缩算法压缩数据，可以减少60%~80%数据存储空间，大大节省硬件存储成本。

关于CarbonData的架构和详细原理介绍，请参见：**https://carbondata.apache.org/**。

1.4.3 ClickHouse

ClickHouse 简介

ClickHouse是一款开源的面向联机分析处理的列式数据库，其独立于Hadoop大数据体系，最核心的特点是压缩率和极速查询性能。同时，ClickHouse支持SQL查询，且查询性能好，特别是基于大宽表的聚合分析查询性能非常优异，比其他分析型数据库速度快一个数量级。

ClickHouse核心的功能特性介绍如下：

完备的DBMS功能

ClickHouse拥有完备的数据库管理功能，具备一个DBMS（Database Management System，数据库管理系统）基本的功能，如下所示。

- DDL（数据定义语言）：可以动态地创建、修改或删除数据库、表和视图，而无须重启服务。
- DML（数据操作语言）：可以动态查询、插入、修改或删除数据。
- 权限控制：可以按照用户粒度设置数据库或者表的操作权限，保障数据的安全性。

- 数据备份与恢复：提供了数据备份导出与导入恢复机制，满足生产环境的要求。
- 分布式管理：提供集群模式，能够自动管理多个数据库节点。

列式存储与数据压缩

ClickHouse是一款使用列式存储的数据库，数据按列进行组织，属于同一列的数据会被保存在一起，列与列之间也会由不同的文件分别保存。

在执行数据查询时，列式存储可以减少数据扫描范围和数据传输时的大小，提高了数据查询的效率。

例如在传统的行式数据库系统中，数据按如下表1-1顺序存储：

表 1-1 行式数据库

row	ID	Flag	Name	Event	Time
0	12345678901	0	name1	1	2020/1/11 15:19
1	32345678901	1	name2	1	2020/5/12 18:10
2	42345678901	1	name3	1	2020/6/13 17:38
N	…	…	…	…	…

行式数据库中处于同一行中的数据总是被物理的存储在一起，而在列式数据库系统中，数据按如下表1-2顺序存储：

表 1-2 列式数据库

row:	0	1	2	N
ID:	12345678901	32345678901	42345678901	…
Flag:	0	1	1	…
Name:	name1	name2	name3	…
Event:	1	1	1	…
Time:	2020/1/11 15:19	2020/5/12 18:10	2020/6/13 17:38	…

该示例中只展示了数据在列式数据库中数据的排列方式。对于存储而言，列式数据库总是将同一列的数据存储在一起，不同列的数据也总是分开存储，列式数据库更适合于OLAP（Online Analytical Processing）场景。

向量化执行引擎

ClickHouse利用CPU的SIMD指令实现了向量化执行。SIMD的全称是Single Instruction Multiple Data，即用单条指令操作多条数据，通过数据并行以提高性能的一种实现方

式（其他的还有指令级并行和线程级并行），它的原理是在CPU寄存器层面实现数据的并行操作。

关系模型与SQL查询

ClickHouse完全使用SQL作为查询语言，提供了标准协议的SQL查询接口，使得现有的第三方分析可视化系统可以轻松与它集成对接。

同时ClickHouse使用了关系模型，所以将构建在传统关系型数据库或数据仓库之上的系统迁移到ClickHouse的成本会变得更低。

数据分片与分布式查询

ClickHouse集群由1到多个分片组成，而每个分片则对应了ClickHouse的1个服务节点。分片的数量上限取决于节点数量（1个分片只能对应1个服务节点）。

ClickHouse提供了本地表（Local Table）与分布式表（Distributed Table）的概念。一张本地表等同于一份数据的分片。而分布式表本身不存储任何数据，它是本地表的访问代理，其作用类似分库中间件。借助分布式表，能够代理访问多个数据分片，从而实现分布式查询。

ClickHouse 应用场景

ClickHouse是Click Stream + Data WareHouse的缩写，起初应用于一款Web流量分析工具，基于页面的单击事件流，面向数据仓库进行OLAP分析。当前ClickHouse被广泛的应用于互联网广告、App和Web流量、电信、金融、物联网等众多领域，非常适用于商业智能化应用场景，在国内外有大量的应用和实践，具体请参考：**https://clickhouse.tech/docs/en/introduction/adopters/**。

ClickHouse 开源增强特性

MRS ClickHouse具备“手动挡”集群模式升级、平滑弹性扩容、高可用HA部署架构等优势能力，具体详情如下：

- 手动挡集群模式升级

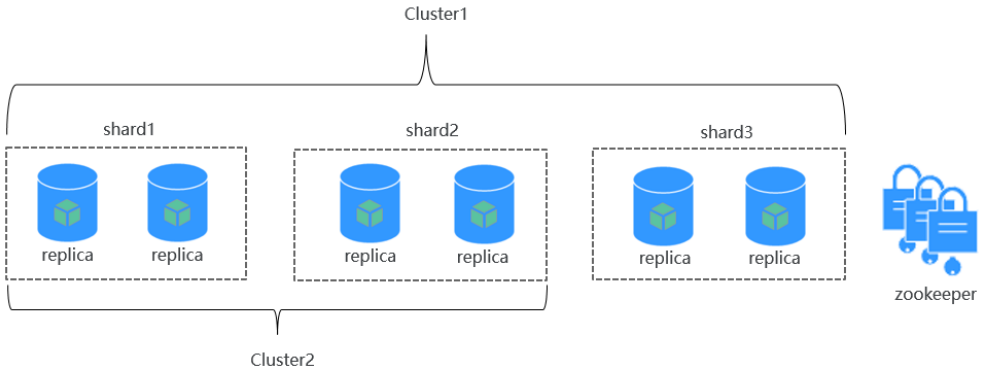
  如图1-7所示，多个ClickHouse节点组成的集群，没有中心节点，更多的是一个静态资源池的概念，业务要使用ClickHouse集群模式，需要预先在各个节点的配置文件中定义cluster信息，等所有参与的节点达成共识，业务才可以正确的交互访问，也就是说配置文件中的cluster是通常理解的“集群”概念。

  图 1-7 ClickHouse 集群

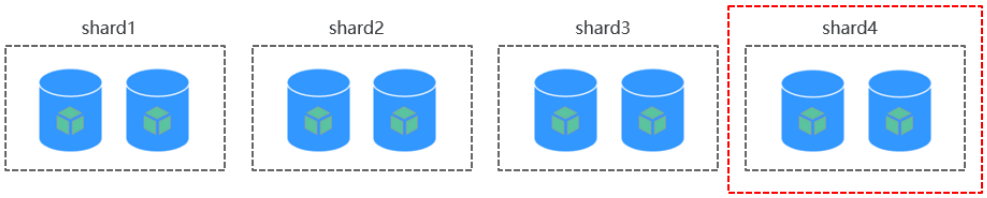
  常见的数据库系统，隐藏了表级以下的数据分区、副本存储等细节，用户是无感知的，而ClickHouse则要求用户主动来规划和定义数据分片（shard）、分区（partition）、副本（replica）位置等详细配置。它的这种类似“手动挡”的属性，给用户带来极不友好的体验，所以MRS服务的ClickHouse实例对这些工作做了统一的打包处理，适配成了“自动挡”，实现了统一管理，灵活易用。具体部署形态上，一个ClickHouse实例将包含3个Zookeeper节点和多个ClickHouse节点，采用Dedicated Replica模式，数据双副本高可靠。

图 1-8 ClickHouse 的 cluster 结构



- 平滑的弹性扩容能力

  随着业务的快速增长，面对集群存储容量或者CPU计算资源接近极限等场景，MRS服务提供了平滑的弹性扩容能力，能快速满足客户业务增长诉求。在用户对集群进行扩容ClickHouse节点时，MRS提供了一键式数据Balance均衡工具，并把数据均衡的主动权交给用户，由用户根据业务的特点，自由决定数据均衡的方式和时间点，以便保障业务可用性，实现了更加平滑的扩容能力。

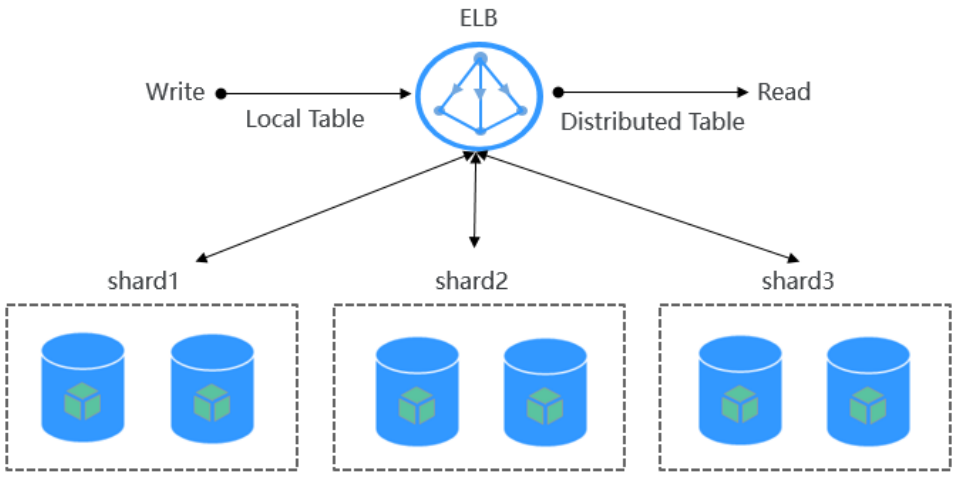


- 高可用HA部署架构

  MRS服务提供了基于ELB的HA部署架构，可以将用户访问流量自动分发到多台后端节点，扩展系统对外的服务能力，实现更高水平的应用容错。如图1-9所示，客户端应用请求集群时，使用ELB（Elastic Load Balance）来进行流量分发，通过ELB的轮询机制，写不同节点上的本地表（Local Table），读不同节点上的分布式表（Distributed Table），这样，无论集群写入的负载、读的负载以及应用接入的高可用性都具备了有力的保障。

  ClickHouse集群发放成功后，每个ClickHouse实例节点对应一个副本replica，两个副本组成一个shard逻辑分片。如创建ReplicatedMergeTree引擎表时，可以指定分片，相同分片内的两个副本数据就可以自动进行同步。

图 1-9 高可用 HA 部署架构图



1.4.4 DBService

1.4.4.1 DBService 基本原理

DBService 简介

DBService是一个高可用性的关系型数据库存储系统，适用于存储小量数据（10GB左右），比如：组件元数据。DBService仅提供给集群内部的组件使用，提供数据存储、查询、删除等功能。

DBService是集群的基础组件，Hive、Hue、Oozie、Loader和Redis组件将元数据存储在DBService上，并由DBService提供这些元数据的备份与恢复功能。

DBService 结构

DBService组件在集群中采用主备模式部署两个DBServer实例，每个DBServer实例包含三个模块：HA、Database和Floatip。

其逻辑结构如图1-10所示。

图 1-10 DBService 结构

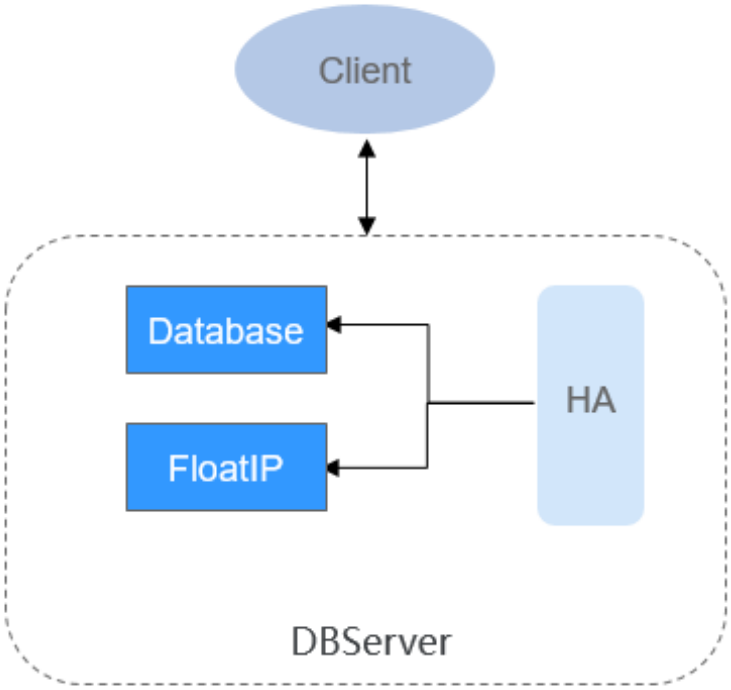


图1-10中各模块的说明如表1-3所示。

表 1-3 模块说明

名称	描述
HA	高可用性管理模块，主备DBServer通过HA进行管理。
Database	数据库模块，存储Client模块的元数据。
FloatIP	浮动IP，对外提供访问功能，只在主DBServer实例上启动浮动IP，Client模块通过该IP访问Database。
Client	使用DBService组件的客户端，部署在组件实例节点上，通过Floatip连接数据库，执行元数据的增加、删除、修改等操作。

1.4.4.2 DBService 与其他组件的关系

DBService是集群的基础组件，Hive、Hue、Oozie、Loader、Metadata和Redis组件将元数据存储在DBService上，并由DBService提供这些元数据的备份与恢复功能。

1.4.5 Flink

1.4.5.1 Flink 基本原理

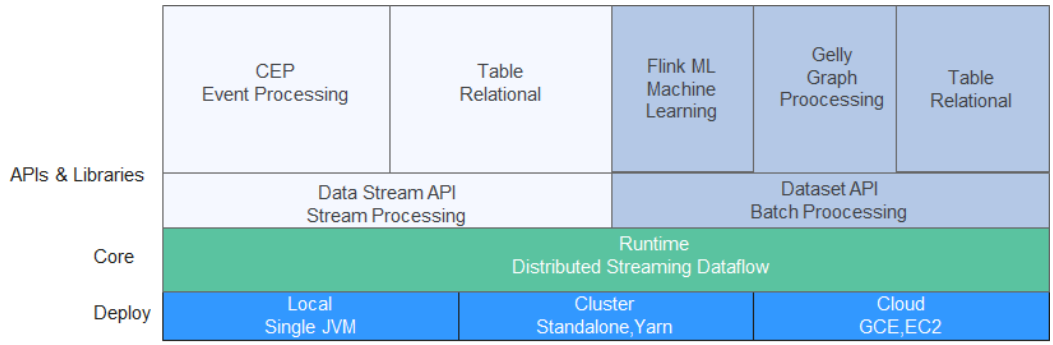
Flink 简介

Flink是一个批处理和流处理结合的统一计算框架，其核心是一个提供了数据分发以及并行化计算的流数据处理引擎。它的最大亮点是流处理，是业界主流的开源流处理引擎。

Flink最适合的应用场景是低时延的数据处理（Data Processing）场景：高并发pipeline处理数据，时延毫秒级，且兼具可靠性。

Flink技术栈如图1-11所示。

图 1-11 Flink 技术栈



Flink在当前版本中重点构建如下特性：

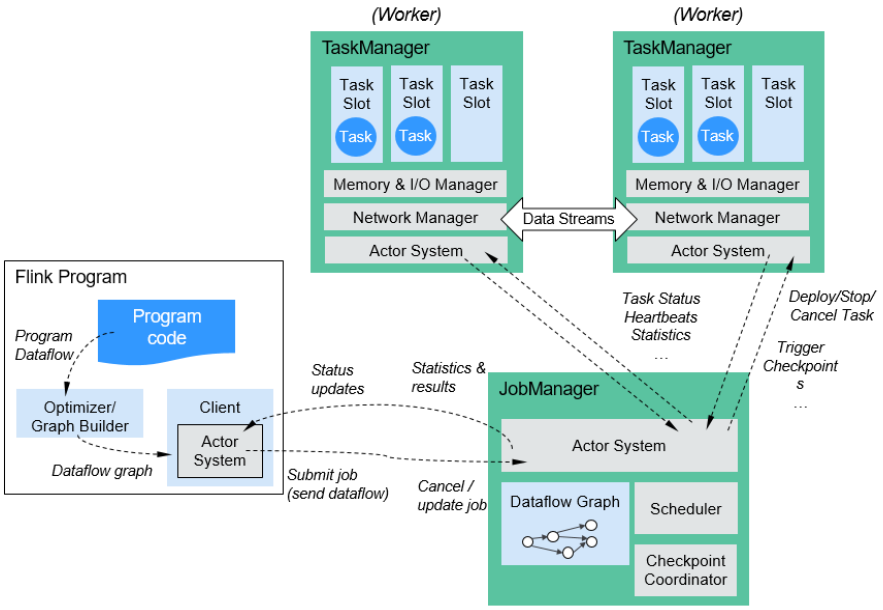
- DataStream
- Checkpoint
- 窗口
- Job Pipeline
- 配置表

其他特性继承开源社区，不做增强，具体请参考：**https://ci.apache.org/projects/flink/flink-docs-release-1.12/**。

Flink 结构

Flink结构如图1-12所示。

图 1-12 Flink 结构



Flink整个系统包含三个部分：

- Client

  Flink Client主要给用户提供向Flink系统提交用户任务（流式作业）的能力。

- TaskManager

  Flink系统的业务执行节点，执行具体的用户任务。TaskManager可以有多个，各个TaskManager都平等。

- JobManager

  Flink系统的管理节点，管理所有的TaskManager，并决策用户任务在哪些Taskmanager执行。JobManager在HA模式下可以有多个，但只有一个主JobManager。

如果您想了解更多关于Flink架构的信息，请参考链接：**https://ci.apache.org/projects/flink/flink-docs-master/docs/concepts/flink-architecture/**。

Flink 原理

- **Stream & Transformation & Operator**

  用户实现的Flink程序是由Stream和Transformation这两个基本构建块组成。

  a. Stream是一个中间结果数据，而Transformation是一个操作，它对一个或多个输入Stream进行计算处理，输出一个或多个结果Stream。

  b. 当一个Flink程序被执行的时候，它会被映射为Streaming Dataflow。一个Streaming Dataflow是由一组Stream和Transformation Operator组成，它类似于一个DAG图，在启动的时候从一个或多个Source Operator开始，结束于一个或多个Sink Operator。

  图1-13为一个由Flink程序映射为Streaming Dataflow的示意图。

图 1-13 Flink DataStream 示例

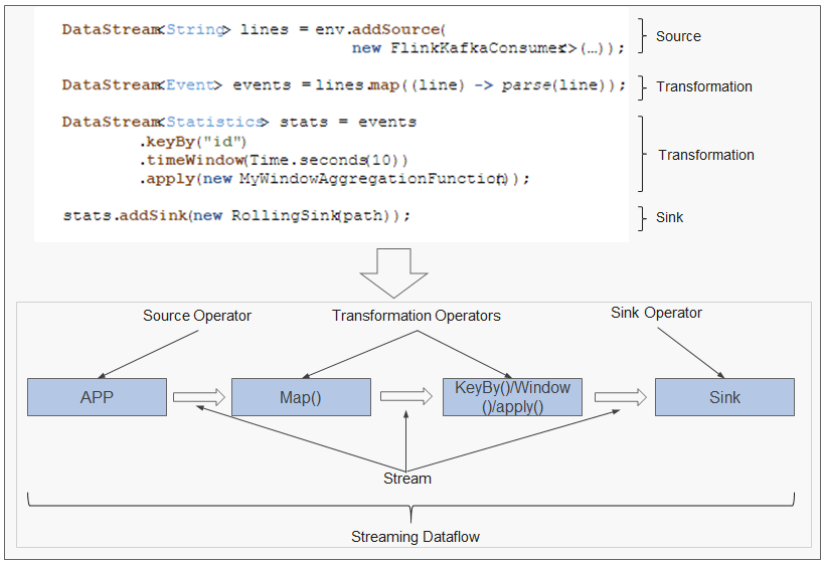


图1-13中“FlinkKafkaConsumer”是一个Source Operator，Map、KeyBy、TimeWindow、Apply是Transformation Operator，RollingSink是一个Sink Operator。

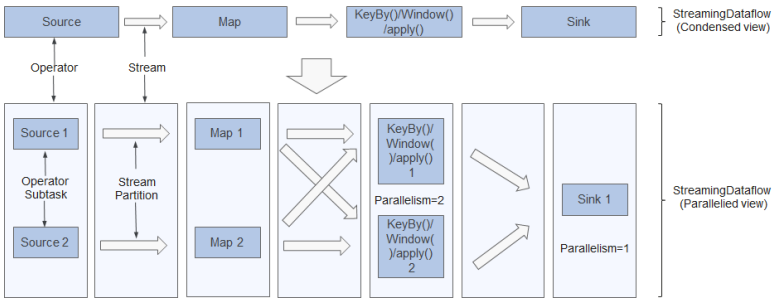
- **Pipeline Dataflow**

  在Flink中，程序是并行和分布式的方式运行。一个Stream可以被分成多个Stream分区（Stream Partitions），一个Operator可以被分成多个Operator Subtask。

  Flink内部有一个优化的功能，根据上下游算子的紧密程度来进行优化。

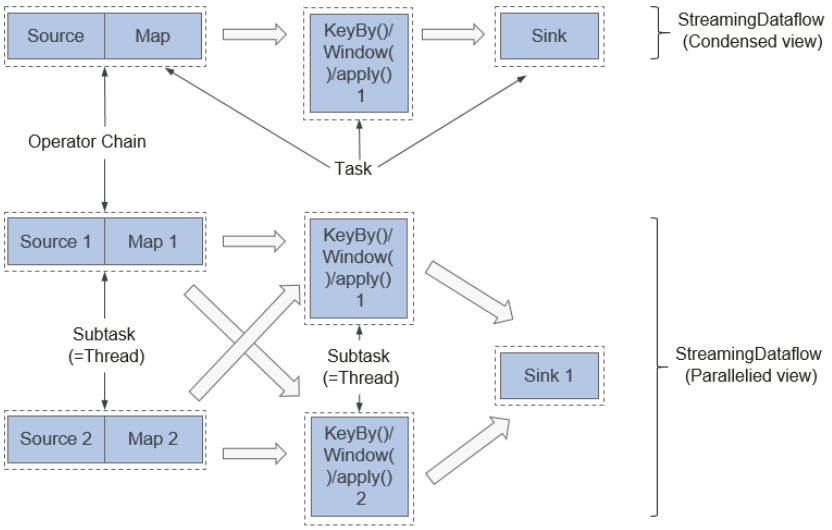
  - 紧密度低的算子则不能进行优化，而是将每一个Operator Subtask放在不同的线程中独立执行。一个Operator的并行度，等于Operator Subtask的个数，一个Stream的并行度（分区总数）等于生成它的Operator的并行度，如图1-14所示。

    图 1-14 Operator

  - 紧密度高的算子可以进行优化，优化后可以将多个Operator Subtask串起来组成一个Operator Chain，实际上就是一个执行链，每个执行链会在TaskManager上一个独立的线程中执行，如图1-15所示。

图 1-15 Operator chain



- 图1-15中上半部分表示的是将Source和map两个紧密度高的算子优化后串成一个Operator Chain，实际上一个Operator Chain就是一个大的Operator的概念。图中的Operator Chain表示一个Operator，keyBy表示一个Operator，Sink表示一个Operator，它们通过Stream连接，而每个Operator在运行时对应一个Task，也就是说图中的上半部分有3个Operator对应的是3个Task。
- 图1-15中下半部分是上半部分的一个并行版本，对每一个Task都并行化为多个Subtask，这里只是演示了2个并行度，Sink算子是1个并行度。

Flink 关键特性

- 流式处理

  高吞吐、高性能、低时延的实时流处理引擎，能够提供ms级时延处理能力。

- 丰富的状态管理

  流处理应用需要在一定时间内存储所接收到的事件或中间结果，以供后续某个时间点访问并进行后续处理。Flink提供了丰富的状态管理相关的特性支持，其中包括

  - 多种基础状态类型：Flink提供了多种不同数据结构的状态支持，如ValueState、ListState、MapState等。用户可以基于业务模型选择最高效、合适状态类型。
  - 丰富的State Backend：State Backend负责管理应用程序的状态，并根据需要进行Checkpoint。Flink提供了不同State Backend，State可以存储在内存上或RocksDB等上，并支持异步以及增量的Checkpoint机制。
  - 精确一次语义：Flink的Checkpoint和故障恢复能力保证了任务在故障发生前后的应用状态一致性，为某些特定的存储支持了事务型输出的功能，即使在发生故障的情况下，也能够保证精确一次的输出。

- 丰富的时间语义支持

  时间是流处理应用的重要组成部分，对于实时流处理应用来说，基于时间语义的窗口聚合、检测、匹配等运算是非常常见的。Flink提供了丰富的时间语义支持。

  - Event-time：使用事件本身自带的时间戳进行计算，使乱序到达或延迟到达的事件处理变得更加简单。
  - Watermark支持：Flink引入Watermark概念，用以衡量事件时间的发展。Watermark也为平衡处理时延和数据完整性提供了灵活的保障。当处理带有Watermark的事件流时，在计算完成之后仍然有相关数据到达时，Flink提供了多种处理选项，如将数据重定向（side output）或更新之前完成的计算结果。
  - Processing-time和Ingestion-time支持。
  - 高度灵活的流式窗口支持：Flink能够支持时间窗口、计数窗口、会话窗口，以及数据驱动的自定义窗口，可以通过灵活的触发条件定制，实现复杂的流式计算模式。

- 容错机制

  分布式系统，单个task或节点的崩溃或故障，往往会导致整个任务的失败。Flink提供了任务级别的容错机制，保证任务在异常发生时不会丢失用户数据，并且能够自动恢复。

  - Checkpoint：Flink基于Checkpoint实现容错，用户可以自定义对整个任务的Checkpoint策略，当任务出现失败时，可以将任务恢复到最近一次Checkpoint的状态，从数据源重发快照之后的数据。
  - Savepoint：一个Savepoint就是应用状态的一致性快照，Savepoint与Checkpoint机制相似，但Savepoint需要手动触发，Savepoint保证了任务在升级或迁移时，不丢失掉当前流应用的状态信息，便于任何时间点的任务暂停和恢复。

- Flink SQL

  Table API和SQL借助了Apache Calcite来进行查询的解析，校验以及优化，可以与DataStream和DataSet API无缝集成，并支持用户自定义的标量函数，聚合函数以及表值函数。简化数据分析、ETL等应用的定义。下面代码实例展示了如何使用Flink SQL语句定义一个会话单击量的计数应用。

  ```
  SELECT userId, COUNT(*)
  FROM clicks
  GROUP BY SESSION(clicktime, INTERVAL '30' MINUTE), userId
  ```

  有关Flink SQL的更多信息，请参见：**https://ci.apache.org/projects/flink/flink-docs-master/dev/table/sqlClient.html**。

- CEP in SQL

  Flink允许用户在SQL中表示CEP（Complex Event Processing）查询结果以用于模式匹配，并在Flink上对事件流进行评估。

  CEP SQL 通过MATCH_RECOGNIZE的SQL语法实现。MATCH_RECOGNIZE子句自Oracle Database 12c起由Oracle SQL支持，用于在SQL中表示事件模式匹配。CEP SQL使用举例如下：

  ```
  SELECT T.aid, T.bid, T.cid
  FROM MyTable
    MATCH_RECOGNIZE (
      PARTITION BY userid
      ORDER BY proctime
      MEASURES
        A.id AS aid,
        B.id AS bid,
        C.id AS cid
      PATTERN (A B C)
      DEFINE
        A AS name = 'a',
        B AS name = 'b',
        C AS name = 'c'
    ) AS T
  ```

1.4.5.2 Flink HA 方案介绍

Flink HA 方案介绍

每个Flink集群只有单个JobManager，存在单点失败的情况。Flink有YARN、Standalone和Local三种模式，其中YARN和Standalone是集群模式，Local是指单机模式。但Flink对于YARN模式和Standalone模式提供HA机制，使集群能够从失败中恢复。这里主要介绍YARN模式下的HA方案。

Flink支持HA模式和Job的异常恢复。这两项功能高度依赖ZooKeeper，在使用之前用户需要在“flink-conf.yaml”配置文件中配置ZooKeeper，配置ZooKeeper的参数如下：

```
high-availability: zookeeper
high-availability.zookeeper.quorum: ZooKeeperIP地址:2181
high-availability.storageDir: hdfs:///flink/recovery
```

YARN模式

Flink的JobManager与YARN的Application Master（简称AM）是在同一个进程下。YARN的ResourceManager对AM有监控，当AM异常时，YARN会将AM重新启动，启动后，所有JobManager的元数据从HDFS恢复。但恢复期间，旧的业务不能运行，新的业务不能提交。ZooKeeper上还是存有JobManager的元数据，比如运行Job的信息，会提供给新的JobManager使用。对于TaskManager的失败，由JobManager上Akka的DeathWatch机制监听处理。当TaskManager失败后，重新向YARN申请容器，创建TaskManager。

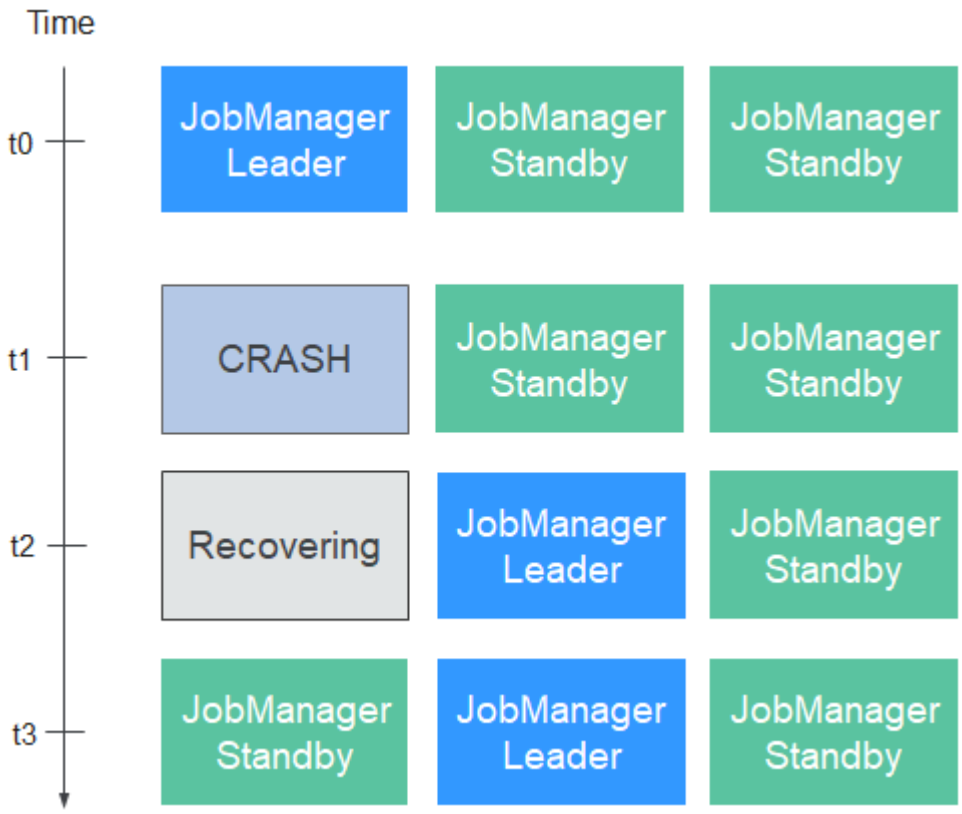
YARN模式的HA方案的更多信息，可参考链接：**http://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/ResourceManagerHA.html**。

关于YARN的**yarn-site.xml**设置，可参考链接：**https://ci.apache.org/projects/flink/flink-docs-release-1.12/ops/jobmanager_high_availability.html**。

Standalone模式

对于Standalone模式的集群，可以启动多个JobManager，然后通过ZooKeeper选举出leader作为实际使用的JobManager。该模式下可以配置一个主JobManager（Leader JobManager）和多个备JobManager（Standby JobManager），这能够保证当主JobManager失败后，备的某个JobManager可以承担主的职责。**图1-16**为主备JobManager的恢复过程。

图 1-16 恢复过程



TaskManager恢复

对于TaskManager的失败，由JobManager上Akka的DeathWatch机制监听处理。当TaskManager失败后，由JobManager负责创建一个新TaskManager，并把业务迁移到新的TaskManager上。

JobManager恢复

Flink的JobManager与YARN的Application Master（简称AM）是在同一个进程下。YARN的ResourceManager对AM有监控，当AM异常时，YARN会将AM重新启动，启动后，所有JobManager的元数据从HDFS恢复。但恢复期间，旧的业务不能运行，新的业务不能提交。

Job恢复

Job的恢复必须在Flink的配置文件中配置重启策略。当前包含三种重启策略：fixed-delay、failure-rate和none。只有配置fixed-delay、failure-rate，job才可以恢复。另外，如果配置了重启策略为none，但job设置了Checkpoint，默认会将重启策略改为fixed-delay，且重试次数是配置项“restart-strategy.fixed-delay.attempts”配置为“Integer.MAX_VALUE”。

三种策略的具体信息请参考Flink官网：**https://ci.apache.org/projects/flink/flink-docs-release-1.12/dev/task_failure_recovery.html**。配置策略的参考如下：

```
restart-strategy: fixed-delay
restart-strategy.fixed-delay.attempts: 3
restart-strategy.fixed-delay.delay: 10 s
```

以下场景的异常，都会导致job重新恢复：

- 当JobManager失败后，所有Job会停止，直到新的JobManager起来后，所有Job恢复。
- 当某一TaskManager失败后，这个TaskManager上的所有作业都将停止，然后等待有可用资源后重启。
- 当某个Job的Task失败后，整个Job也会重启。

说明

有关Job的配置重启策略，具体内容请参见https://ci.apache.org/projects/flink/flink-docs-release-1.12/ops/jobmanager_high_availability.html。

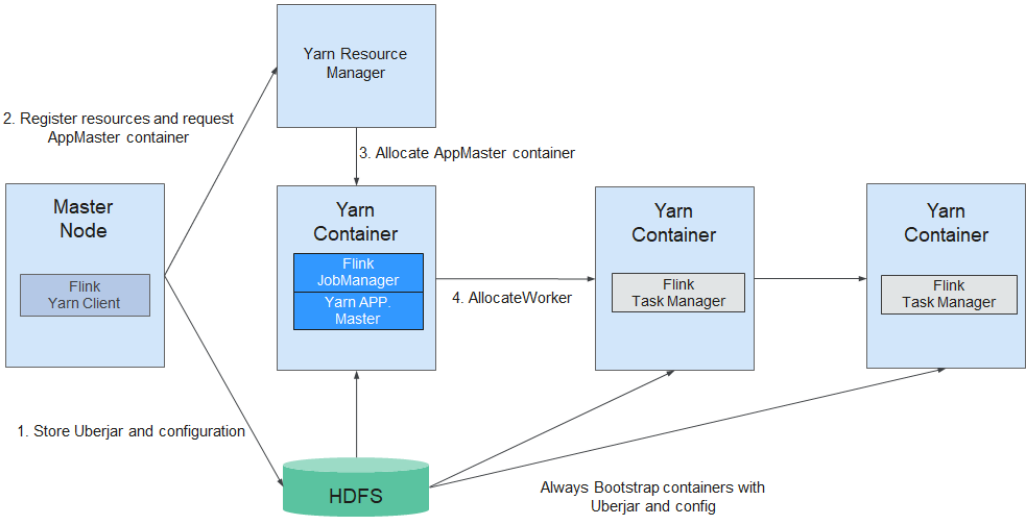
1.4.5.3 Flink 与其他组件的关系

Flink 与 YARN 的关系

Flink支持基于YARN管理的集群模式，在该模式下，Flink作为YARN上的一个应用，提交到YARN上执行。

Flink基于YARN的集群部署如图1-17所示。

图 1-17 Flink 基于 YARN 的集群部署



1. Flink YARN Client首先会检验是否有足够的资源来启动YARN集群，如果资源足够的话，会将jar包、配置文件等上传到HDFS。
2. Flink YARN Client首先与YARN Resource Manager进行通信，申请启动Application Master（以下简称AM）的Container，并启动AM。等所有的YARN的Node Manager将HDFS上的jar包、配置文件下载后，则表示AM启动成功。
3. AM在启动的过程中会和YARN的RM进行交互，向RM申请需要的Task Manager Container，申请到Task Manager Container后，启动TaskManager进程。
4. 在Flink YARN的集群中，AM与Flink JobManager在同一个Container中。AM会将JobManager的RPC地址通过HDFS共享的方式通知各个TaskManager，TaskManager启动成功后，会向JobManager注册。
5. 等所有TaskManager都向JobManager注册成功后，Flink基于YARN的集群启动成功，Flink YARN Client就可以提交Flink Job到Flink JobManager，并进行后续的映射、调度和计算处理。

1.4.5.4 Flink 开源增强特性

1.4.5.4.1 窗口

Flink 开源特性增强：窗口

本节主要介绍滑动窗口，并提供滑动窗口优化方式。窗口的详细内容请参见官网：**https://ci.apache.org/projects/flink/flink-docs-release-1.12/dev/stream/operators/windows.html**。

窗口介绍

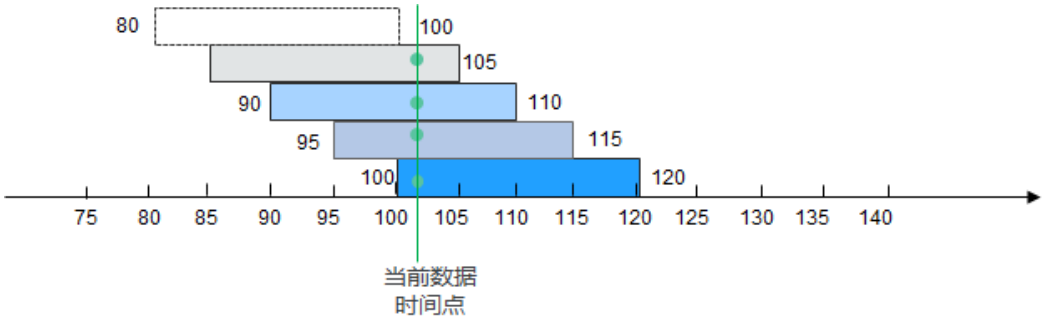
窗口中数据的保存形式主要有中间结果和原始数据两种，对窗口中的数据使用公共算子，如sum等操作时（window(SlidingEventTimeWindows.of(Time.seconds(20), Time.seconds(5))).sum）仅会保留中间结果；当用户使用自定义窗口时（window(SlidingEventTimeWindows.of(Time.seconds(20), Time.seconds(5))).apply(new UDF)）时保存所有的原始数据。

用户使用自定义SlidingEventTimeWindow和SlidingProcessingTimeWindow时，数据以多备份的形式保存。假设窗口的定义如下：

```
window(SlidingEventTimeWindows.of(Time.seconds(20), Time.seconds(5))).apply(new UDFWindowFunction)
```

当一个数据到来时，会被分配到20/5=4个不同的窗口中，即：数据在内存中保存了4份。当窗口大小/滑动周期非常大时，冗余现象非常严重，难以接受。

图 1-18 窗口原始结构示例



假设一个数据在102秒时到来，它将会被分配到[85, 105)、[90, 110)、[95, 115)以及[100, 120)四个不同的窗口中。

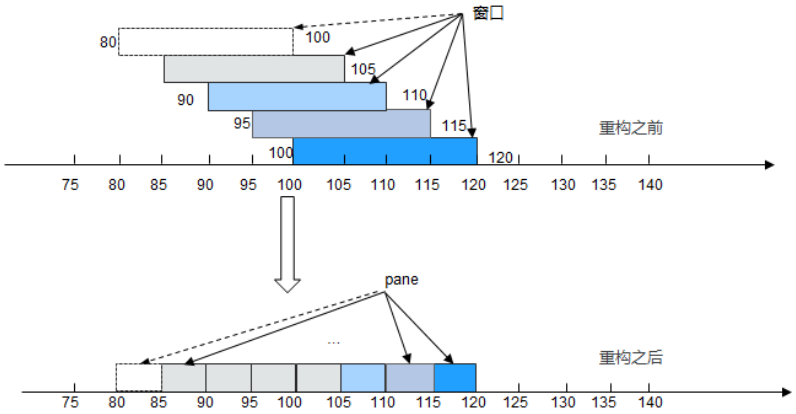
窗口优化

针对上述SlidingEventTimeWindow和SlidingProcessingTimeWindow在保存原始数据时存在的数据冗余问题，对保存原始数据的窗口进行重构，优化存储，使其存储空间大大降低，具体思路如下：

1. 以滑动周期为单位，将窗口划分为若干相互不重合的pane。

   每个窗口由一到多个pane组成，多个pane对窗口构成了覆盖关系。所谓一个pane即一个滑动周期，如：在窗口window(SlidingEventTimeWindows.of(Time.seconds(20), Time.seconds.of(5)))中pane的大小为5秒，假设这个窗口为[100, 120)，则包含的pane为[100, 105), [105, 110), [110, 115), [115, 120)。

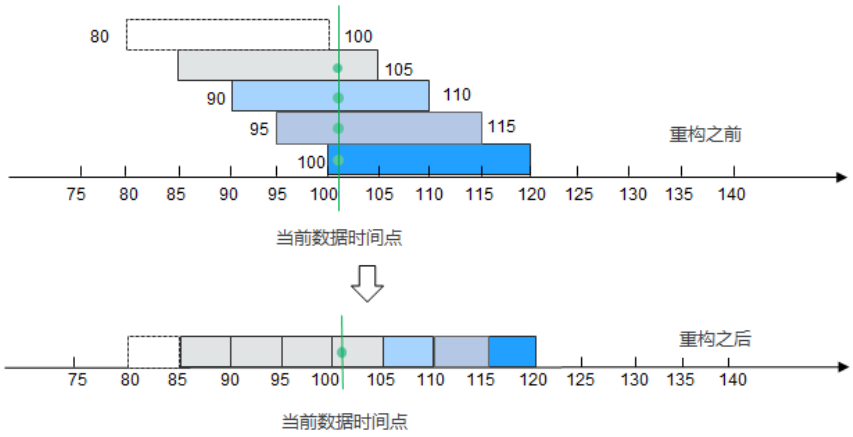
图 1-19 窗口重构示例



2. 当某个数据到来时，并不分配到具体的窗口中，而是根据自己的时间戳计算出该数据所属的pane，并将其保存到对应的pane中。

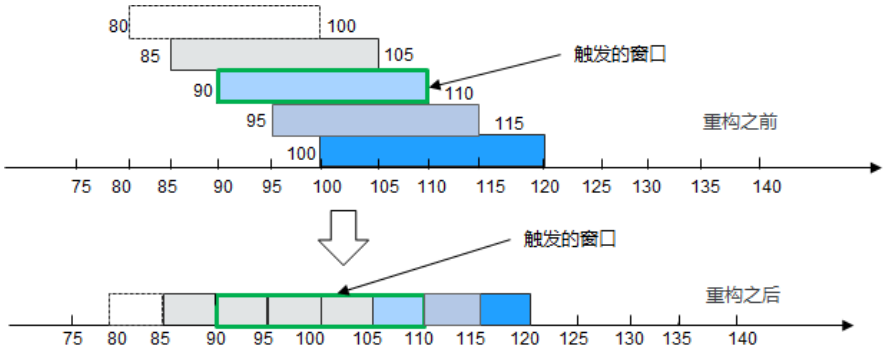
   一个数据仅保存在一个pane中，内存中只有一份。

图 1-20 窗口保存数据示例



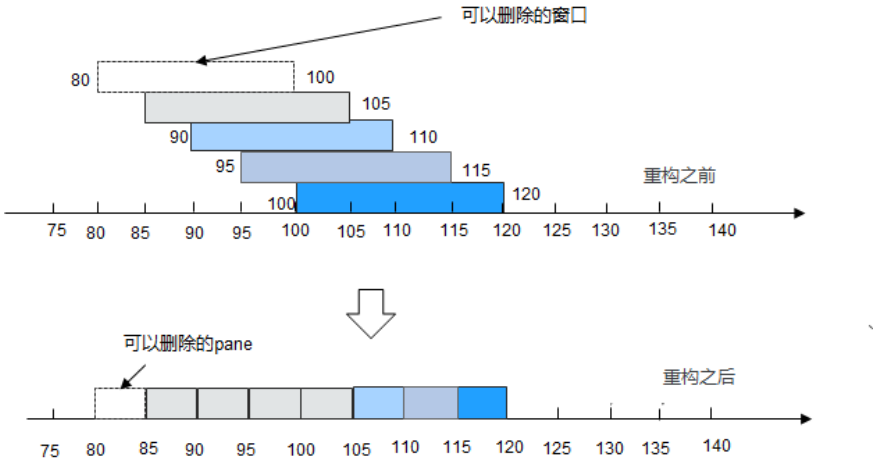
3. 当需要触发某个窗口时，计算该窗口包含的所有pane，并取出合并成一个完整的窗口计算。

图 1-21 窗口触发计算示例



4. 当某个pane不再需要时，将其从内存中删除。

图 1-22 窗口删除示例



通过优化，可以大幅度降低数据在内存以及快照中的数量。

1.4.5.4.2 Job Pipeline

Flink 开源增强特性：Job Pipeline

通常情况下，会将与某一方面业务相关的逻辑代码放在一个比较大的Jar包中，这种Jar包称为Fat Jar。Fat Jar具有以下缺点：

- 随着业务逻辑越来越复杂，Jar包的大小也不断增加。
- 协调难度增大，所有的业务开发人员都在同一套业务逻辑上开发，虽然可以将整个业务逻辑划分为几个模块，单各模块之间是一种紧耦合的关系，当需求更改时，需要重新规划整个流图。

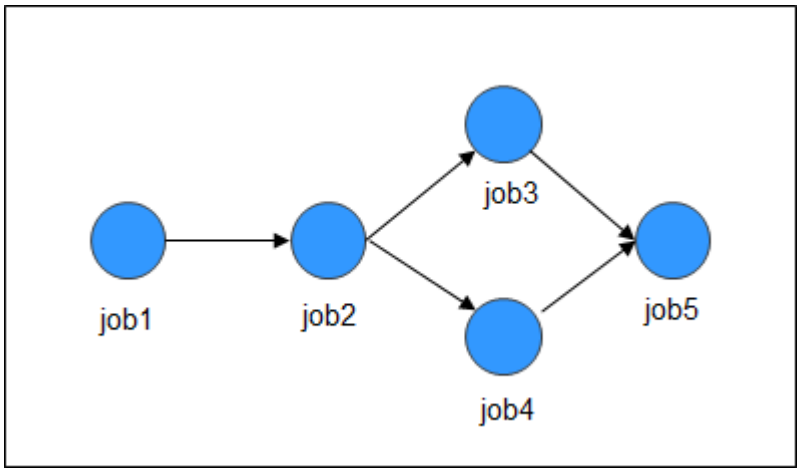
拆分成多个作业目前还存在问题。

- 通常情况下，作业之间可以通过Kafka实现数据传输，如作业A可以将数据发送到Kafka的Topic A下，然后作业B和作业C可以从Topic A下读取数据。该方案简单易行，但是延迟很难做到100ms以内。
- 采用TCP直接相连的方式，算子在分布式环境下，可能会调度到任意节点，上下游之间无法感知其存在。

Job Pipeline流图结构

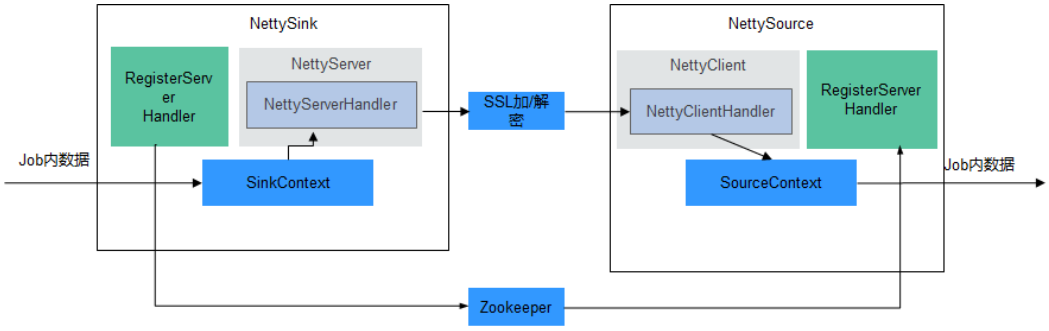
Pipeline是由Flink的多个Job通过TCP连接起来，上游Job可以直接向下游Job发送数据。这种发送数据的流图称为Job Pipeline，如图1-23所示。

图 1-23 Job Pipeline 流图



Job Pipeline原理介绍

图 1-24 Job Pipeline 原理图



- NettySink和NettySource

  Pipeline中上下游Job是直接通过Netty进行通信，上游Job的Sink算子作为Server，下游Job的Source算子作为Client。上游Job的Sink算子命名为NettySink，下游Job的Source算子命名为NettySource。
- NettyServer和NettyClient

  NettySink作为Netty的服务器端，内部NettyServer实现服务器功能；NettySource作为Netty的客户端，内部NettyClient实现客户端功能。
- 发布者

  通过NettySink向下游Job发送数据的Job称为发布者。
- 订阅者

  通过NettySource接收上游Job发送的数据的Job称为订阅者。
- 注册服务器

  保存NettyServer的IP、端口以及NettySink的并发度信息的第三方存储器。
- 总体架构是一个三层结构，由外到里依次是：
  - NettySink->NettyServer->NettyServerHandler
  - NettySource->NettyClient->NettyClientHandler

Job Pipeline功能介绍

- **NettySink**

  NettySink由以下几个重要模块组成：

  - RichParallelSinkFunction

    NettySink继承了RichParallelSinkFunction，使其具有Sink算子的属性。主要通过RichParallelSinkFunction的接口来实现以下功能：

    - 启动NettySink算子。
    - 运行NettySink算子，从本job的上游算子接收数据。
    - 取消NettySink算子运行等。

    也可以通过其属性获取以下信息：

    - NettySink算子各个并发度的subtaskIndex信息。
    - NettySink算子的并发度是多少。

  - RegisterServerHandler

    该组件主要是与注册服务器交互的部件，在平台上定义了一系列接口，包括以下几种接口：

    - “start();”：启动RegisterServerHandler，与第三方RegisterServer建立联系。
    - “createTopicNode();”：创建Topic节点。
    - “register();”： 将IP、端口及并发度信息注册到Topic节点下。
    - “deleteTopicNode();”： 删除Topic节点。
    - “unregister();”： 删除注册信息。
    - “query(); ”：查询注册信息。
    - “isExist();”： 查找某个信息是否存在。
    - “shutdown(); ”：关闭RegisterServerHandler，与第三方RegisterServer断开连接。

说明

- RegisterServerHandler接口实现了ZooKeeper作为RegisterServer的Handler，用户可以根据自己的需求，实现自己的Handler，ZooKeeper中信息的保存形式如下图所示：

```
Namespace
|---Topic-1
  |---parallel-1
  |---parallel-2
  |....
  |---parallel-n
|---Topic-2
  |---parallel-1
  |---parallel-2
  |....
  |---parallel-m
|...
```

- Namespace的信息通过“flink-conf.yaml”的以下配置项获取：

```
nettyconnector.registerserver.topic.storage: /flink/nettyconnector
```

- ZookeeperRegisterServerHandler与ZooKeeper之间的SASL认证通过Flink的框架实现。
- 用户必须自己保证每个Job有一个唯一的TOPIC，否则会引起作业间订阅关系的混乱。
- 在ZookeeperRegisterServerHandler调用shutdown()时，首先删除本并发度的注册信息，然后尝试删除TOPIC节点，如果TOPIC节点为非空，则放弃删除TOPIC节点，说明其他并发度还未退出。

– NettyServer

该模块是NettySink算子的核心之一，主要作用是创建一个NettyServer并接收NettyClient的连接申请。将同一Job中上游算子发送过来的数据，经由NettyServerHandler发送出去。 另外，NettyServer的端口及子网需要在“flink-conf.yaml”配置文件中配置：

- 端口范围

```
nettyconnector.sinkserver.port.range: 28444-28943
```

- 子网

```
nettyconnector.sinkserver.subnet: 10.162.222.123/24
```

说明

nettyconnector.sinkserver.subnet默认配置为Flink客户端所在节点子网，若客户端与TaskManager不在同一个子网则有可能导致错误，需手动配置为TaskManager所在网络子网（业务IP）。

– NettyServerHandler

该Handler是NettySink与订阅者交互的通道，当NettySink接收到消息时，该Handler负责将消息发送出去。为保证数据传输的安全性，该通道通过SSL加密。另外设置一个Netty Connector的功能开关，只有当Flink的SSL总开关被打开以及配置“nettyconnector.ssl.enabled”为“true”的时候才开启SSL加密，否则不开启。

- **NettySource**

NettySource由以下几个重要模块组成：

– RichParallelSourceFunction

NettySource继承了RichParallelSinkFunction，使其具有Source算子的属性，主要通过RichParallelSourceFunction接口来实现以下功能：

- 启动NettySink算子。
- 运行NettySink算子，接收来自订阅者的数据并注入到所在Job中。
- 取消Source算子运行等。

也可以通过其属性获取以下信息：

- NettySource算子各个并发度的subtaskIndex信息。
- NettySource算子的并发度是多少。

当NettySource算子进入run阶段后，平台内部会不断监控其NettyClient状态是否健康，一旦发现其出现异常，即会重启NettyClient，重新与NettyServer建立连接并接收数据，以防接收的数据混乱。

- RegisterServerHandler

  该组件与NettySink的RegisterServerHandler功能相同，在NettySource算子中仅获取所订阅Job的各个并发算子的IP、端口及并发算子信息。

- NettyClient

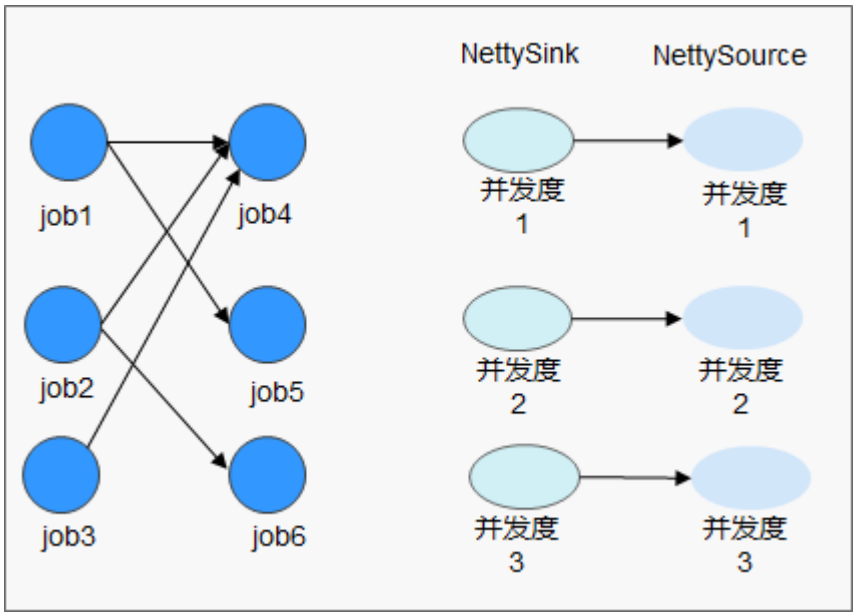
  NettyClient与NettyServer建立连接，并通过NettyClientHandler接收数据。每个NettySource算子必须具有唯一的name（由用户来保障）。NettyServer通过唯一的name确定每个Client来自不同的NettySource。当NettyClient与NettyServer建立连接时，首先向NettyServer注册NettyClient，将NettyClient的NettySource name传递给NettyServer。

- NettyClientHandler

  该模块是与发布者交互的通道，也是与Job的其他算子交互的通道。当该通道中接收到消息时，该Handler负责将消息注入到Job内部。另外，为保证数据安全传输，该通道通过SSL加密，与NettySink进行通信。另外设置一个NettyConnector的功能开关，只有当Flink的SSL总开关被打开以及“nettyconnector.ssl.enabled”为“true”的时候才开启SSL加密，否则不开启。

Job与Job之间的联系可能是多对多的关系，对于每个NettySink和NettySource算子的并发度而言，是一对多的关系，如**图1-25**所示。

图 1-25 关系图



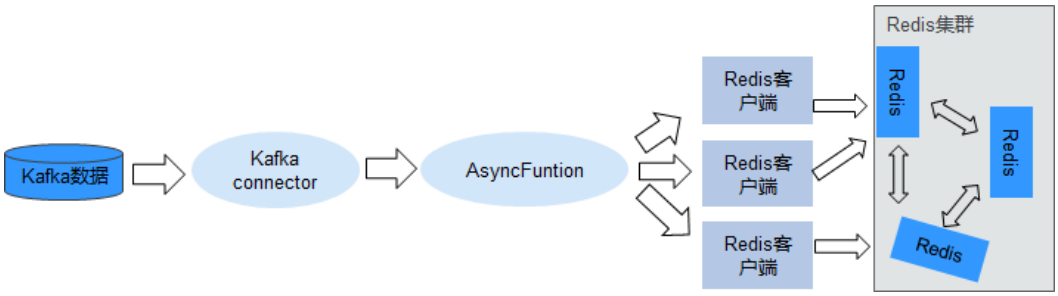
1.4.5.4.3 配置表

Flink 开源特性增强：配置表

在某些场景下，用户存在固定的配置表，存储了基础信息；当平台接收流数据并处理时，需要与配置表进行匹配操作。由于配置表可能较大，考虑使用Redis存储，Redis是一个高性能的key-value数据库，流数据查询时延较低。

具体流程如下：

图 1-26 流程图



Redis存储数据

Redis并不是简单的key-value存储，实际上它是一个数据结构服务器，支持不同类型的值。支持数据类型存储如下：

- 二进制安全的字符串。
- Lists: 按插入顺序排序的字符串元素的集合。基本上就是链表（*linked lists*）。
- Sets: 不重复且无序的字符串元素的集合。
- Sorted sets：每个字符串元素都关联到一个叫*score*浮动数值（floating number value）。里面的元素是通过score进行排序，它是可以检索的一系列元素。

- Hashes：由field和关联的value组成的map，field和value都是字符串。
- Bit arrays: 通过特殊的命令，用户可以将String值当作一系列bits处理。例如用户可以设置和清除单独的bits，统计出所有设为1的bits的数量，或找到第一个被设为1或0的bit等等。
- HyperLogLogs: 这是被用于估计一个set中元素数量的概率性的数据结构。

为满足最大5亿条数据配置表的存储并及时响应查询，使用Redis集群存储配置表，并使用流的异步IO作消息查询，提高数据处理的吞吐量。

说明

- Redis集群：在集群环境上的各个节点上部署Redis，并将数据分散存储在各个节点上，提升了存储容量，目前MRS中已有Redis组件。
- 异步IO：处理流数据，最大化数据处理的吞吐量，提高处理效率。

涉及Redis主要有两部分，Redis安装部署以及配置表数据导入：

1. Redis安装。

   MRS已经有Redis组件，在集群安装时可以勾选安装。

2. 配置表导入Redis。

   用户可以按照配置表的特征选取主键或者关键某几列作为key值，当需要存储的配置表的属性较多时，建议以Hashes的数据形式存储。

   MRS的Redis组件提供了Redis客户端对数据进行插入查询，可以参考Redis组件样例代码。

说明

Redis数据类型详细信息请参见官网：**https://redis.io/topics/data-types-intro**。

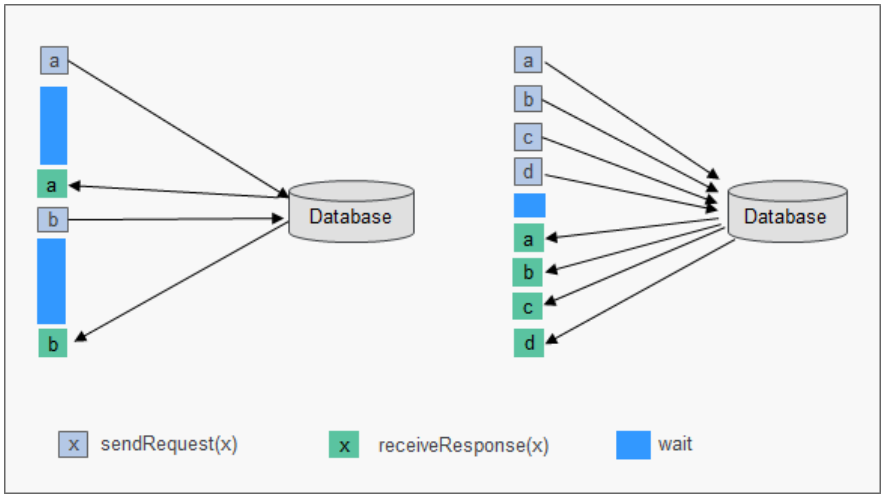
Flink异步IO

当与外部系统进行交互时，如外部的数据库，访问等待时间过长导致数据处理效率低。异步IO实现了不需要等待请求返回就可以同时发送其他请求，以此提高数据吞吐量。

异步IO的API实现需要注意三点：

- AsyncFunction函数实现了数据处理的异步处理，需要重写asyncInvoke方法。
- 回调函数获取算子的结果，并且通过AsyncCollector收集起来。

图 1-27 Async.I/O 的比较



- 超时时间和最大容量设置。

  超时时间定义了一个异步请求失败的最大时间。最大容量设置是指同时可以存在多少个异步请求，过多导致资源消耗加大；过小导致并行数小，吞吐量不能提高；建议针对数据源特点进行合理适配。

1.4.5.4.4 Stream SQL Join

Flink 开源增强特性：Stream SQL Join

Flink的Table API&SQL是一种用于Scala 和Java的语言集成式查询API，它支持非常直观的从关系运算符（如选择、筛选和连接）进行组合查询。Table API&SQL详细内容请参见官网：**https://ci.apache.org/projects/flink/flink-docs-release-1.12/dev/table/index.html**。

Stream SQL Join介绍

SQL Join用于根据两个或多个表中的列之间的关系，从这些表中查询数据。Flink Stream SQL Join允许对两个流式table进行join，并从中查询结果。支持类似于以下内容的查询：

```
SELECT o.proctime, o.productId, o.orderId, s.proctime AS shipTime
FROM Orders AS o
JOIN Shipments AS s
ON o.orderId = s.orderId
AND o.proctime BETWEEN s.proctime AND s.proctime + INTERVAL '1' HOUR;
```

目前，Stream SQL Join需在指定的窗口范围内进行。对窗口范围内的数据进行连接，需要至少一个相等连接谓词和一个绑定双方时间的条件。这个条件可以由两个适当的范围谓词（**<, <=, >=, >**），一个**BETWEEN**谓词或者一个单一的相等谓词来定义。这个相等谓词主要是比较两个输入表的同类型时间属性（比如处理时间或者事件时间）。

以下是一个关于在收到订单后四小时内发货, 将所有订单及其相应的货件进行Join的示例：

```
SELECT *
FROM Orders o, Shipments s
WHERE o.id = s.orderId AND
o.ordertime BETWEEN s.shiptime - INTERVAL '4' HOUR AND s.shiptime
```

说明

1. Stream SQL Join仅支持Inner Join。
2. ON子句应包括相等连接条件。
3. 时间属性只支持处理时间和事件时间。
4. 窗口条件只支持有界的时间范围，如 **o.proctime BETWEEN s.proctime - INTERVAL '1' HOUR AND s.proctime + INTERVAL '1' HOUR**，不支持像**o. proctime > s.proctime**这样无界的范围，并应包括两个流的proctime属性，不支持o.proctime BETWEEN **proctime ()** AND **proctime () + 1**。

1.4.5.4.5 Flink CEP in SQL

SQL 中的 Flink CEP

CloudStream扩展为允许用户在SQL中表示CEP查询结果以用于模式匹配，并在Flink引擎上对事件流进行评估。

SQL 查询语法

通过MATCH_RECOGNIZE的SQL语法实现。MATCH_RECOGNIZE子句自Oracle Database 12c起由Oracle SQL支持，用于在SQL中表示事件模式匹配。Apache Calcite同样支持MATCH_RECOGNIZE子句。

由于Flink通过Calcite分析SQL查询结果，本操作遵循Apache Calcite语法。

```
MATCH_RECOGNIZE (
    [ PARTITION BY expression [, expression ]* ]
    [ ORDER BY orderItem [, orderItem ]* ]
    [ MEASURES measureColumn [, measureColumn ]* ]
    [ ONE ROW PER MATCH | ALL ROWS PER MATCH ]
    [ AFTER MATCH
          ( SKIP TO NEXT ROW
          | SKIP PAST LAST ROW
          | SKIP TO FIRST variable
          | SKIP TO LAST variable
          | SKIP TO variable )
    ]
    PATTERN ( pattern )
    [ WITHIN intervalLiteral ]
    [ SUBSET subsetItem [, subsetItem ]* ]
    DEFINE variable AS condition [, variable AS condition ]*
    )
```

MATCH_RECOGNIZE子句的语法元素定义如下：

-PARTITION BY [可选]：定义分区列。该子句为可选子句。如果未定义，则使用并行度1。

-ORDER BY [可选]：定义数据流中事件的顺序。ORDER BY子句为可选子句，如果忽略则使用非确定性排序。由于事件顺序在模式匹配中很重要，因此大多数情况下应指定该子句。

-MEASURES [可选]：指定匹配成功的事件的属性值。

-ONE ROW PER MATCH | ALL ROWS PER MATCH [可选]：定义如何输出结果。ONE ROW PER MATCH表示每次匹配只输出一行，ALL ROWS PER MATCH表示每次匹配的每一个事件输出一行。

-AFTER MATCH [可选]：指定从何处开始对下一个模式匹配进行匹配成功后的处理。

-PATTERN： 将匹配模式定义为正则表达式格式。PATTERN子句中可使用以下运算符： 连接运算符，量词运算符(*，+，?，{n}，{n,}, {n,m}, {,m}),分支运算符（使用竖线‘|’），以及异运算符（ ‘{- -}’）。

-WITHIN [可选]：当且仅当匹配发生在指定时间内，则输出模式子句匹配。

-SUBSET [可选]：将DEFINE子句中定义的一个或多个关联变量组合在一起。

-DEFINE：指定boolean条件，该条件定义了PATTERN子句中使用的变量。

此外，还支持以下函数：

-MATCH_NUMBER()：可用于MEASURES子句中，为同一成功匹配的每一行分配相同编号。

-CLASSIFIER()：可用于MEASURES子句中，以指示匹配的行与变量之间的映射关系。

-FIRST()和LAST()：可用于MEASURES子句中，返回在映射到模式变量的行集的第一行或最后一行中评估的表达式的值。

-NEXT()和PREV()：可用于DEFINE子句中，通过分区中的前一行或下一行来评估表达式。

-RUNNING和FINAL关键字：可用于确定聚合的所需语义。RUNNING可用于MEASURES和DEFINE子句中，而FINAL只能用于MEASURES子句中。

-聚合函数(COUNT，SUM，AVG，MAX，MIN)：这些聚合函数可用于MEASURES子句和DEFINE子句中。

查询示例

以下查询发现股票价格数据流中的V型模式。

```
SELECT *
   FROM MyTable
   MATCH_RECOGNIZE (
     ORDER BY rowtime
     MEASURES
       STRT.name as s_name,
       LAST(DOWN.name) as down_name,
       LAST(UP.name) as up_name
     ONE ROW PER MATCH
     PATTERN (STRT DOWN+ UP+)
     DEFINE
       DOWN AS DOWN.v < PREV(DOWN.v),
       UP AS UP.v > PREV(UP.v)
   )
```

在以下查询中，聚合函数AVG应用于A和C相关变量组成的SUBSET E的MEASURES子句中。

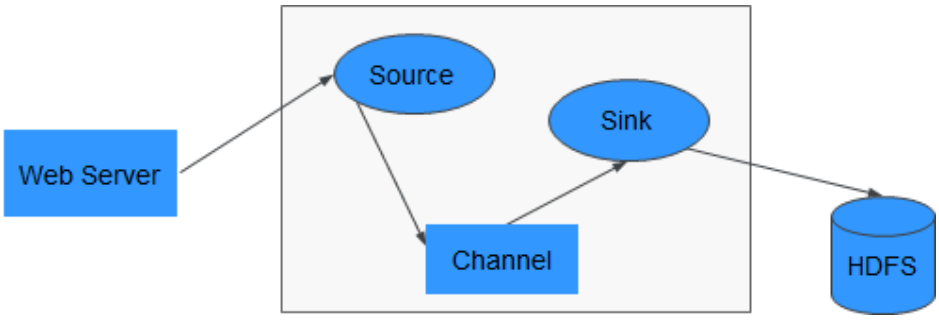
```
SELECT *
   FROM Ticker
   MATCH_RECOGNIZE (
     MEASURES
       AVG(E.price) AS avgPrice
     ONE ROW PER MATCH
     AFTER MATCH SKIP PAST LAST ROW
     PATTERN (A B+ C)
     SUBSET E = (A,C)
     DEFINE
       A AS A.price < 30,
       B AS B.price < 20,
       C AS C.price < 30
   )
```

1.4.6 Flume

1.4.6.1 Flume 基本原理

Flume是一个高可用、高可靠，分布式的海量日志采集、聚合和传输的系统。Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。其中Flume-NG是Flume的一个分支，其目的是要明显简单，体积更小，更容易部署，其最基本的架构如下图所示：

图 1-28 Flume-NG 架构



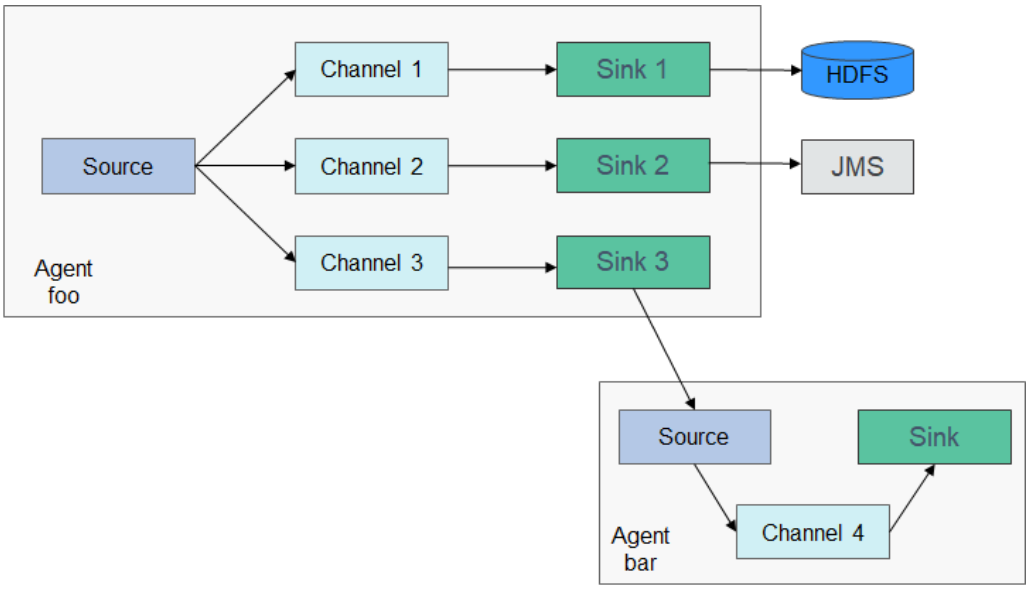
Flume-NG由一个个Agent来组成，而每个Agent由Source、Channel、Sink三个模块组成，其中Source负责接收数据，Channel负责数据的传输，Sink则负责数据向下一端的发送。

表 1-4 模块说明

名称	说明
Source	Source负责接收数据或通过特殊机制产生数据，并将数据批量放到一个或多个Channel。Source的类型有数据驱动和轮询两种。 典型的Source类型如下： ● 和系统集成的Sources：Syslog、Netcat。 ● 自动生成事件的Sources：Exec、SEQ。 ● 用于Agent和Agent之间通信的IPC Sources：Avro。 Source必须至少和一个Channel关联。
Channel	Channel位于Source和Sink之间，用于缓存来自Source的数据，当Sink成功将数据发送到下一跳的Channel或最终目的地时，数据从Channel移除。 Channel提供的持久化水平与Channel的类型相关，有以下三类： ● Memory Channel：非持久化。 ● File Channel：基于WAL（预写式日志Write-Ahead Logging）的持久化实现。 ● JDBC Channel：基于嵌入Database的持久化实现。 Channel支持事务，可提供较弱的顺序保证，可以和任何数量的Source和Sink工作。
Sink	Sink负责将数据传输到下一跳或最终目的，成功完成后将数据从Channel移除。 典型的Sink类型如下： ● 存储数据到最终目的终端Sink，比如：HDFS、HBase。 ● 自动消耗的Sink，比如：Null Sink。 ● 用于Agent间通信的IPC sink：Avro。 Sink必须作用于一个确切的Channel。

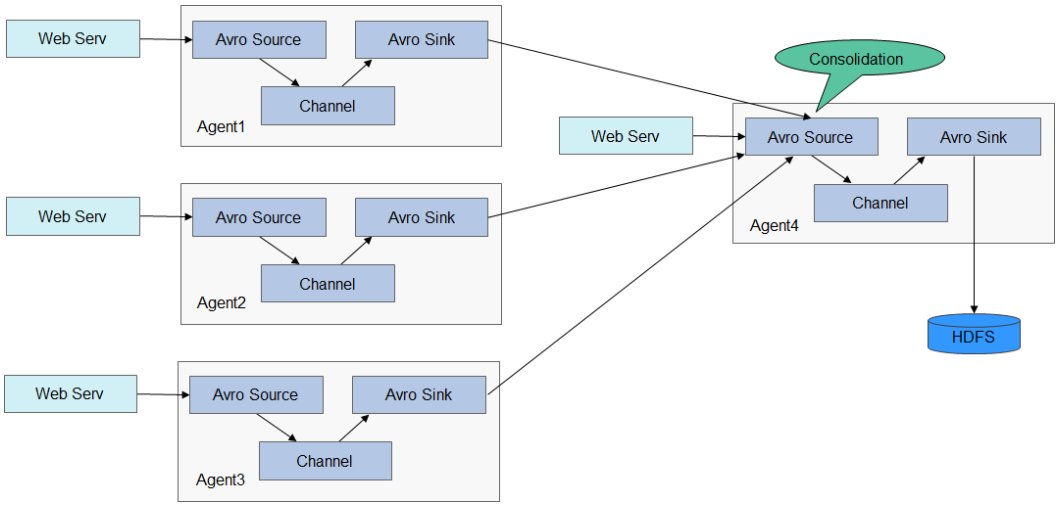
Flume也可以配置成多个Source、Channel、Sink，如图1-29所示：

图 1-29 Flume 结构图



Flume的可靠性基于Agent间事务的交换，下一个Agent down掉，Channel可以持久化数据，Agent恢复后再传输。Flume的可用性则基于内建的Load Balancing和Failover机制。Channel及Agent都可以配多个实体，实体之间可以使用负载分担等策略。每个Agent为一个JVM进程，同一台服务器可以有多个Agent。收集节点（Agent1，2，3）负责处理日志，汇聚节点（Agent4）负责写入HDFS，每个收集节点的Agent可以选择多个汇聚节点，这样可以实现负载均衡。

图 1-30 Flume 级联结构图



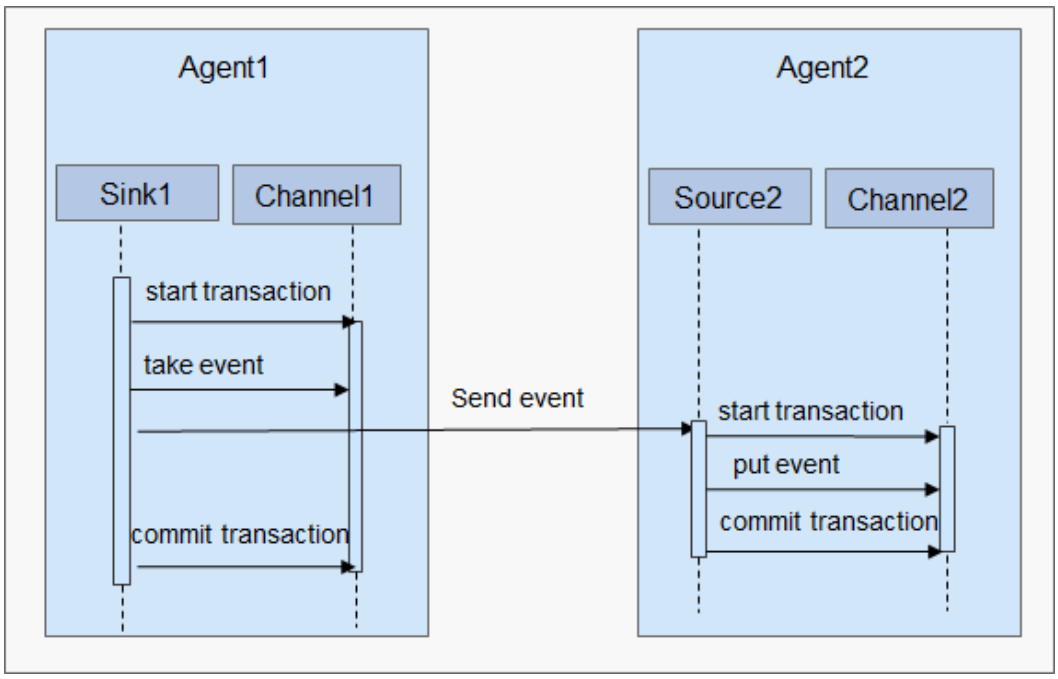
Flume的架构和详细原理介绍，请参见：https://flume.apache.org/releases/1.9.0.html。

Flume 原理

Agent之间的可靠性

Agent之间数据交换流程如图1-31所示。

图 1-31 Agent 数据传输流程



1. Flume采用基于Transactions的方式保证数据传输的可靠性，当数据从一个Agent流向另外一个Agent时，两个Transactions已经开始生效。发送Agent的Sink首先从Channel取出一条消息，并且将该消息发送给另外一个Agent。如果接受消息的Agent成功地接受并处理消息，那么发送Agent将会提交Transactions，标识一次数据传输成功可靠地完成。
2. 当接收Agent接受到发送Agent发送的消息时，开始一个新的Transactions，当该数据被成功处理（写入Channel中），那么接收Agent提交该Transactions，并向发送Agent发送成功响应。
3. 如果在某次提交（commit）之前，数据传输出现了失败，将会再次开始上一次Transactions，并将上次发送失败的数据重新传输。因为commit操作已经将Transactions写入了磁盘，那么在进程故障退出并恢复业务之后，仍然可以继续上次的Transactions。

1.4.6.2 Flume 与其他组件的关系

Flume 与 HDFS 的关系

当用户配置HDFS作为Flume的Sink时，HDFS就作为Flume的最终数据存储系统，Flume将传输的数据全部按照配置写入HDFS中。

Flume 与 HBase 的关系

当用户配置HBase作为Flume的Sink时，HBase就作为Flume的最终数据存储系统，Flume将传输的数据全部按照配置写入HBase中。

1.4.6.3 Flume 开源增强特性

Flume 开源增强特性

- 提升传输速度。可以配置将指定的行数作为一个Event，而不仅是一行，提高了代码的执行效率以及减少写入磁盘的次数。
- 传输超大二进制文件。Flume根据当前内存情况，自动调整传输超大二进制文件的内存占用情况，不会导致Out of Memory（OOM）的出现。
- 支持定制传输前后准备工作。Flume支持定制脚本，指定在传输前或者传输后执行指定的脚本，用于执行准备工作。
- 管理客户端告警。Flume通过MonitorServer接收Flume客户端告警，并上报Manager告警管理中心。

1.4.7 HBase

1.4.7.1 HBase 基本原理

数据存储使用HBase来承接，HBase是一个开源的、面向列（Column-Oriented）、适合存储海量非结构化数据或半结构化数据的、具备高可靠性、高性能、可灵活扩展伸缩的、支持实时数据读写的分布式存储系统。更多关于HBase的信息，请参见：**https://hbase.apache.org/**。

存储在HBase中的表的典型特征：

- 大表（BigTable）：一个表可以有上亿行，上百万列
- 面向列：面向列（族）的存储、检索与权限控制
- 稀疏：表中为空（null）的列不占用存储空间

MRS服务的HBase组件支持计算存储分离，数据可以存储在低成本的云存储服务中，包含对象存储服务，并支持跨AZ数据备份。并且MRS服务支持HBase组件的二级索引，支持为列值添加索引，提供使用原生的HBase接口的高性能基于列过滤查询的能力。

HBase 结构

HBase集群由主备Master进程和多个RegionServer进程组成。如**图 HBase结构**所示。

图 1-32 HBase 结构

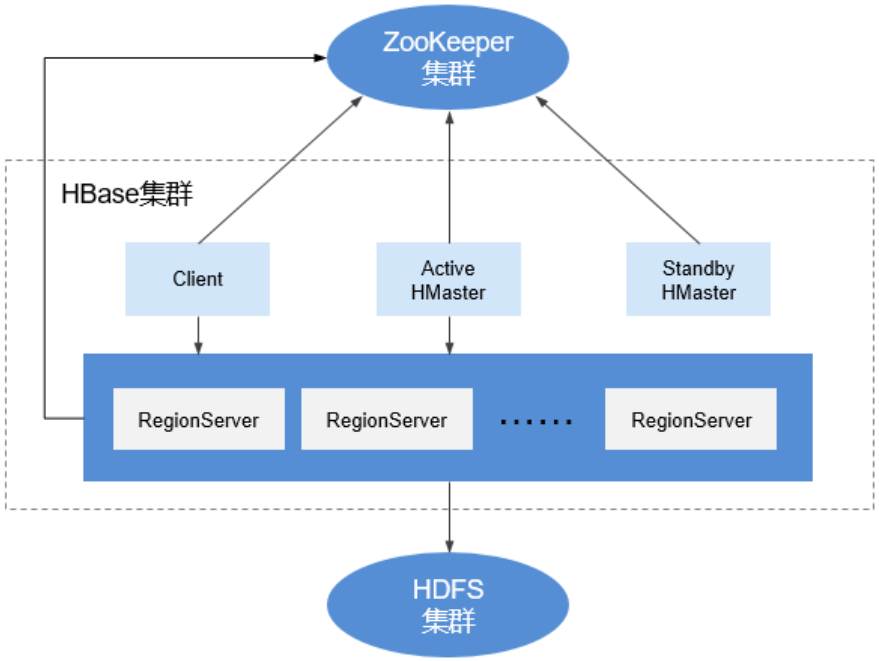


表 1-5 模块说明

名称	描述
Master	又叫HMaster，在HA模式下，包含主用Master和备用Master。 • 主用Master：负责HBase中RegionServer的管理，包括表的增删改查；RegionServer的负载均衡，Region分布调整；Region分裂以及分裂后的Region分配；RegionServer失效后的Region迁移等。 • 备用Master：当主用Master故障时，备用Master将取代主用Master对外提供服务。故障恢复后，原主用Master降为备用。
Client	Client使用HBase的RPC机制与Master、RegionServer进行通信。Client与Master进行管理类通信，与RegionServer进行数据操作类通信。
RegionServer	RegionServer负责提供表数据读写等服务，是HBase的数据处理和计算单元。 RegionServer一般与HDFS集群的DataNode部署在一起，实现数据的存储功能。
ZooKeeper集群	ZooKeeper为HBase集群中各进程提供分布式协作服务。各RegionServer将自己的信息注册到ZooKeeper中，主用Master据此感知各个RegionServer的健康状态。
HDFS集群	HDFS为HBase提供高可靠的文件存储服务，HBase的数据全部存储在HDFS中。

HBase 原理

- **HBase数据模型**

  HBase以表的形式存储数据，数据模型如图 HBase数据模型所示。表中的数据划分为多个Region，并由Master分配给对应的RegionServer进行管理。

  每个Region包含了表中一段RowKey区间范围内的数据，HBase的一张数据表开始只包含一个Region，随着表中数据的增多，当一个Region的大小达到容量上限后会分裂成两个Region。您可以在创建表时定义Region的RowKey区间，或者在配置文件中定义Region的大小。

  图 1-33 HBase 数据模型

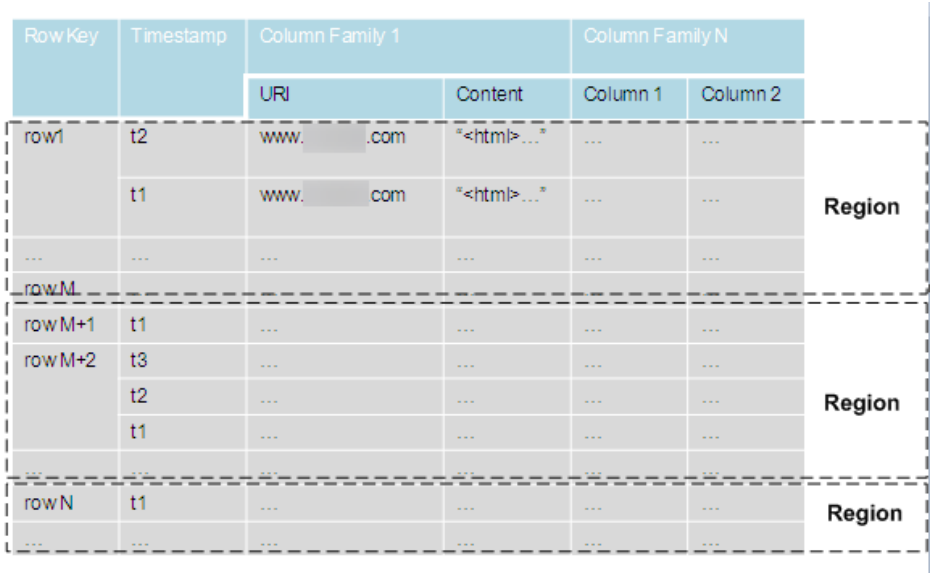


表 1-6 概念介绍

名称	描述
RowKey	行键，相当于关系表的主键，每一行数据的唯一标识。字符串、整数、二进制串都可以作为RowKey。所有记录按照RowKey排序后存储。
Timestamp	每次数据操作对应的时间戳，数据按时间戳区分版本，每个Cell的多个版本的数据按时间倒序存储。
Cell	HBase最小的存储单元，由Key和Value组成。Key由row、column family、column qualifier、timestamp、type、MVCC version这6个字段组成。Value就是对应存储的二进制数据对象。
Column Family	列族，一个表在水平方向上由一个或多个Column Family组成。一个CF（Column Family）可以由任意多个Column组成。Column是CF下的一个标签，可以在写入数据时任意添加，因此CF支持动态扩展，无需预先定义Column的数量和类型。HBase中表的列非常稀疏，不同行的列的个数和类型都可以不同。此外，每个CF都有独立的生存周期（TTL）。可以只对行上锁，对行的操作始终是原始的。

名称	描述
Column	列，与传统的数据库类似，HBase的表中也有列的概念，列用于表示相同类型的数据。

- **RegionServer数据存储**

  RegionServer主要负责管理由HMaster分配的Region，RegionServer的数据存储结构如图 RegionServer的数据存储结构所示。

  图 1-34 RegionServer 的数据存储结构

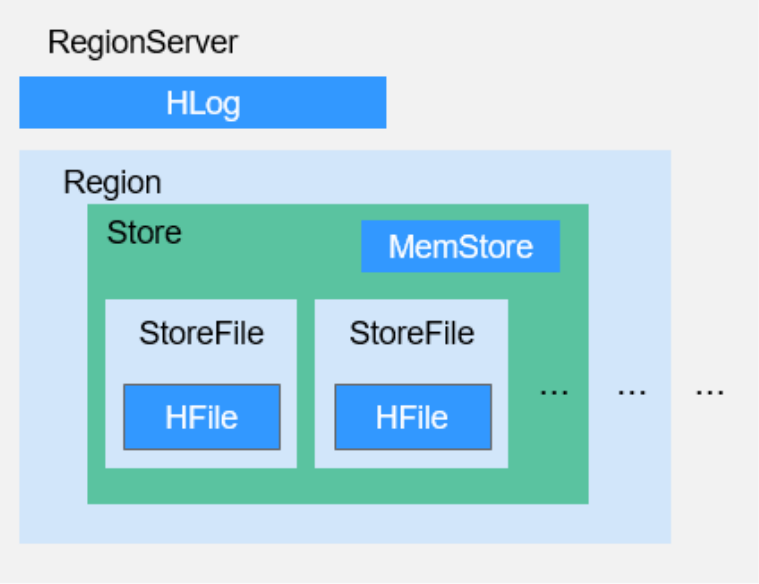


  图 RegionServer的数据存储结构中Region的各部分的说明如表 Region结构说明所示。

  表 1-7 Region 结构说明

名称	描述
Store	一个Region由一个或多个Store组成，每个Store对应图 HBase数据模型中的一个Column Family。
MemStore	一个Store包含一个MemStore，MemStore缓存客户端向Region插入的数据，当RegionServer中的MemStore大小达到配置的容量上限时，RegionServer会将MemStore中的数据“flush”到HDFS中。
StoreFile	MemStore的数据flush到HDFS后成为StoreFile，随着数据的插入，一个Store会产生多个StoreFile，当StoreFile的个数达到配置的最大值时，RegionServer会将多个StoreFile合并为一个大的StoreFile。
HFile	HFile定义了StoreFile在文件系统中的存储格式，它是当前HBase系统中StoreFile的具体实现。

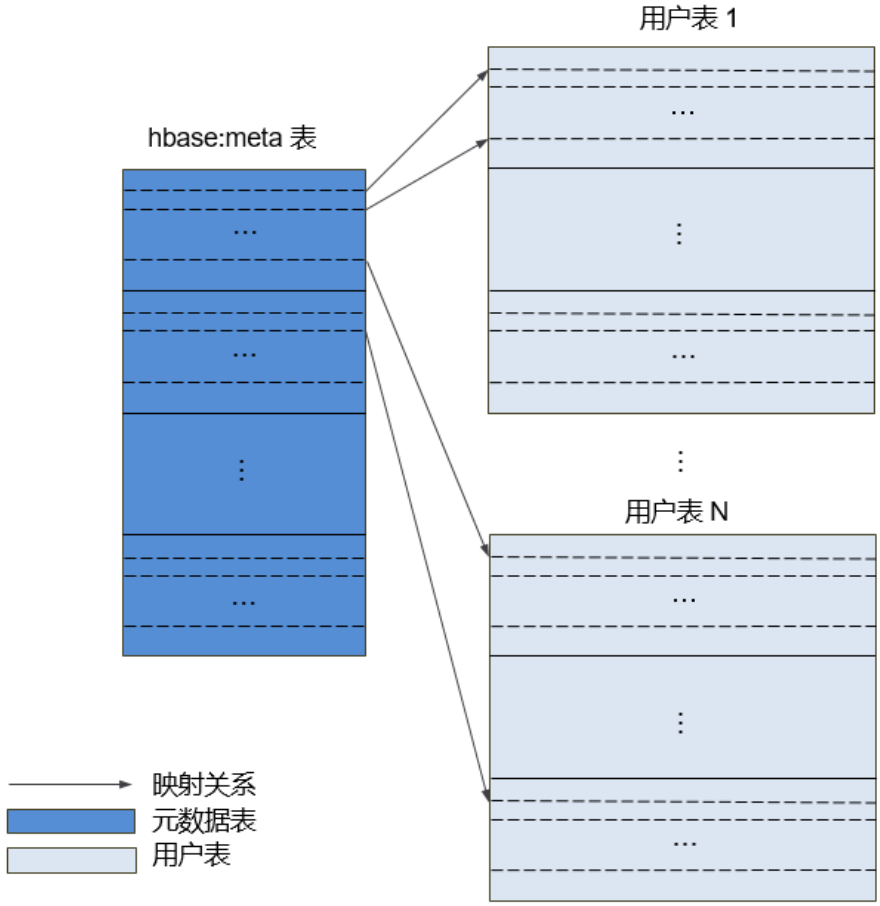
名称	描述
HLog	HLog日志保证了当RegionServer故障的情况下用户写入的数据不丢失，RegionServer的多个Region共享一个相同的HLog。

- **元数据表**

  元数据表是HBase中一种特殊的表，用来帮助Client定位到具体的Region。元数据表包括“hbase:meta”表，用来记录用户表的Region信息，例如，Region位置、起始RowKey及结束RowKey等信息。

  元数据表和用户表的映射关系如图 元数据表和用户表的映射关系所示。

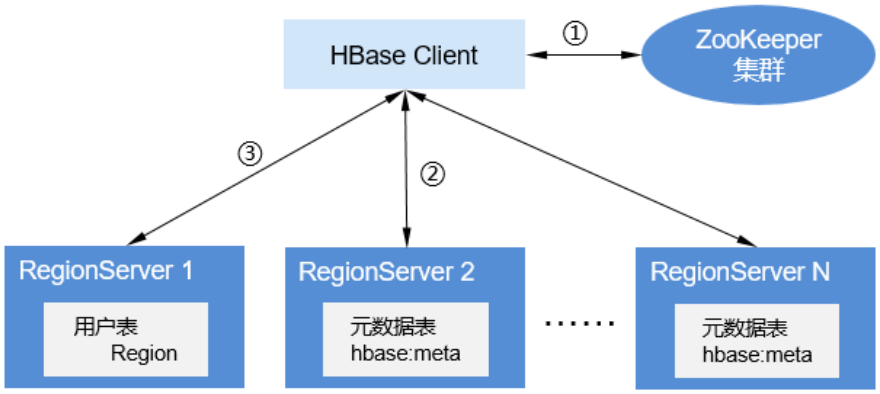
  图 1-35 元数据表和用户表的映射关系



- **数据操作流程**

  HBase数据操作流程如图 数据操作流程所示。

图 1-36 数据操作流程



a. 对HBase进行增、删、改、查数据操作时，HBase Client首先连接ZooKeeper获得“hbase:meta”表所在的RegionServer的信息(涉及namespace级别修改的，比如创建表、删除表需要访问HMaster更新meta信息)。

b. HBase Client连接到包含对应的“hbase:meta”表的Region所在的RegionServer，并获得相应的用户表的Region所在的RegionServer位置信息。

c. HBase Client连接到对应的用户表Region所在的RegionServer，并将数据操作命令发送给该RegionServer，RegionServer接收并执行该命令从而完成本次数据操作。

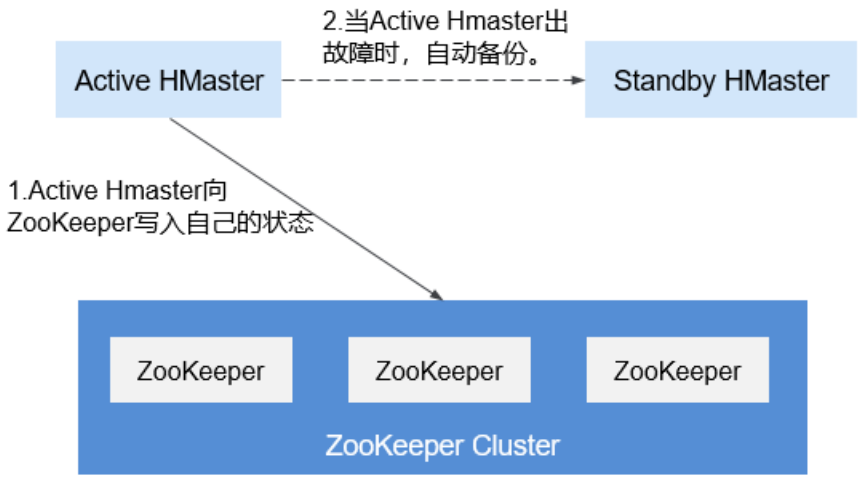
为了提升数据操作的效率，HBase Client会在内存中缓存“hbase:meta”和用户表Region的信息，当应用程序发起下一次数据操作时，HBase Client会首先从内存中获取这些信息；当未在内存缓存中找到对应数据信息时，HBase Client会重复上述操作。

1.4.7.2 HBase HA 方案介绍

HBase HA 原理与实现方案

HBase中的HMaster负责region分配，当regionserver服务停止后，HMaster把相应region迁移到其他RegionServer。为了解决HMaster单点故障导致HBase正常功能受到影响的问题，引入HMaster HA模式。

图 1-37 HMaster 高可用性实现架构



HMaster高可用性架构通过在ZooKeeper集群创建ephemeral zookeeper node实现的。

当HMaster两个节点启动时都会尝试在ZooKeeper集群上创建一个znode节点master，先创建的成为Active HMaster，后创建的成为Standby HMaster。

Standby HMaster会在master节点添加监听事件。如果主节点服务停止，就会和zooKeeper集群失去联系，session过期之后master节点会消失。Standby节点通过监听事件（watch event）感知到节点消失，会去创建master节点自己成为Active HMaster，主备倒换完成。如果后续停止服务的节点重新启动，发现master节点已经存在，则进入Standby模式，并对master znode创建监听事件。

当客户端访问HBase时，会首先通过ZooKeeper上的master节点信息找到HMaster的地址，然后与Active HMaster进行连接。

1.4.7.3 HBase 与其他组件的关系

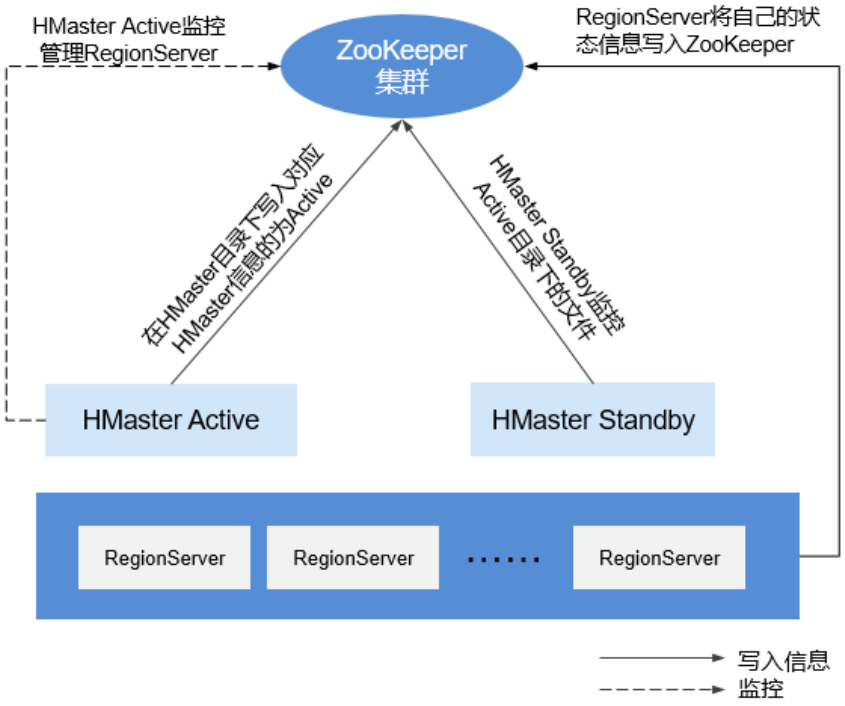
HBase 和 HDFS 的关系

HDFS是Apache的Hadoop项目的子项目，HBase利用Hadoop HDFS作为其文件存储系统。HBase位于结构化存储层，Hadoop HDFS为HBase提供了高可靠性的底层存储支持。除了HBase产生的一些日志文件，HBase中的所有数据文件都可以存储在Hadoop HDFS文件系统上。

HBase 和 ZooKeeper 的关系

HBase和ZooKeeper的关系如**图 ZooKeeper和HBase的关系**所示。

图 1-38 HBase 和 ZooKeeper 的关系



1. HRegionServer以Ephemeral node的方式注册到ZooKeeper中。其中ZooKeeper存储HBase的如下信息：HBase元数据、HMaster地址。
2. HMaster通过ZooKeeper随时感知各个HRegionServer的健康状况，以便进行控制管理。
3. HBase也可以部署多个HMaster，类似HDFS NameNode，当HMaster主节点出现故障时，HMaster备用节点会通过ZooKeeper获取主HMaster存储的整个HBase集群状态信息。即通过ZooKeeper实现避免HBase单点故障问题的问题。

1.4.7.4 HBase 开源增强特性

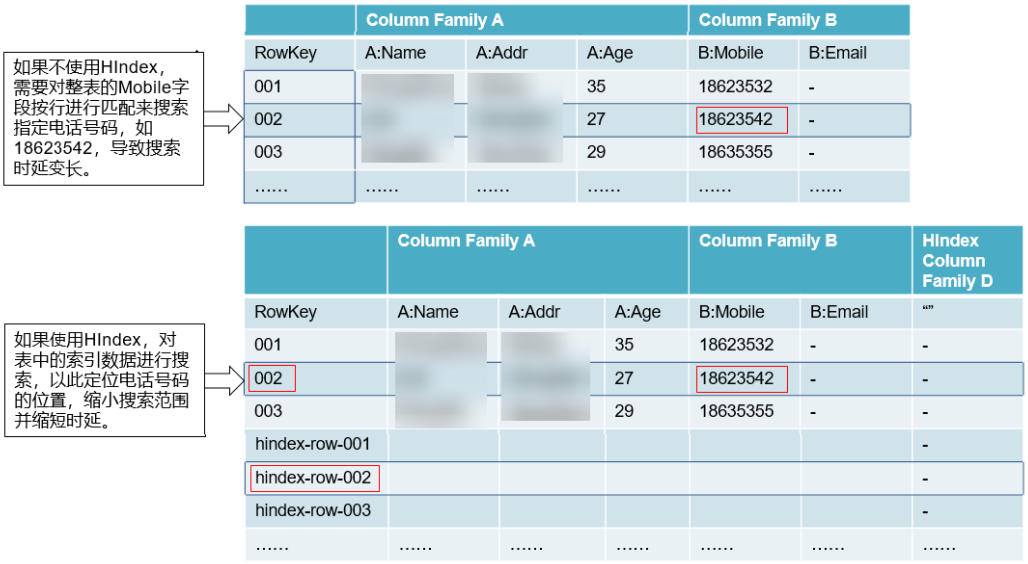
HBase 开源增强特性：HIndex

HBase是一个Key-Value类型的分布式存储数据库。每张表的数据按照RowKey的字典顺序排序，因此，如果按照某个指定的RowKey去查询数据，或者指定某一个RowKey范围去扫描数据时，HBase可以快速定位到需要读取的数据位置，从而可以高效地获取到所需要的数据。

在实际应用中，很多场景是查询某一个列值为“XXX”的数据。HBase提供了Filter特性去支持这样的查询，它的原理是：按照RowKey的顺序，去遍历所有可能的数据，再依次去匹配那一列的值，直到获取到所需要的数据。可以看出，可能只是为了获取一行数据，它却扫描了很多不必要的数据。因此，如果对于这样的查询请求非常频繁并且对查询性能要求较高，使用Filter无法满足这个需求。

这就是HBase HIndex产生的背景。HIndex为HBase提供了按照某些列的值进行索引的能力。

图 1-39 HIndex



- 索引数据不支持滚动升级。
- 组合索引限制。
  - 用户必须在单次mutation中输入或删除参与组合索引的所有列。否则会导致不一致问题。

    索引：IDX1=>cf1:[q1->datatype],[q2];cf2:[q2->datatype]

    正确的写操作：

    ```
    Put put = new Put(Bytes.toBytes("row"));
    put.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q1"), Bytes.toBytes("valueA"));
    put.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q2"), Bytes.toBytes("valueB"));
    put.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q2"), Bytes.toBytes("valueC"));
    table.put(put);
    ```

    错误的写操作：

    ```
    Put put1 = new Put(Bytes.toBytes("row"));
    put1.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q1"), Bytes.toBytes("valueA"));
    table.put(put1);
    Put put2 = new Put(Bytes.toBytes("row"));
    put2.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q2"), Bytes.toBytes("valueB"));
    table.put(put2);
    Put put3 = new Put(Bytes.toBytes("row"));
    put3.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q2"), Bytes.toBytes("valueC"));
    table.put(put3);
    ```

  - 使用组合条件查询，仅支持组合索引列包含过滤条件的查询，或者不指定StartRow和StopRow的部分索引列的查询。

    索引：IDX1=>cf1:[q1->datatype],[q2];cf2:[q1->datatype]

    正确的查询操作：

    ```
    scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',>=,'binary:valueA',true,true) AND
    SingleColumnValueFilter('cf1','q2',>=,'binary:valueB',true,true) AND
    SingleColumnValueFilter('cf2','q1',>=,'binary:valueC',true,true) "}

    scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',=,'binary:valueA',true,true) AND
    SingleColumnValueFilter('cf1','q2',>=,'binary:valueB',true,true)" }

    scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',>=,'binary:valueA',true,true) AND
    SingleColumnValueFilter('cf1','q2',>=,'binary:valueB',true,true) AND
    SingleColumnValueFilter('cf2','q1',>=,'binary:valueC',true,true)",STARTROW=>'row001',STOPROW
    =>'row100'}
    ```

错误的查询操作：

```
scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',>=,'binary:valueA',true,true) AND
SingleColumnValueFilter('cf1','q2',>=,'binary:valueB',true,true) AND
SingleColumnValueFilter('cf2','q1',>=,'binary:valueC',true,true)  AND
SingleColumnValueFilter('cf2','q2',>=,'binary:valueD',true,true)"}

scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',=,'binary:valueA',true,true) AND
SingleColumnValueFilter('cf2','q1',>=,'binary:valueC',true,true)" }

scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',=,'binary:valueA',true,true) AND
SingleColumnValueFilter('cf2','q2',>=,'binary:valueD',true,true)" }

scan 'table', {FILTER=>"SingleColumnValueFilter('cf1','q1',=,'binary:valueA',true,true) AND
SingleColumnValueFilter('cf1','q2',>=,'binary:valueB',true,true)" ,STARTROW=>'row001',STOPROW
=>'row100' }
```

- 用户不要明确地为有索引数据的表配置任何分裂策略。
- 不支持其他的mutation操作，如increment和append。
- 不支持maxVersions>1的列的索引。
- 不支持一行数据索引列的更新操作。

  索引1：IDX1=>cf1:[q1->datatype],[q2];cf2:[q1->datatype]

  索引2：IDX2=>cf2:[q2->datatype]

  正确的更新操作：

```
Put put1 = new Put(Bytes.toBytes("row"));
put1.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q1"), Bytes.toBytes("valueA"));
put1.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q2"), Bytes.toBytes("valueB"));
put1.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q1"), Bytes.toBytes("valueC"));
put1.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q2"), Bytes.toBytes("valueD"));
table.put(put1);

Put put2 = new Put(Bytes.toBytes("row"));
put2.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q3"), Bytes.toBytes("valueE"));
put2.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q3"), Bytes.toBytes("valueF"));
table.put(put2);
```

  错误的更新操作：

```
Put put1 = new Put(Bytes.toBytes("row"));
put1.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q1"), Bytes.toBytes("valueA"));
put1.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q2"), Bytes.toBytes("valueB"));
put1.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q1"), Bytes.toBytes("valueC"));
put1.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q2"), Bytes.toBytes("valueD"));
table.put(put1);

Put put2 = new Put(Bytes.toBytes("row"));
put2.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q1"), Bytes.toBytes("valueA_new"));
put2.addColumn(Bytes.toBytes("cf1"), Bytes.toBytes("q2"), Bytes.toBytes("valueB_new"));
put2.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q1"), Bytes.toBytes("valueC_new"));
put2.addColumn(Bytes.toBytes("cf2"), Bytes.toBytes("q2"), Bytes.toBytes("valueD_new"));
table.put(put2);
```

- 添加索引的表不应拥有大于32KB的值。
- 当由于列族级TTL（生存周期）过期而导致用户数据删除时，对应的索引数据不会立即删除。索引数据会在进行major compaction操作时被删除。
- 用户列族的TTL在索引创建后不能修改。
  - 如果在创建索引之后，列族的TTL值变大，应该删除并重新创建该索引。否则，一些已经生成的索引数据会先于用户数据被删除。
  - 如果在创建索引之后，列族的TTL值变小。索引数据会晚于用户数据被删除。
- 索引查询不支持reverse；且查询结果是无序的。
- 索引不支持clone snapshot操作。

- 索引表必须使用HIndexWALPlayer回放日志，不支持WALPlayer回放日志。

```
hbase org.apache.hadoop.hbase.hindex.mapreduce.HIndexWALPlayer
Usage: WALPlayer [options] <wal inputdir> <tables> [<tableMappings>]
Read all WAL entries for <tables>.
If no tables ("") are specific, all tables are imported.
(Careful, even -ROOT- and hbase:meta entries will be imported in that case.)
Otherwise <tables> is a comma separated list of tables.

The WAL entries can be mapped to new set of tables via <tableMapping>.
<tableMapping> is a command separated list of targettables.
If specified, each table in <tables> must have a mapping.

By default WALPlayer will load data directly into HBase.
To generate HFiles for a bulk data load instead, pass the option:
  -Dwal.bulk.output=/path/for/output
  (Only one table can be specified, and no mapping is allowed!)
Other options: (specify time range to WAL edit to consider)
  -Dwal.start.time=[date|ms]
  -Dwal.end.time=[date|ms]
For performance also consider the following options:
  -Dmapreduce.map.speculative=false
  -Dmapreduce.reduce.speculative=false
```

- 使用deleteall操作索引表存在性能慢问题。
- 索引表不支持HBCK；如需使用HBCK修复索引表，需先删除索引数据后，再进行修复。

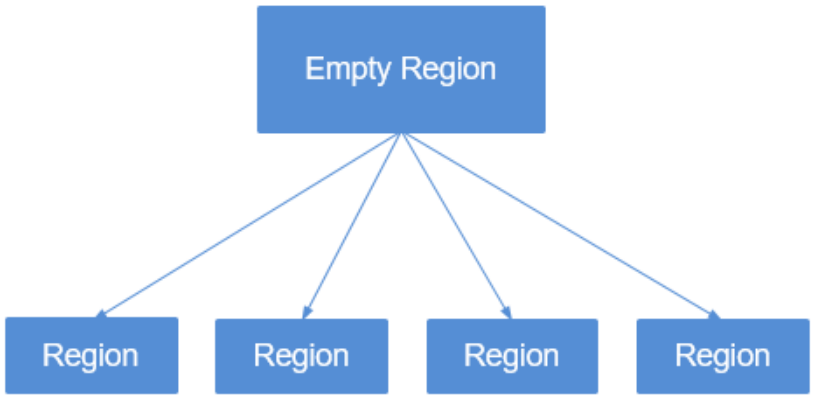
HBase 开源增强特性：支持多点分割

当用户在HBase创建Region预先分割的表时，用户可能不知道数据的分布趋势，所以Region的分割可能不合适，所以当系统运行一段时间后，Region需要重新分割以获得更好的查询性能，HBase只会分割空的Region。

HBase自带的Region分割只有当Region到达设定的Threshold后才会进行分割，这种分割被称为单点分割。

为了实现根据用户的需要动态分割Region以获得更好的性能这一目标，开发了多点分割又称动态分割，即把空的Region预先分割成多个Region。通过预先分割，避免了因为Region空间不足出现Region分割导致性能下降的现象。

图 1-40 多点分割



HBase 开源增强特性：连接数限制

过多的session连接意味着过多的查询和MR任务跑在HBase上，这会导致HBase性能下降以至于导致HBase拒绝服务。通过配置参数来限制客户端连接到HBase服务器端的session数目，来实现HBase过载保护。

HBase 开源增强特性：容灾增强

主备集群之间的容灾能力可以增强HBase数据的高可用性，主集群提供数据服务，备用集群提供数据备份，当主集群出现故障时，备集群可以提供数据服务。相比开源Replication功能，做了如下增强：

1. 备集群白名单功能，只接受指定集群ip的数据推送。
2. 开源版本中replication是基于WAL同步，在备集群回放WAL实现数据备份的。对于BulkLoad，由于没有WAL产生，BulkLoad的数据不会replicate到备集群。通过将BulkLoad操作记录在WAL上，同步至备集群，备集群通过WAL读取BulkLoad操作记录，将对应的主集群的HFile加载到备集群，完成数据的备份。
3. 开源版本中HBase对于系统表ACL做了过滤，ACL信息不会同步至备集群，通过新加一个过滤器 *org.apache.hadoop.hbase.replication.SystemTableWALEntryFilterAllowACL*，允许ACL信息同步至备集群，用户可以通过配置 *hbase.replication.filter.sytemWALEntryFilter*使用该过滤其实现ACL同步。
4. 备集群只读限制，备集群只接受备集群节点内的super user对备集群的HBase进行修改操作，即备集群节点之外的HBase客户端只能对备集群的HBase进行读操作。

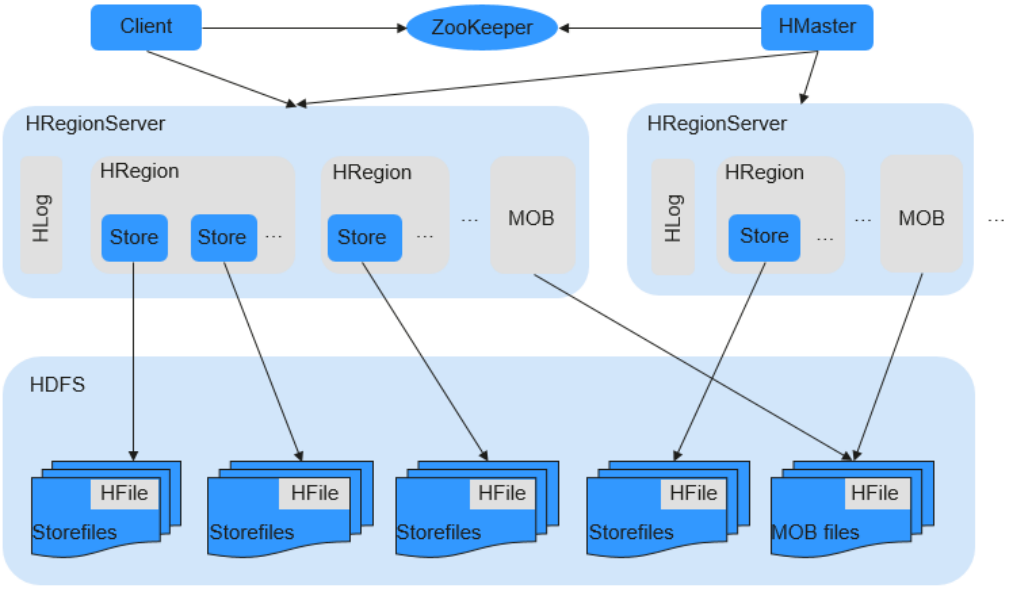
HBase 开源增强特性：HBase MOB

在实际应用中，用户需要存储大大小小的数据，比如图像数据、文档。小于10MB的数据一般都可以存储在HBase上，对于小于100KB的数据，HBase的读写性能是最优的。如果存放在HBase的数据大于100KB甚至到10MB时，插入同样个数的数据文件，其数据量很大，会导致频繁的compaction和split，占用很多CPU，磁盘IO频率很高，性能严重下降。

将MOB数据（即100KB到10MB大小的数据）直接以HFile的格式存储在文件系统上（例如HDFS文件系统），然后把这个文件的地址信息及大小信息作为value存储在普通HBase的store上，通过expiredMobFileCleaner和Sweeper工具集中管理这些文件。这样就可以大大降低HBase的compation和split频率，提升性能。

如**图1-41**所示，图中MOB模块表示存储在HRegion上的mobstore，mobstore存储的是key-value，key即为HBase中对应的key，value对应的就是存储在文件系统上的引用地址以及数据偏移量。读取数据时，mobstore会用自己的scanner，先读取mobstore中的key-value数据对象，然后通过value中的地址及数据大小信息，从文件系统中读取真正的数据。

图 1-41 MOB 数据存储原理



HBase 开源增强特性：HFS

HBase文件存储模块（HBase FileStream，简称HFS）是HBase的独立模块，它作为对HBase与HDFS接口的封装，应用在MRS的上层应用，为上层应用提供文件的存储、读取、删除等功能。

在Hadoop生态系统中，无论是HDFS，还是HBase，均在面对海量文件的存储的时候，在某些场景下，都会存在一些很难解决的问题：

- 如果把海量小文件直接保存在HDFS中，会给NameNode带来极大的压力。
- 由于HBase接口以及内部机制的原因，一些较大的文件也不适合直接保存到HBase中。

HFS的出现，就是为了解决需要在Hadoop中存储海量小文件，同时也要存储一些大文件的混合的场景。简单来说，就是在HBase表中，需要存放大量的小文件（10MB以下），同时又需要存放一些比较大的文件（10MB以上）。

HFS为以上场景提供了统一的操作接口，这些操作接口与HBase的函数接口类似。

HBase 开源增强特性：多 RegionServer 共机部署

HBase支持一个节点部署多个RegionServer，提升HBase资源利用率。

单RegionServer资源利用率低：

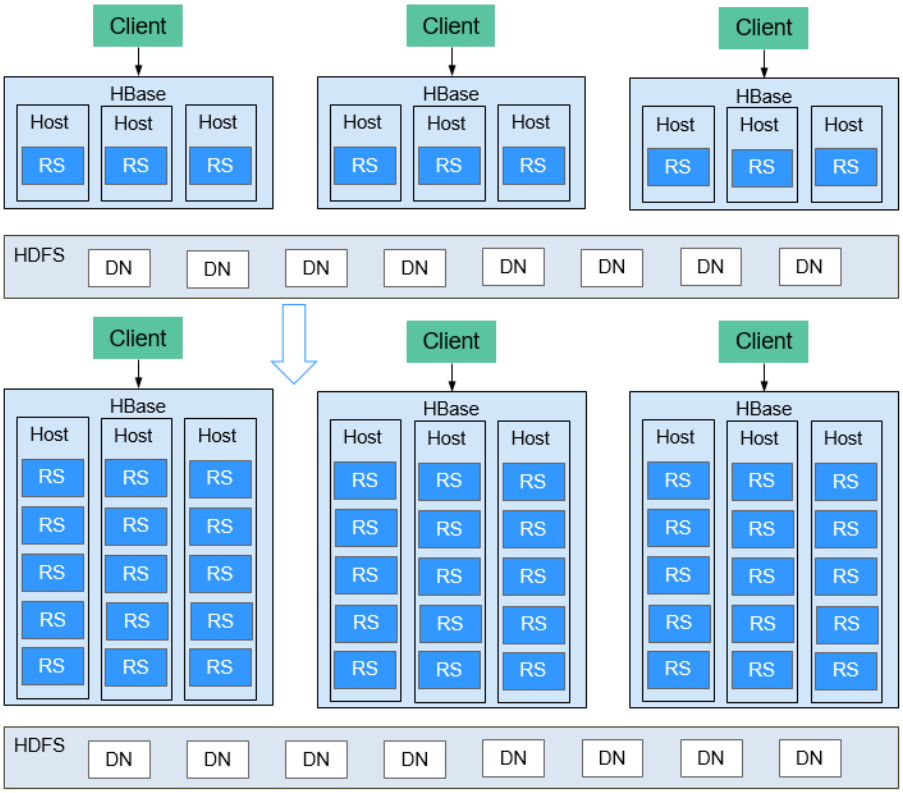
1. 单个RegionServer支持的Region数量有限，无法充分利用内存、CPU资源。
2. 单个RegionServer数据量为20T，两副本为40T，三副本60T，无法用完96T的磁盘。
3. 写入性能差：一台物理机一个RegionServer，只有一个HLog，只能同时写三块盘。

多RegionServer共机部署，提升HBase资源利用率：

1. 一台物理机最多可以部署5个RegionServer，每台物理机上部署的RegionServer个数可以根据需要自由选择。

2. 充分利用内存、磁盘、CPU等资源。
3. 一台物理机最多5个HLog ，可以同时写15块盘，大幅提升写入性能。

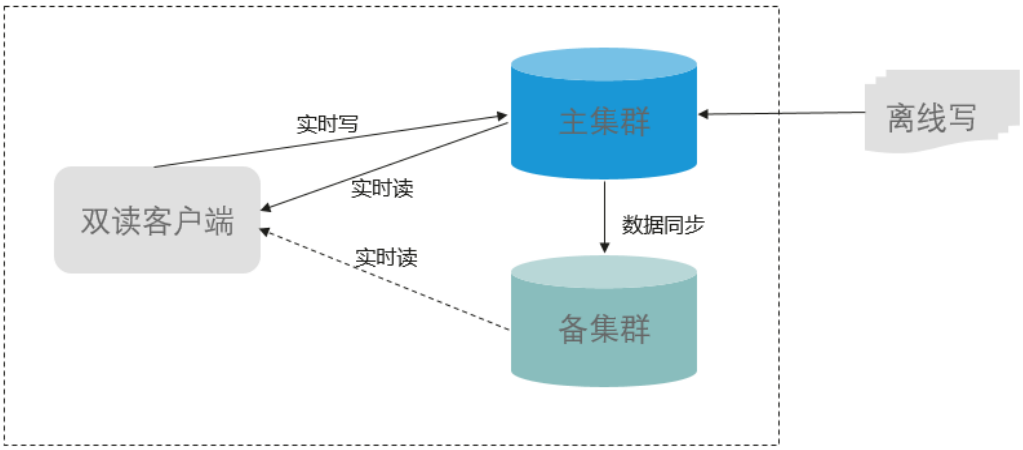
图 1-42 HBase 资源利用率提升



HBase 开源增强特性：HBase 双读

在HBase存储场景下，因为GC、网络抖动、磁盘坏道等原因，很难保证99.9%的查询稳定性。为了满足用户大数据量随机读低毛刺的要求，新增了HBase双读特性。

HBase双读特性是建立在主备集群容灾能力之上，两套集群同时产生毛刺的概率要远远小于一套集群，即采用双集群并发访问的方式，保证查询的稳定性。当用户发起查询请求时，同时查询两个集群的HBase服务，在等待一段时间（最大容忍的毛刺时间）后，如果主集群没有返回结果，则可以使用响应最快的集群数据。原理图如下：



1.4.8 HDFS

1.4.8.1 HDFS 基本原理

HDFS是Hadoop的分布式文件系统（Hadoop Distributed File System），实现大规模数据可靠的分布式读写。HDFS针对的使用场景是数据读写具有“一次写，多次读”的特征，而数据“写”操作是顺序写，也就是在文件创建时的写入或者在现有文件之后的添加操作。HDFS保证一个文件在一个时刻只被一个调用者执行写操作，而可以被多个调用者执行读操作。

HDFS 结构

HDFS包含主、备NameNode和多个DataNode，如图1-43所示。

HDFS是一个Master/Slave的架构，在Master上运行NameNode，而在每一个Slave上运行DataNode，ZKFC需要和NameNode一起运行。

NameNode和DataNode之间的通信都是建立在TCP/IP的基础之上的。NameNode、DataNode、ZKFC和JournalNode能部署在运行Linux的服务器上。

图 1-43 HA HDFS 结构

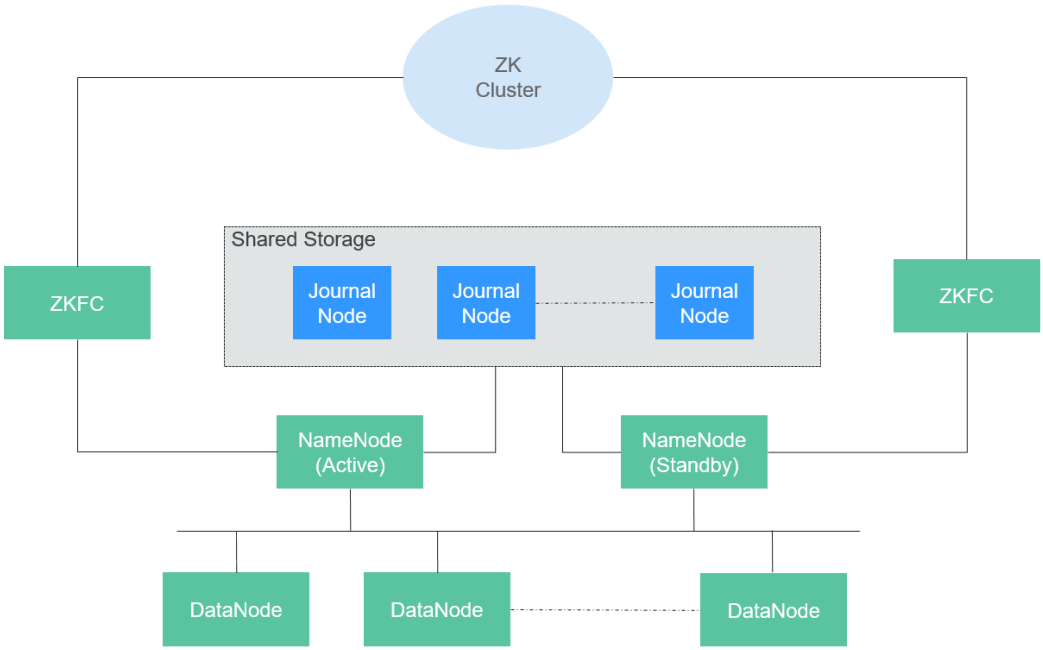


图1-43中各模块的功能说明如表1-8所示。

表 1-8 模块说明

名称	描述
NameNode	用于管理文件系统的命名空间、目录结构、元数据信息以及提供备份机制等，分为： ● Active NameNode：管理文件系统的命名空间、维护文件系统的目录结构树以及元数据信息；记录写入的每个“数据块”与其归属文件的对应关系。 ● Standby NameNode：与Active NameNode中的数据保持同步；随时准备在Active NameNode出现异常时接管其服务。 ● Observer NameNode：与Active NameNode中的数据保持同步，处理来自客户端的读请求。
DataNode	用于存储每个文件的“数据块”数据，并且会周期性地向NameNode报告该DataNode的数据存放情况。
JournalNode	HA集群下，用于同步主备NameNode之间的元数据信息。
ZKFC	ZKFC是需要和NameNode一一对应的服务，即每个NameNode都需要部署ZKFC。它负责监控NameNode的状态，并及时把状态写入ZooKeeper。ZKFC也有选择谁作为Active NameNode的权利。
ZK Cluster	ZooKeeper是一个协调服务，帮助ZKFC执行主NameNode的选举。
HttpFS gateway	HttpFS是个单独无状态的gateway进程，对外提供webHDFS接口，对HDFS使用FileSystem接口对接。可用于不同Hadoop版本间的数据传输，及用于访问在防火墙后的HDFS（HttpFS用作gateway）。

- **HDFS HA架构**

  HA即为High Availability，用于解决NameNode单点故障问题，该特性通过主备的方式为主NameNode提供一个备用者，一旦主NameNode出现故障，可以迅速切换至备NameNode，从而不间断对外提供服务。

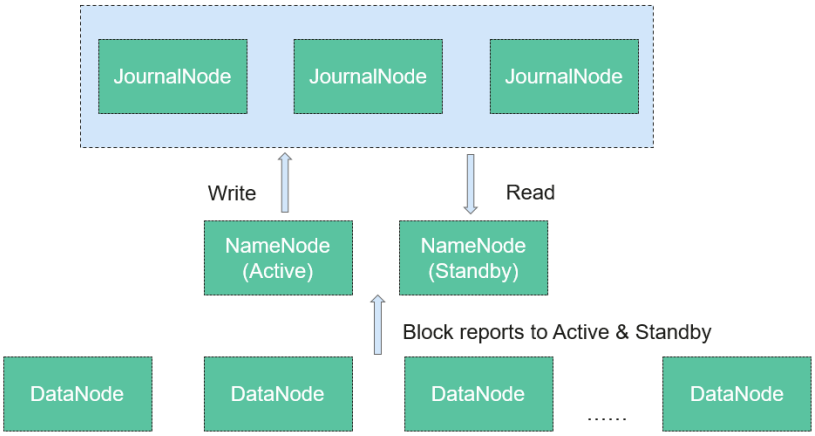
  在一个典型HDFS HA场景中，通常由两个NameNode组成，一个处于Active状态，另一个处于Standby状态。

  为了能实现Active和Standby两个NameNode的元数据信息同步，需提供一个共享存储系统。本版本提供基于QJM（Quorum Journal Manager）的HA解决方案，如图1-44所示。主备NameNode之间通过一组JournalNode同步元数据信息。

  通常配置奇数个（2N+1个）JournalNode，且最少要运行3个JournalNode。这样，一条元数据更新消息只要有N+1个JournalNode写入成功就认为数据写入成功，此时最多容忍N个JournalNode写入失败。比如，3个JournalNode时，最多允许1个JournalNode写入失败，5个JournalNode时，最多允许2个JournalNode写入失败。

  由于JournalNode是一个轻量级的守护进程，可以与Hadoop其它服务共用机器。建议将JournalNode部署在控制节点上，以避免数据节点在进行大数据量传输时引起JournalNode写入失败。

图 1-44 基于 QJM 的 HDFS 架构

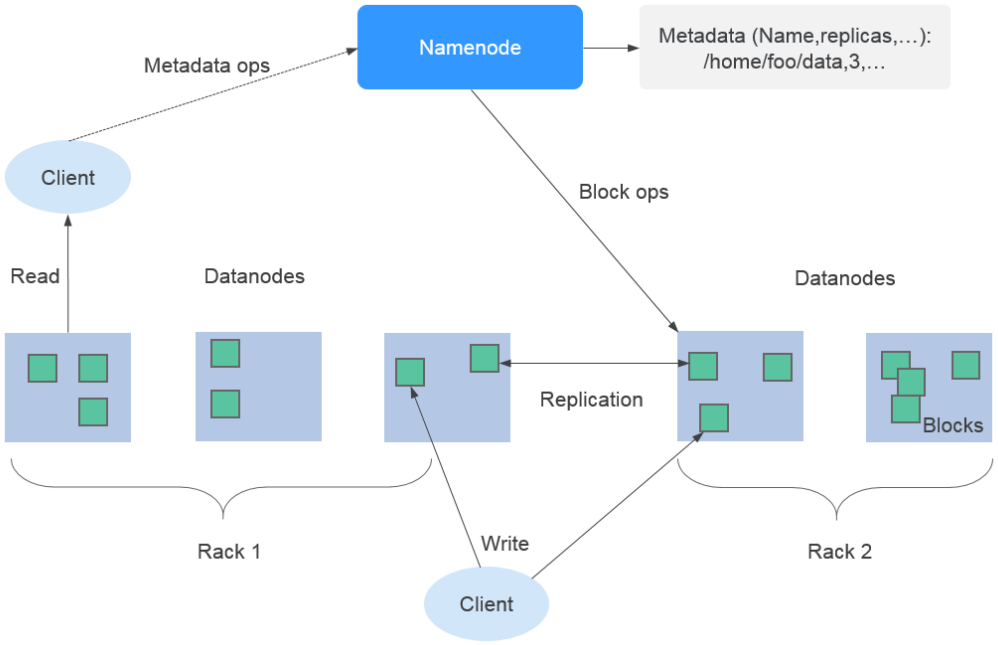


HDFS 原理

MRS使用HDFS的副本机制来保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即共2个副本。HDFS副本数可通过“dfs.replication”参数查询。

- 当MRS集群中Core节点规格选择为非本地盘（hdd）时，若集群中只有一个Core节点，则HDFS默认副本数为1。若集群中Core节点数大于等于2，则HDFS默认副本数为2。
- 当MRS集群中Core节点规格选择为本地盘（hdd）时，若集群中只有一个Core节点，则HDFS默认副本数为1。若集群中有两个Core节点，则HDFS默认副本数为2。若集群中Core节点数大于等于3，则HDFS默认副本数为3。

图 1-45 HDFS 架构



MRS服务的HDFS组件支持以下部分特性：

- HDFS组件支持纠删码，使得数据冗余减少到50%，且可靠性更高，并引入条带化的块存储结构，最大化的利用现有集群单节点多磁盘的能力，使得数据写入性能在引入编码过程后，仍和原来多副本冗余的性能接近。
- 支持HDFS组件上节点均衡调度和单节点内的磁盘均衡调度，有助于扩容节点或扩容磁盘后的HDFS存储性能提升。

关于Hadoop的架构和详细原理介绍，请参见：**http://hadoop.apache.org/**。

1.4.8.2 HDFS HA 方案介绍

HDFS HA 方案背景

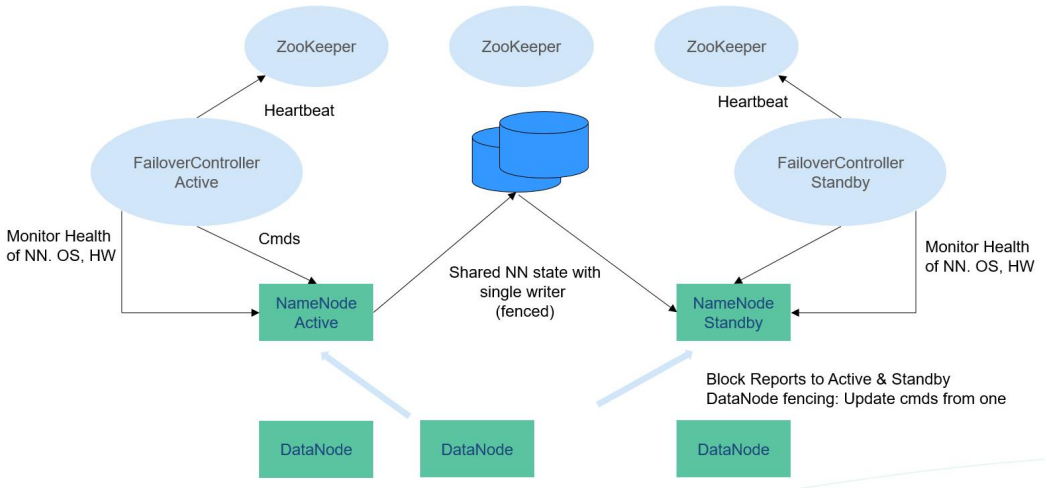
在Hadoop2.0.0之前，HDFS集群中存在单点故障问题。由于每个集群只有一个NameNode，如果NameNode所在机器发生故障，将导致HDFS集群无法使用，除非NameNode重启或者在另一台机器上启动。这在两个方面影响了HDFS的整体可用性：

1. 当异常情况发生时，如机器崩溃，集群将不可用，除非重新启动NameNode。
2. 计划性的维护工作，如软硬件升级等，将导致集群停止工作。

针对以上问题，HDFS高可用性方案通过自动或手动（可配置）的方式，在一个集群中为NameNode启动一个热替换的NameNode备份。当一台机器故障时，可以迅速地自动进行NameNode主备切换。或者当主NameNode节点需要进行维护时，通过MRS集群管理员控制，可以手动进行NameNode主备切换，从而保证集群在维护期间的可用性。有关HDFS自动故障转移功能，请参阅**http://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-hdfs/HDFSHighAvailabilityWithQJM.html#Automatic_Failover**。

HDFS HA 实现方案

图 1-46 典型的 HA 部署方式



在一个典型的HA集群中（如**图1-46**），需要把两个NameNodes配置在两台独立的机器上。在任何一个时间点，只有一个NameNode处于Active状态，另一个处于Standby状态。Active节点负责处理所有客户端操作，Standby节点时刻保持与Active节点同步的状态以便在必要时进行快速主备切换。

为保持Active和Standby节点的数据一致性，两个节点都要与一组称为JournalNode的节点通信。当Active对文件系统元数据进行修改时，会将其修改日志保存到大多数的

JournalNode节点中，例如有3个JournalNode，则日志会保存在至少2个节点中。Standby节点监控JournalNodes的变化，并同步来自Active节点的修改。根据修改日志，Standby节点将变动应用到本地文件系统元数据中。一旦发生故障转移，Standby节点能够确保与Active节点的状态是一致的。这保证了文件系统元数据在故障转移时在Active和Standby之间是完全同步的。

为保证故障转移快速进行，Standby需要时刻保持最新的块信息，为此DataNodes同时向两个NameNodes发送块信息和心跳。

对一个HA集群，保证任何时刻只有一个NameNode是Active状态至关重要。否则，命名空间会分为两部分，有数据丢失和产生其他错误的风险。为保证这个属性，防止“split-brain”问题的产生，JournalNodes在任何时刻都只允许一个NameNode写入。在故障转移时，将变为Active状态的NameNode获得写入JournalNodes的权限，这会有效防止其他NameNode的Active状态，使得切换安全进行。

关于HDFS高可用性方案的更多信息，可参考如下链接：

http://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-hdfs/HDFSHighAvailabilityWithQJM.html

1.4.8.3 HDFS 与其他组件的关系

HDFS 和 HBase 的关系

HDFS是Apache的Hadoop项目的子项目，HBase利用Hadoop HDFS作为其文件存储系统。HBase位于结构化存储层，Hadoop HDFS为HBase提供了高可靠性的底层存储支持。除了HBase产生的一些日志文件，HBase中的所有数据文件都可以存储在Hadoop HDFS文件系统上。

HDFS 和 MapReduce 的关系

- HDFS是Hadoop分布式文件系统，具有高容错和高吞吐量的特性，可以部署在价格低廉的硬件上，存储应用程序的数据，适合有超大数据集的应用程序。
- 而MapReduce是一种编程模型，用于大数据集（大于1TB）的并行运算。在MapReduce程序中计算的数据可以来自多个数据源，如Local FileSystem、HDFS、数据库等。最常用的是HDFS，可以利用HDFS的高吞吐性能读取大规模的数据进行计算。同时在计算完成后，也可以将数据存储到HDFS。

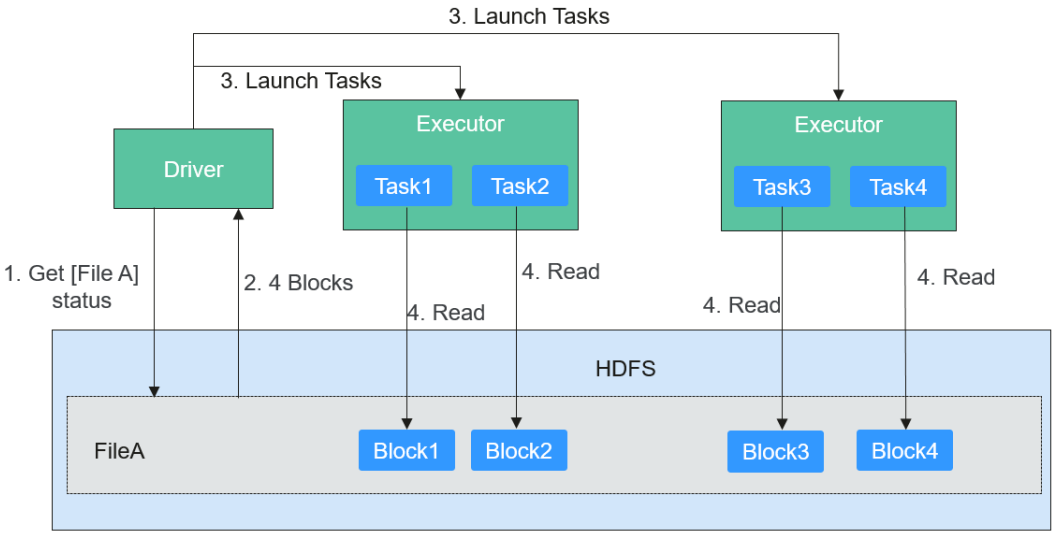
HDFS 和 Spark 的关系

通常，Spark中计算的数据可以来自多个数据源，如Local File、HDFS等。最常用的是HDFS，用户可以一次读取大规模的数据进行并行计算。在计算完成后，也可以将数据存储到HDFS。

分解来看，Spark分成控制端（Driver）和执行端（Executor）。控制端负责任务调度，执行端负责任务执行。

读取文件的过程如**图1-47**所示。

图 1-47 读取文件过程

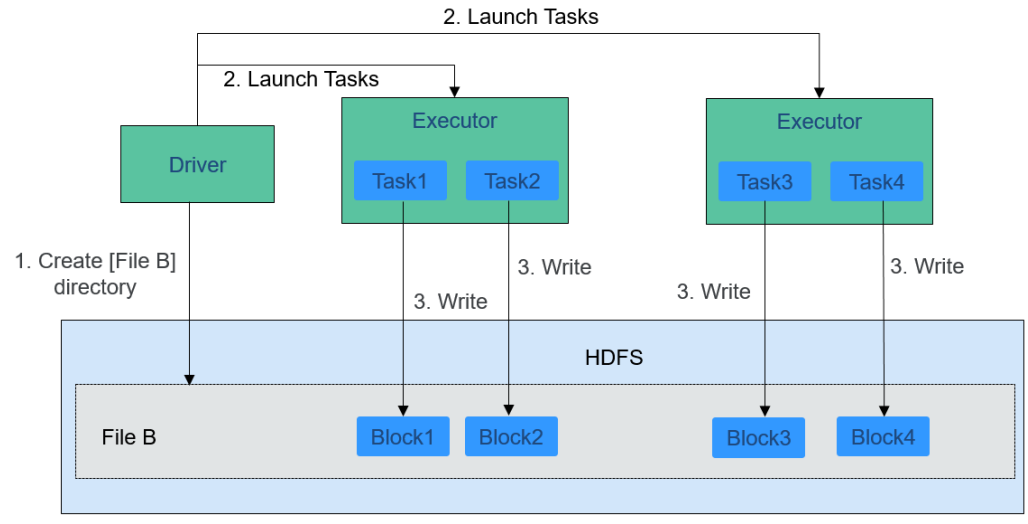


读取文件步骤的详细描述如下所示：

1. Driver与HDFS交互获取File A的文件信息。
2. HDFS返回该文件具体的Block信息。
3. Driver根据具体的Block数据量，决定一个并行度，创建多个Task去读取这些文件Block。
4. 在Executor端执行Task并读取具体的Block，作为RDD（弹性分布数据集）的一部分。

写入文件的过程如图1-48所示。

图 1-48 写入文件过程



HDFS文件写入的详细步骤如下所示：

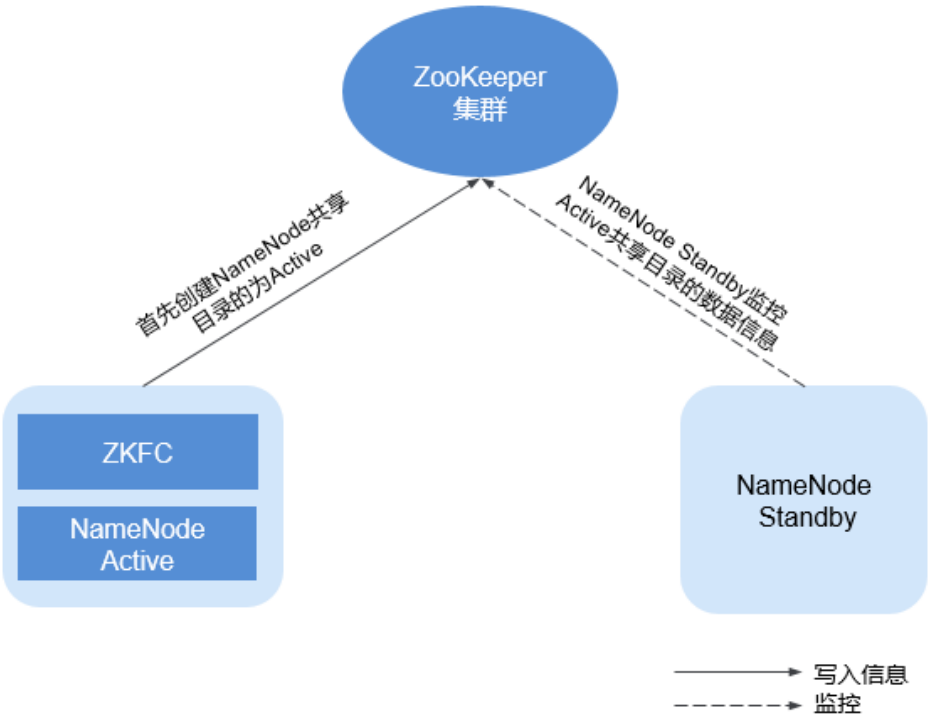
1. Driver创建要写入文件的目录。
2. 根据RDD分区分块情况，计算出写数据的Task数，并下发这些任务到Executor。

3. Executor执行这些Task，将具体RDD的数据写入到步骤1创建的目录下。

HDFS 和 ZooKeeper 的关系

ZooKeeper与HDFS的关系如图1-49所示。

图 1-49 ZooKeeper 和 HDFS 的关系



ZKFC（ZKFailoverController）作为一个ZooKeeper集群的客户端，用来监控NameNode的状态信息。ZKFC进程仅在部署了NameNode的节点中存在。HDFS NameNode的Active和Standby节点均部署有zkfc进程。

1. HDFS NameNode的ZKFC连接到ZooKeeper，把主机名等信息保存到ZooKeeper中，即“/hadoop-ha”下的znode目录里。先创建znode目录的NameNode节点为主节点，另一个为备节点。HDFS NameNode Standby通过ZooKeeper定时读取NameNode信息。
2. 当主节点进程异常结束时，HDFS NameNode Standby通过ZooKeeper感知“/hadoop-ha”目录下发生了变化，NameNode会进行主备切换。

1.4.8.4 HDFS 开源增强特性

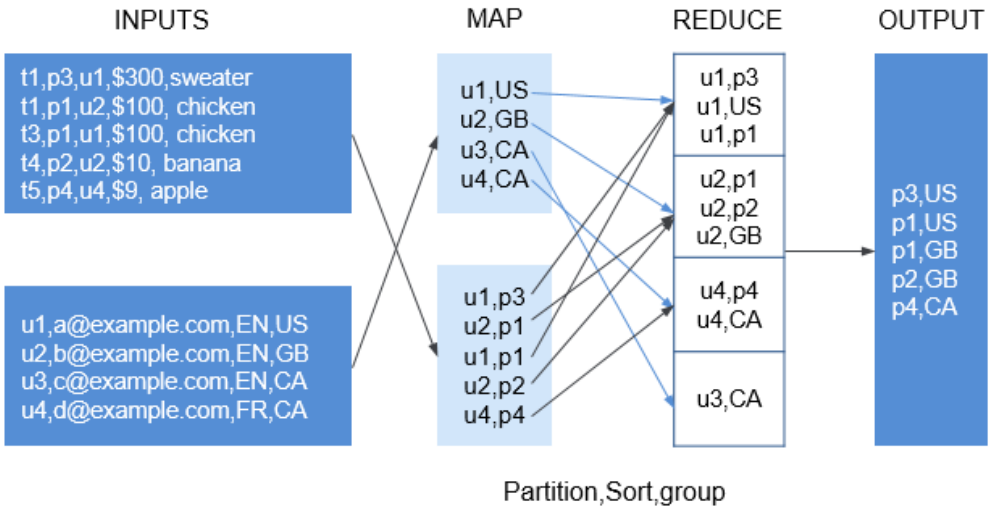
HDFS 开源增强特性：文件块同分布（Colocation）

离线数据汇总统计场景中，Join是一个经常用到的计算功能，在MapReduce中的实现方式大体如下：

1. Map任务分别将两表文件的记录处理成（Join Key，Value），然后按照Join Key做Hash分区后，送到不同的Reduce任务里去处理。
2. Reduce任务一般使用Nested Loop方式递归左表的数据，并遍历右表的每一行，对于相等的Join Key，处理Join结果并输出。

以上方式的最大问题在于，由于数据分散在各节点上，所以在Map到Reduce过程中，需要大量的网络数据传输，使得Join计算的性能大大降低，该过程如图1-50所示：

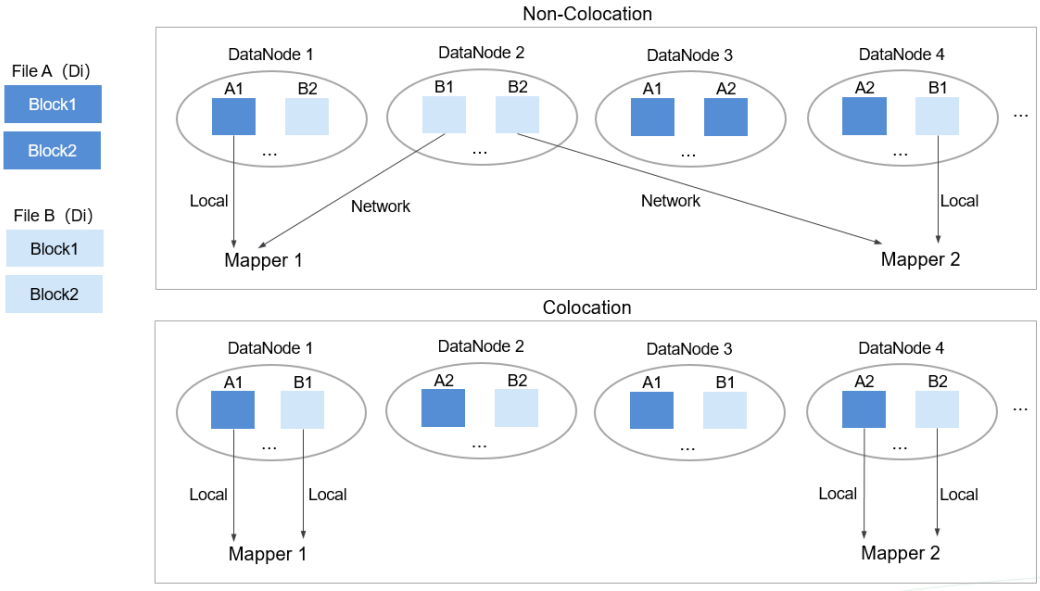
图 1-50 无同分布数据传输流程



由于数据表文件是以HDFS Block方式存放在物理文件系统中，如果能把两个需要Join的文件数据块按Join Key分区后，一一对应地放在同一台机器上，则在Join计算的Reduce过程中无需传递数据，直接在节点本地做Map Join后就能得到结果，性能显著提升。

HDFS数据同分布特性，使得需要做关联和汇总计算的两个文件FileA和FileB，通过指定同一个分布ID，使其所有的Block分布在一起，不再需要跨节点读取数据就能完成计算，极大提高MapReduce Join性能。

图 1-51 无同分布与同分布数据块分布对比

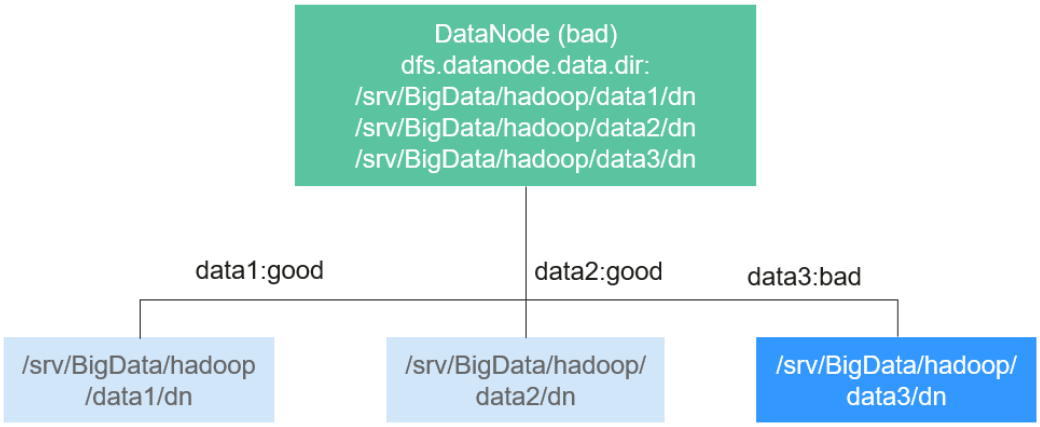


HDFS 开源增强特性：硬盘坏卷设置

在开源版本中，如果为DataNode配置多个数据存放卷，默认情况下其中一个卷损坏，则DataNode将不再提供服务。配置项“dfs.datanode.failed.volumes.tolerated”可以指定失败的个数，小于该个数，DataNode可以继续提供服务。

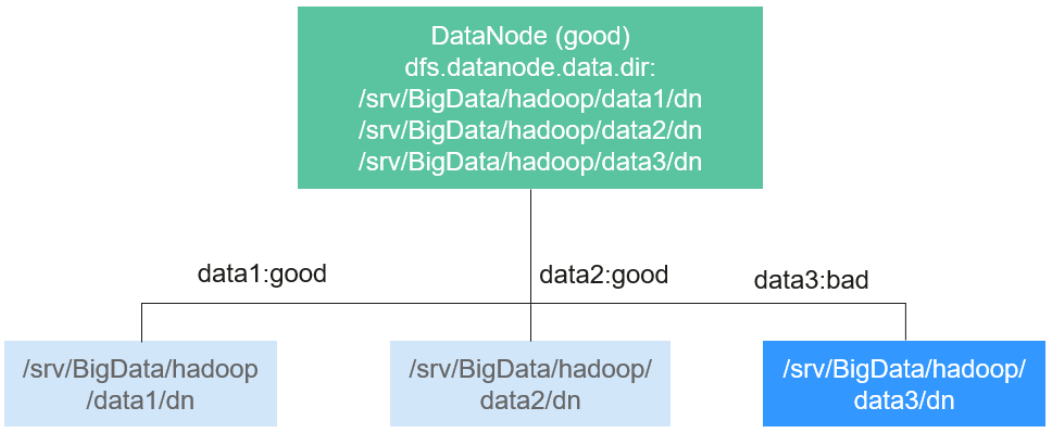
“dfs.datanode.failed.volumes.tolerated”取值范围为-1~DataNode上配置的磁盘卷数，默认值为-1，效果如图1-52所示。

图 1-52 选项设置为 0



例如：某个DataNode中挂载了3个数据存放卷，
“dfs.datanode.failed.volumes.tolerated”配置为1，则当该DataNode中的其中一个数据存放卷不能使用的时候，该DataNode会继续提供服务。如图1-53所示。

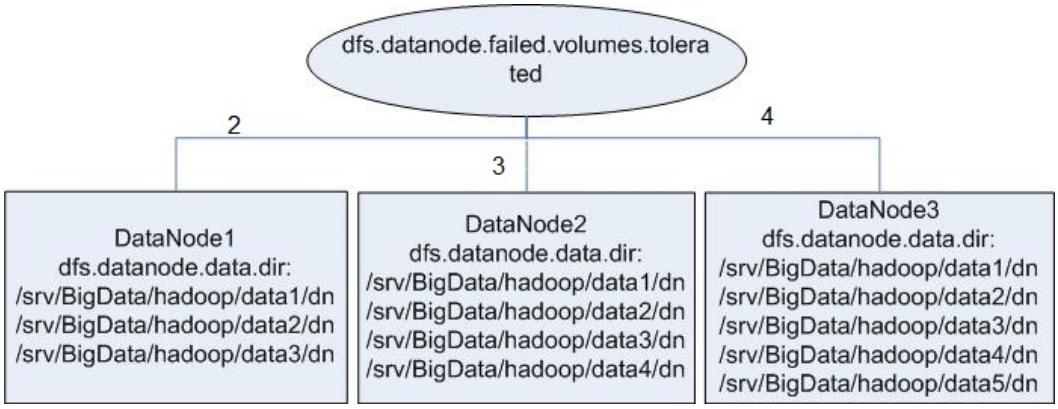
图 1-53 选项设置为 1



这个原生的配置项，存在一定的缺陷。当DataNode的数据存放卷数量不一致的时候，就需要对每个DataNode进行单独配置，而无法配置为所有节点统一生成配置文件，造成用户使用的不便。

例如：集群中存在3个DataNode节点，第一个节点有3个数据目录，第二个节点有4个数据目录，第三个节点有5个数据目录，如果需要实现当节点有一个目录还可用的时候DataNode服务依然可用的效果，就需要如图1-54所示进行设置。

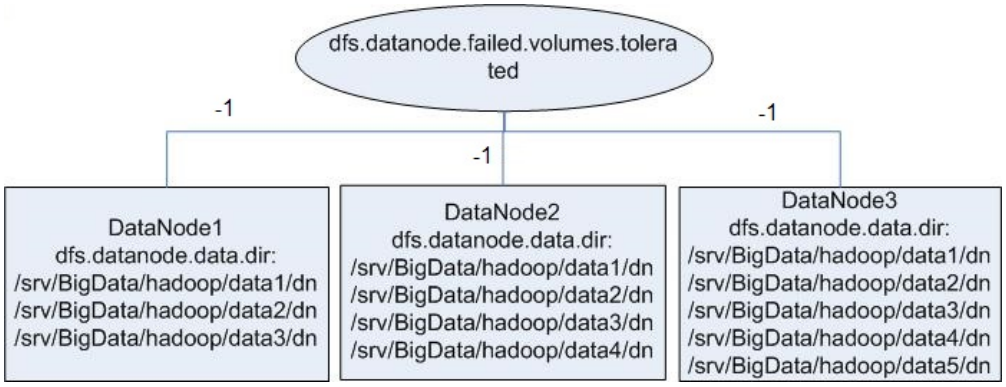
图 1-54 未增强前属性设置



在自研增强版本的HDFS中，对该配置项进行了增强，增加了-1的值选项。当配置成-1的时候，所有DataNode节点只要还有一个数据存放卷，DataNode就能继续提供服务。

所以对于上面提到的例子，该属性的配置将统一成-1，如图1-55所示。

图 1-55 增强后属性配置



HDFS 开源增强特性：HDFS 启动加速

在HDFS中，NameNode启动需要加载元数据文件fsimage，然后等待DataNode完成启动并上报数据块信息。当DataNode上报的数据块信息达到设定百分比时，NameNode退出Safemode，完成启动过程。当HDFS上保存的文件数量达到千万甚至亿级以后，以上两个过程都要耗费大量的时间，致使NameNode的启动过程变得非常漫长。该版本对加载元数据fsimage这一过程进行了优化。

在开源HDFS中，fsimage里保存了所有类型的元数据信息，每一类元数据信息（如文件元数据信息和文件夹元数据信息）分别保存在一个section块里，这些section块在启动时是串行加载的。当HDFS上存储了大量的文件和文件夹时，这两个section的加载就会非常耗时，影响HDFS文件系统的启动时间。HDFS NameNode在生成fsimage时可以将同一类型的元数据信息分段保存在多个section里，当NameNode启动时并行加载fsimage中的section以加快加载速度。

HDFS 开源增强特性：基于标签的数据块摆放策略（HDFS Nodelabel）

用户需要通过数据特征灵活配置HDFS文件数据块的存储节点。通过设置HDFS目录/文件对应一个标签表达式，同时设置每个DataNode对应一个或多个标签，从而给文件的

数据块存储指定了特定范围的DataNode。当使用基于标签的数据块摆放策略，为指定的文件选择DataNode节点进行存放时，会根据文件的标签表达式选择出将要存放的Datanode节点范围，然后在这些Datanode节点范围内，选择出合适的存放节点。

- 支持用户将数据块的各个副本存放在指定具有不同标签的节点，如某个文件的数据块的2个副本放置在标签L1对应节点中，该数据块的其他副本放置在标签L2对应的节点中。
- 支持选择节点失败情况下的策略，如随机从全部节点中选一个。

如图1-56所示。

- /HBase下的数据存储在A，B，D
- /Spark下的数据存储在A，B，D，E，F
- /user下的数据存储在C，D，F
- /user/shl下的数据存储在A，E，F

图 1-56 基于标签的数据块摆放策略样例

HDFS 开源增强特性：HDFS Load Balance

HDFS的现有读写策略主要以数据本地性优先为主，并未考虑节点或磁盘的实际负载情况。HDFS Load Balance功能是基于不同节点的I/O负载情况，在HDFS客户端进行读写操作时，尽可能地选择I/O负载较低的节点进行读写，以此达到I/O负载均衡，以及充分利用集群整体吞吐能力。

写文件时，如果开启写文件的HDFS Load Balance功能，NameNode仍然是根据正常顺序（本地节点—本机架—远端机架）进行DataNode节点的选取，只是在每次选择节点后，如果该节点I/O负载较高，会舍弃并从其他节点中重新选取。

读文件时，Client会向NameNode请求所读Block所在的DataNode列表。NameNode会返回根据网络拓扑距离进行排序的DataNode列表。开启读取的HDFS Load Balance功能时，NameNode会在原先网络拓扑距离排序的基础上，根据每个节点的平均I/O负载情况进行顺序调整，把高I/O负载的节点顺序调整至后面。

HDFS 开源增强特性：HDFS 冷热数据迁移

Hadoop历来主要被用于批量处理大规模的数据。相比处理低时延，批处理应用更关注原始数据处理的吞吐量，因此，目前已有的HDFS模型都运作良好。

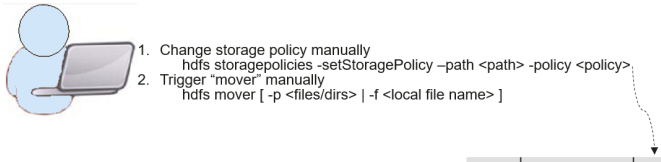
然而，随着技术的发展，Hadoop逐渐被用于以随机I/O访问模式的操作为主的上层应用上，如Hive、HBase等，而这种时延要求较高的场景中，低时延的高速磁盘（如SSD磁盘）可以得到广泛的应用。为了支持这种特性，HDFS现在支持了异构存储类型，这样用户就可以根据自己不同的业务需求场景来选择不同的数据存储类型。

因此，HDFS可以根据数据的热度，选择不同的存储策略。如将HDFS上频繁访问多次的数据被标识为ALL_SSD或HOT，被访问几次的可以标识为WARM，而只有访问1～2次甚至更少的可以被标识为COLD等，如下图为不同的数据热度，可以选择不同的数据存储策略。

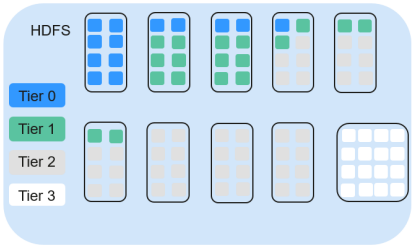
但是，这些高速低时延磁盘，例如SSD磁盘，通常比机械磁盘贵很多。大部分用户希望只有那些经常被访问的热数据才能一直被存储在昂贵的高速磁盘上，而随着数据的访问热度下降以及时间的老化，这些数据应该被迁移到价格低廉的存储介质上。

以详单查询场景作为典型的用例场景，进行说明：当最新详单数据刚刚被导入HDFS上时，会被上层业务人员频繁查询，所以为了提高查询性能，可以将这些详单数据最先导入到SSD磁盘中；但是随着时间的迁移，这些数据逐渐被老化，访问频度越来越低，这时便不适合继续存储在高速硬盘上，需要迁移到廉价的存储介质，节省成本。

目前，如下图所示，HDFS无法很好的支持这些操作，需要自己根据业务类型手动识别数据的热度，并且手动设定数据的存储策略，最后手动触发HDFS Auto Data Movement工具进行数据迁移。



Policy ID	PolicyName	Block Placement (n replacas)	Fallback storages for creation	Fallback storages for replication
15	Lazy_Persist	RAN_DISK:1 DISK:n-1	DISK	DISK
12	All_SSD	SSD:n	DISK	DISK
10	One_SSD	SSD:1,DISK:n-1	SSD,DISK	SSD,DISK
7	Hot(default)	DISK:n	<none>	ARCHIVE
5	Warm	DISK:1,ARCHIVE:n-1	ARCHIVE, DISK	ARCHIVE, DISK
2	Cold	ARCHIVE:n	<none>	<none>



因此，能够基于数据的age自动识别出老化的数据，并将它们迁移到价格低廉的存储介质（如Disk/Archive）上，会给用户节省很高的存储成本，提高数据管理效率。

HDFS Auto Data Movement工具是HDFS冷热数据迁移的核心，根据数据的使用频率自动识别数据冷热设定不同的存储策略。该工具主要支持以下功能：

- 根据数据的age，access time和手动迁移规则，将数据存储策略标识为All_SSD/One_SSD/Hot/Warm/Cold。
- 根据数据age，access time和手动迁移规则，定义区分冷热数据的规则。
- 定义基于age的规则匹配时要采取的行为操作。

  MARK，表示只会基于age规则标识出数据的冷热度，并标记出对应的存储策略。MOVE表示基于age规则识别出相应的数据冷热度，并标记出对应的存储策略后，并触发HDFS Auto Data Movement工具进行数据搬迁。
  - MARK：识别数据是否频繁或很少使用的行为操作，并设置数据存储策略。
  - MOVE：调用HDFS冷热数据迁移工具并跨层迁移数据的行为操作。
  - SET_REPL：为文件设置新的副本数的行为操作。
  - MOVE_TO_FOLDER：将文件移动到目标文件夹的行为操作。
  - DELETE：删除文件/目录的行为操作。
  - SET_NODE_LABEL：设置文件节点标签（NodeLabel）的操作。

使用HDFS冷热数据迁移功能，只需要定义age，基于access time的规则。由HDFS冷热数据迁移工具来匹配基于age的规则的数据，设置存储策略和迁移数据。以这种方式，提高了数据管理效率和集群资源效率。

1.4.9 Hive

1.4.9.1 Hive 基本原理

Hive是建立在Hadoop上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载（ETL），这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。Hive定义了简单的类SQL查询语言，称为HiveQL，它允许熟悉SQL的用户查询数据。Hive的数据计算依赖于MapReduce、Spark、Tez。

使用新的执行引擎Tez代替原先的MapReduce，性能有了显著提升。Tez可以将多个有依赖的作业转换为一个作业（这样只需写一次HDFS，且中间节点较少），从而大大提升DAG作业的性能。

Hive主要特点如下：

- 海量结构化数据分析汇总。
- 将复杂的MapReduce编写任务简化为SQL语句。
- 灵活的数据存储格式，支持JSON，CSV，TEXTFILE，RCFILE，SEQUENCEFILE，ORC（Optimized Row Columnar）这几种存储格式。

Hive体系结构：

- 用户接口：用户接口主要有三个：CLI，Client和WebUI。其中最常用的是CLI，CLI启动的时候，会同时启动一个Hive副本。Client是Hive的客户端，用户连接至Hive Server。在启动Client模式的时候，需要指出Hive Server所在节点，并且在该节点启动Hive Server。WebUI是通过浏览器访问Hive。MRS仅支持Client方式访问Hive。
- 元数据存储：Hive将元数据存储在数据库中，如MySQL、Derby。Hive中的元数据包括表的名字，表的列和分区及其属性，表的属性（是否为外部表等），表的数据所在目录等。

Hive 结构

Hive为单实例的服务进程，提供服务的原理是将HQL编译解析成相应的MapReduce或者HDFS任务，图1-57为Hive的结构概图。

图 1-57 Hive 结构

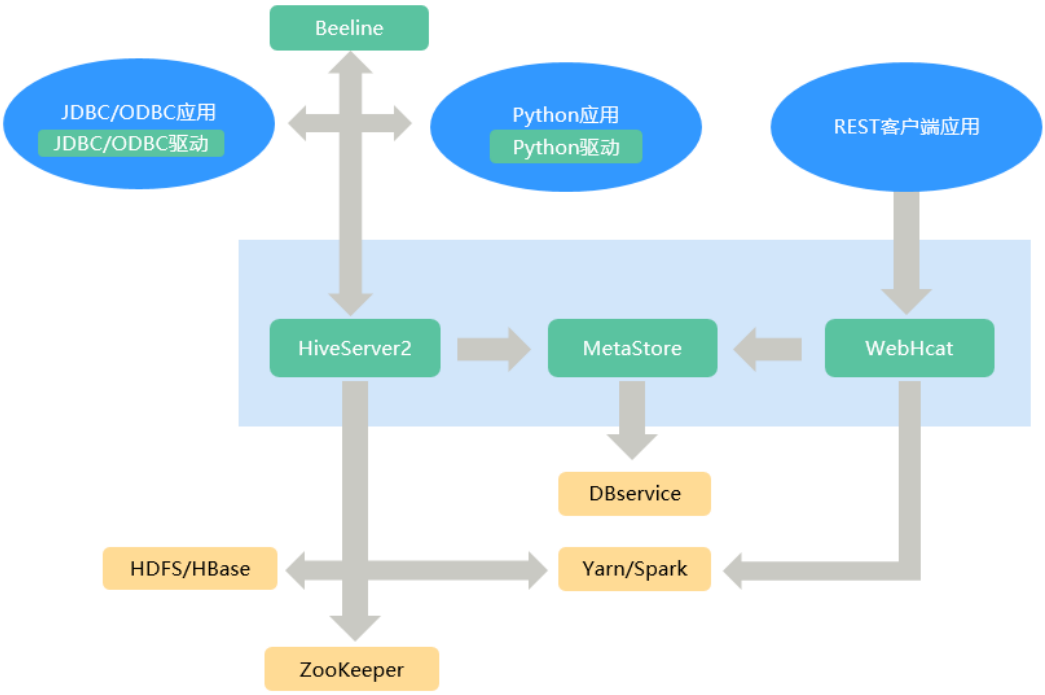
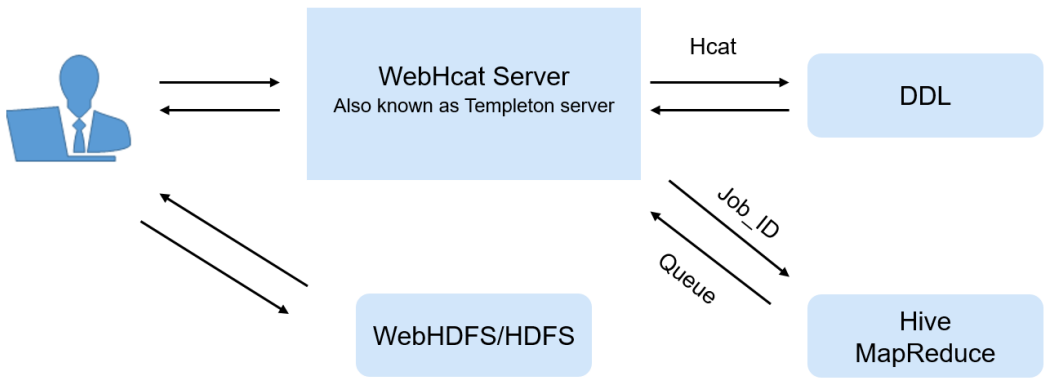


表 1-9 模块说明

名称	说明
HiveServer	一个集群内可部署多个HiveServer，负荷分担。对外提供Hive数据库服务，将用户提交的HQL语句进行编译，解析成对应的Yarn任务或者HDFS操作，从而完成数据的提取、转换、分析。
MetaStore	• 一个集群内可部署多个MetaStore，负荷分担。提供Hive的元数据服务，负责Hive表的结构和属性信息读、写、维护和修改。 • 提供Thrift接口，供HiveServer、Spark、WebHCat等MetaStore客户端来访问，操作元数据。
WebHCat	一个集群内可部署多个WebHCat，负荷分担。提供Rest接口，通过Rest执行Hive命令，提交MapReduce任务。
Hive客户端	包括人机交互命令行Beeline、提供给JDBC应用的JDBC驱动、提供给Python应用的Python驱动、提供给Mapreduce的HCatalog相关JAR包。
ZooKeeper集群	ZooKeeper作为临时节点记录各HiveServer实例的IP地址列表，客户端驱动连接Zookeeper获取该列表，并根据路由机制选取对应的HiveServer实例。
HDFS/HBase集群	Hive表数据存储在HDFS集群中。
MapReduce/Yarn集群	提供分布式计算服务：Hive的大部分数据操作依赖MapReduce，HiveServer的主要功能是将HQL语句转换成MapReduce任务，从而完成对海量数据的处理。

HCatalog建立在Hive Metastore之上，具有Hive的DDL能力。从另外一种意义上说，HCatalog还是Hadoop的表和存储管理层，它使用户能够通过使用不同的数据处理工具（比如MapReduce），更轻松地在网格上读写HDFS上的数据，HCatalog还能为这些数据处理工具提供读写接口，并使用Hive的命令行接口发布数据定义和元数据探索命令。此外，经过封装这些命令，WebHcat Server还对外提供了RESTful接口，如图1-58所示。

图 1-58 WebHCat 的逻辑架构图



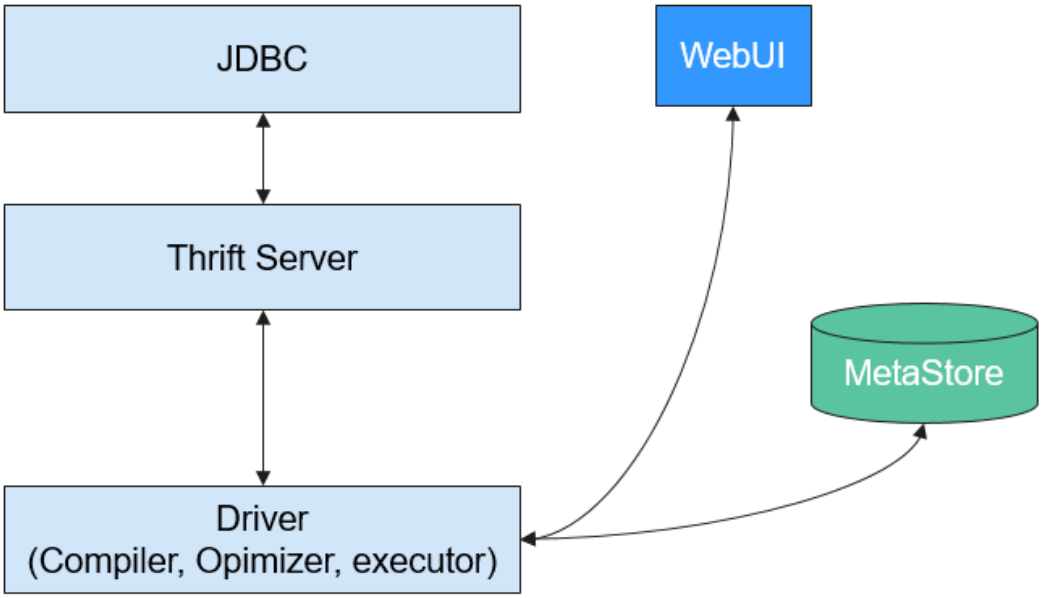
Hive 原理

Hive作为一个基于HDFS和MapReduce架构的数据仓库，其主要能力是通过对HQL（Hive Query Language）编译和解析，生成并执行相应的MapReduce任务或者HDFS操作。Hive与HiveQL相关信息，请参考**HiveQL 语言手册**。

图1-59为Hive的结构简图。

- **Metastore** - 对表，列和Partition等的元数据进行读写及更新操作，其下层为关系型数据库。
- **Driver** - 管理HiveQL执行的生命周期并贯穿Hive任务整个执行期间。
- **Compiler** - 编译HiveQL并将其转化为一系列相互依赖的Map/Reduce任务。
- **Optimizer** - 优化器，分为逻辑优化器和物理优化器，分别对HiveQL生成的执行计划和MapReduce任务进行优化。
- **Executor** - 按照任务的依赖关系分别执行Map/Reduce任务。
- **ThriftServer** - 提供thrift接口，作为JDBC的服务端，并将Hive和其他应用程序集成起来。
- **Clients** - 包含WebUI和JDBC接口，为用户访问提供接口。

图 1-59 Hive 结构



1.4.9.2 Hive CBO 原理介绍

Hive CBO 原理介绍

CBO，全称是Cost Based Optimization，即基于代价的优化器。

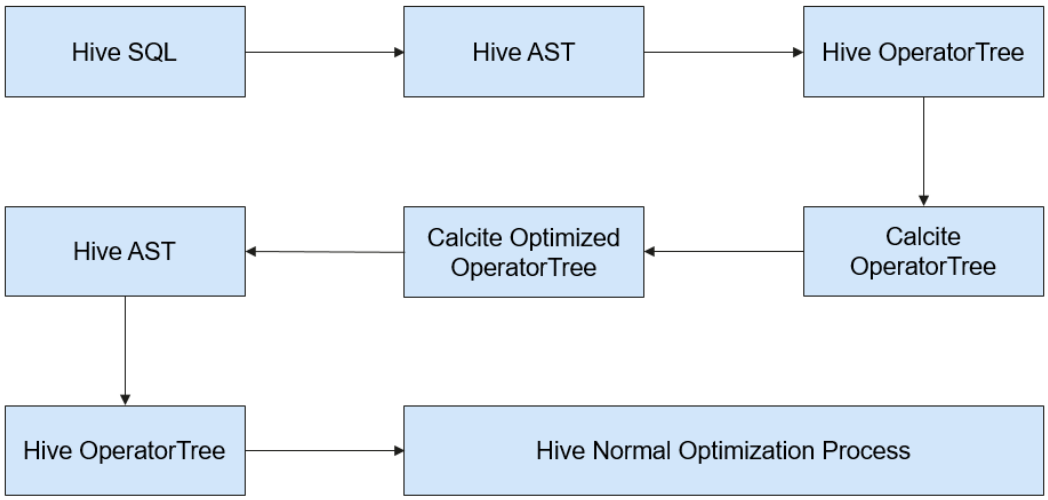
其优化目标是：

在编译阶段，根据查询语句中涉及到的表和查询条件，计算出产生中间结果少的高效join顺序，从而减少查询时间和资源消耗。

Hive中实现CBO的总体过程如下：

Hive使用开源组件Apache Calcite实现CBO。首先SQL语句转化成Hive的AST，然后转成Calcite可以识别的RelNodes。Calcite将RelNode中的Join顺序调整后，再由Hive将RelNode转成AST，继续Hive的逻辑优化和物理优化过程。流程图如图1-60所示：

图 1-60 实现流程图



Calcite调整Join顺序的具体过程如下：

1. 针对所有参与Join的表，依次选取一个表作为第一张表。
2. 依据选取的第一张表，根据代价选择第二张表，第三张表。由此可以得到多个不同的执行计划。
3. 计算出代价最小的一个计划，作为最终的顺序优化结果。

代价的具体计算方法：

当前版本，代价的衡量基于Join出来的数据条数：Join出来的条数越少，代价越小。Join条数的多少，取决于参与join的表的选择率。表的数据条数，取自表级别的统计信息。

过滤条件过滤后的条数，由列级别的统计信息，max，min，以及NDV（Number of Distinct Values）来估算出来。

例如存在一张表table_a，其统计信息如下：数据总条数1000000，NDV 50，查询条件如下：

```
Select * from table_a where colum_a='value1';
```

则估算查询的最终条数为1000000 * 1/50 = 20000条，选择率为2%。

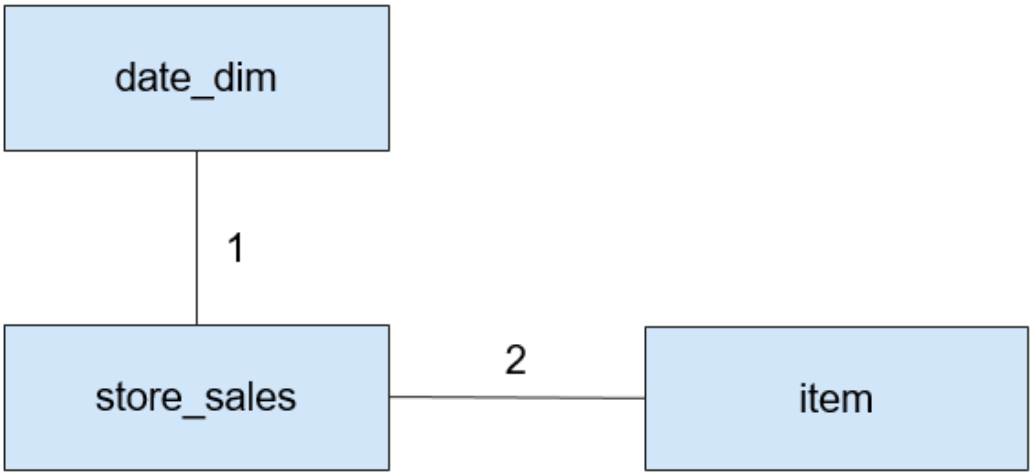
以下以TPC-DS Q3为例来介绍CBO是如何调整Join顺序的。

```
select
    dt.d_year,
    item.i_brand_id brand_id,
    item.i_brand brand,
    sum(ss_ext_sales_price) sum_agg
from
    date_dim dt,
    store_sales,
    item
where
    dt.d_date_sk = store_sales.ss_sold_date_sk
    and store_sales.ss_item_sk = item.i_item_sk
```

```
   and item.i_manufact_id = 436
   and dt.d_moy = 12
group by dt.d_year , item.i_brand , item.i_brand_id
order by dt.d_year , sum_agg desc , brand_id
limit 10;
```

语句解释：这个语句由三张表来做Inner join，其中store_sales是事实表，有约2900000000条数据，date_dim是维度表，有约73000条数据，item是维度表，有约18000条数据。每一个表上都有过滤条件，其Join关系如所图1-61示：

图 1-61 Join 关系



CBO应该先选择能起到最好过滤效果的表来join。

通过分析min，max，NDV，以及数据条数。CBO估算出不同维度表的选择率，详情如表1-10所示。

表 1-10 数据过滤

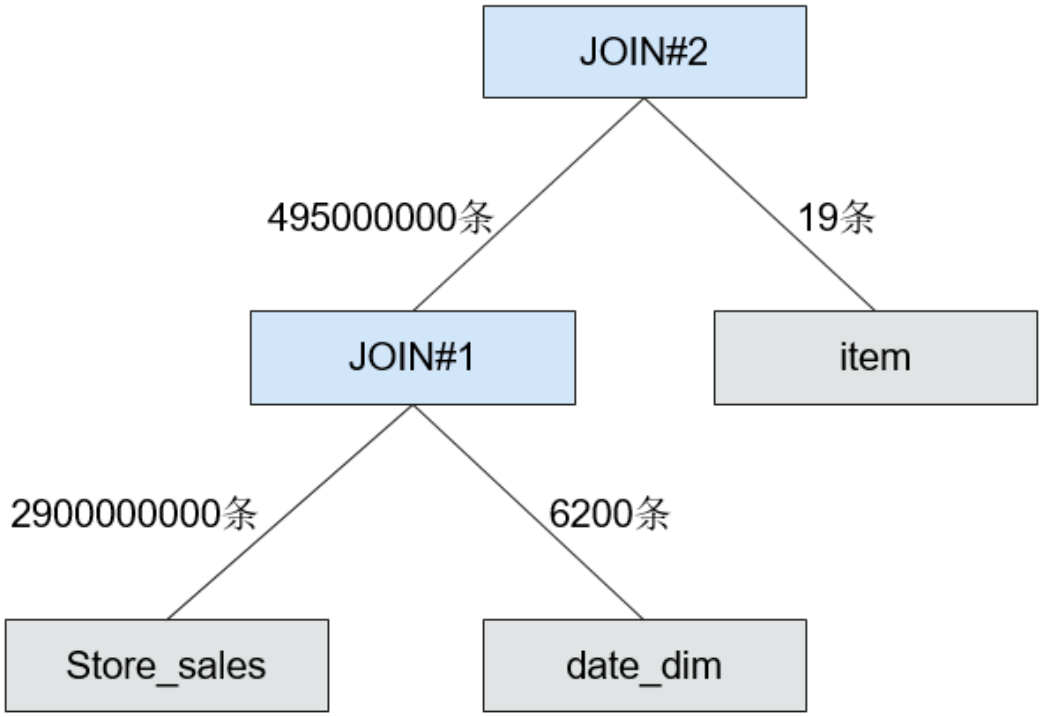
表名	原始数据条数	过滤后数据条数	选择率
date_dim	73000	6200	8.5%
item	18000	19	0.1%

上述表格获取到原始表的数据条数，估算出过滤后的数据条数后，计算出选择率=过滤后条数/原始条数。

从上表可以看出，item表具有较好的过滤效果，因此CBO将item表的join顺序提前。

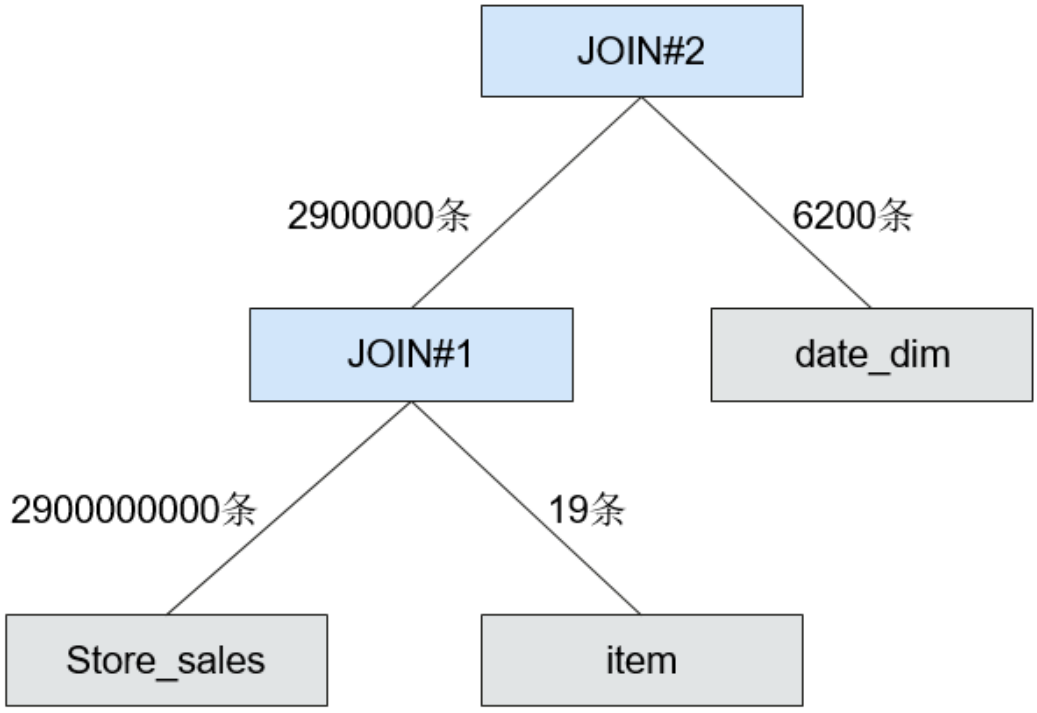
CBO未开启时的Join示意图如图1-62所示：

图 1-62 未开启 CBO



CBO开启后的Join示意图如**图1-63**所示：

图 1-63 开启 CBO



可以看出，优化后中间结果由495000000条减少到了2900000条，执行时间也大幅减少。

1.4.9.3 Hive 与其他组件的关系

Hive 与 HDFS 组件的关系

Hive是Apache的Hadoop项目的子项目，Hive利用HDFS作为其文件存储系统。Hive通过解析和计算处理结构化的数据，Hadoop HDFS则为Hive提供了高可靠性的底层存储支持。Hive数据库中的所有数据文件都可以存储在Hadoop HDFS文件系统上，Hive所有的数据操作也都是通过Hadoop HDFS接口进行的。

Hive 与 MapReduce 组件的关系

Hive的数据计算依赖于MapReduce。MapReduce也是Apache的Hadoop项目的子项目，它是一个基于Hadoop HDFS分布式并行计算框架。Hive进行数据分析时，会将用户提交的HQL语句解析成相应的MapReduce任务并提交MapReduce执行。

Hive 与 Tez 的关系

Tez是Apache的开源项目，它是一个支持有向无环图的分布式计算框架，Hive使用Tez引擎进行数据分析时，会将用户提交的HQL语句解析成相应的Tez任务并提交Tez执行。

Hive 与 DBService 的关系

Hive的MetaStore（元数据服务）处理Hive的数据库、表、分区等的结构和属性信息（即Hive的元数据），这些信息需要存放在一个关系型数据库中，由MetaStore管理和处理。在产品中，Hive的元数据由DBService组件存储和维护，由Metadata组件提供元数据服务。

1.4.9.4 Hive 开源增强特性

Hive 开源增强特性：支持 HDFS Colocation

HDFS Colocation（同分布）是HDFS提供的数据分布控制功能，利用HDFS Colocation接口，可以将存在关联关系或者可能进行关联操作的数据存放在相同的存储节点上。

Hive支持HDFS的Colocation功能，即在创建Hive表时，通过设置表文件分布的locator信息，可以将相关表的数据文件存放在相同的存储节点上，从而使后续的多表关联的数据计算更加方便和高效。

Hive 开源增强特性：支持列加密功能

Hive支持对表的某一列或者多列进行加密。在创建Hive表时，可以指定要加密的列和加密算法。当使用insert语句向表中插入数据时，即可将对应的列进行加密。Hive列加密不支持视图以及Hive over HBase场景。

Hive列加密机制目前支持的加密算法有两种，具体使用的算法在建表时指定。

- AES（对应加密类名称为：org.apache.hadoop.hive.serde2.AESRewriter）
- SMS4（对应加密类名称为：org.apache.hadoop.hive.serde2.SMS4Rewriter）

Hive 开源增强特性：支持 HBase 删除功能

由于底层存储系统的原因，Hive并不能支持对单条表数据进行删除操作，但在Hive on HBase功能中，MRS解决方案中的Hive提供了对HBase表的单条数据的删除功能，通过特定的语法，Hive可以将自己在HBase表中符合条件的一条或者多条数据清除。

Hive 开源增强特性：支持行分隔符

通常情况下，Hive以文本文件存储的表会以回车作为其行分隔符，即在查询过程中，以回车符作为一行表数据的结束符。

但某些数据文件并不是以回车分隔的规则文本格式，而是以某些特殊符号分割其规则文本。

MRS Hive支持指定不同的字符或字符组合作为Hive文本数据的行分隔符。

Hive 开源增强特性：支持基于 HTTPS/HTTP 协议的 REST 接口切换

WebHCat为Hive提供了对外可用的REST接口，开源社区版本默认使用HTTP协议。

MRS Hive支持使用更安全的HTTPS协议，并且可以在两种协议间自由切换。

Hive 开源增强特性：支持开启 Transform 功能

Hive开源社区版本禁止Transform功能。 MRS Hive提供配置开关，Transform功能默认为禁止，与开源社区版本保持一致。

用户可修改配置开关，开启Transform功能，当开启Transform功能时，存在一定的安全风险。

Hive 开源增强特性：支持创建临时函数不需要 ADMIN 权限的功能

Hive开源社区版本创建临时函数需要用户具备ADMIN权限。MRS Hive提供配置开关，默认为创建临时函数需要ADMIN权限，与开源社区版本保持一致。

用户可修改配置开关，实现创建临时函数不需要ADMIN权限。

Hive 开源增强特性：支持数据库授权

Hive开源社区版本只支持数据库的拥有者在数据库中创建表。MRS Hive支持授予用户在数据库中创建表“CREATE”和查询表“SELECT”权限。当授予用户在数据库中查询的权限之后，系统会自动关联数据库中所有表的查询权限。

Hive 开源增强特性：支持列授权

Hive开源社区版本只支持表级别的权限控制。MRS Hive支持列级别的权限控制，可授予用户列级别权限，例如查询“SELECT”、插入“INSERT”、修改“UPDATE”权限。

1.4.10 Hue

1.4.10.1 Hue 基本原理

Hue是一组WEB应用，用于和MRS大数据组件进行交互，能够帮助用户浏览HDFS，进行Hive查询，启动MapReduce任务等，它承载了与所有MRS大数据组件交互的应用。

Hue主要包括了文件浏览器和查询编辑器的功能：

- 文件浏览器能够允许用户直接通过界面浏览以及操作HDFS的不同目录；
- 查询编辑器能够编写简单的SQL，查询存储在Hadoop之上的数据。例如HDFS，HBase，Hive。 用户可以方便地创建、管理、执行SQL，并且能够以Excel的形式下载执行的结果。

通过Hue可以在界面针对组件进行以下操作：

- HDFS：
  - 查看、创建、管理、重命名、移动、删除文件/目录。
  - 上传、下载文件。
  - 搜索文件、目录、文件所有人、所属用户组；修改文件以及目录的属主和权限。
  - 手动配置HDFS目录存储策略，配置动态存储策略等操作。
- Hive：
  - 编辑、执行SQL/HQL语句；保存、复制、编辑SQL/HQL模板；解释SQL/HQL语句；保存SQL/HQL语句并进行查询。
  - 数据库展示，数据表展示。
  - 支持多种Hadoop存储。
  - 通过metastore对数据库及表和视图进行增删改查等操作。

    说明

    如果使用IE浏览器访问Hue界面来执行HiveSQL，由于浏览器存在的功能问题，将导致执行失败。建议使用兼容的浏览器，例如Google Chrome浏览器。
- Impala：
  - 编辑、执行SQL/HQL语句；保存、复制、编辑SQL/HQL模板；解释SQL/HQL语句；保存SQL/HQL语句并进行查询。
  - 数据库展示，数据表展示。
  - 支持多种Hadoop存储。
  - 通过metastore对数据库及表和视图进行增删改查等操作。

    说明

    如果使用IE浏览器访问Hue界面来执行HiveSQL，由于浏览器存在的功能问题，将导致执行失败。建议使用兼容的浏览器，例如Google Chrome浏览器。
- MapReduce：查看集群中正在执行和已经完成的MR任务，包括它们的状态、起始结束时间、运行日志等。
- Oozie：提供了Oozie作业管理器功能，使用户可以通过界面图形化的方式使用Oozie。
- ZooKeeper：提供了ZooKeeper浏览器功能，使用户可以通过界面图形化的方式查看ZooKeeper。

有关Hue的详细信息，请参见：**http://gethue.com/**。

Hue 结构

Hue是建立在Django Python（开放源代码的Web应用框架）的Web框架上的Web应用程序，采用了MTV（模型M-模板T-视图V）的软件设计模式。

Hue由“Supervisor Process”和“WebServer”构成，“Supervisor Process”是Hue的核心进程，负责应用进程管理。“Supervisor Process”和“WebServer”通过“THRIFT/REST”接口与WebServer上的应用进行交互，如图1-64所示。

图 1-64 Hue 架构示意图

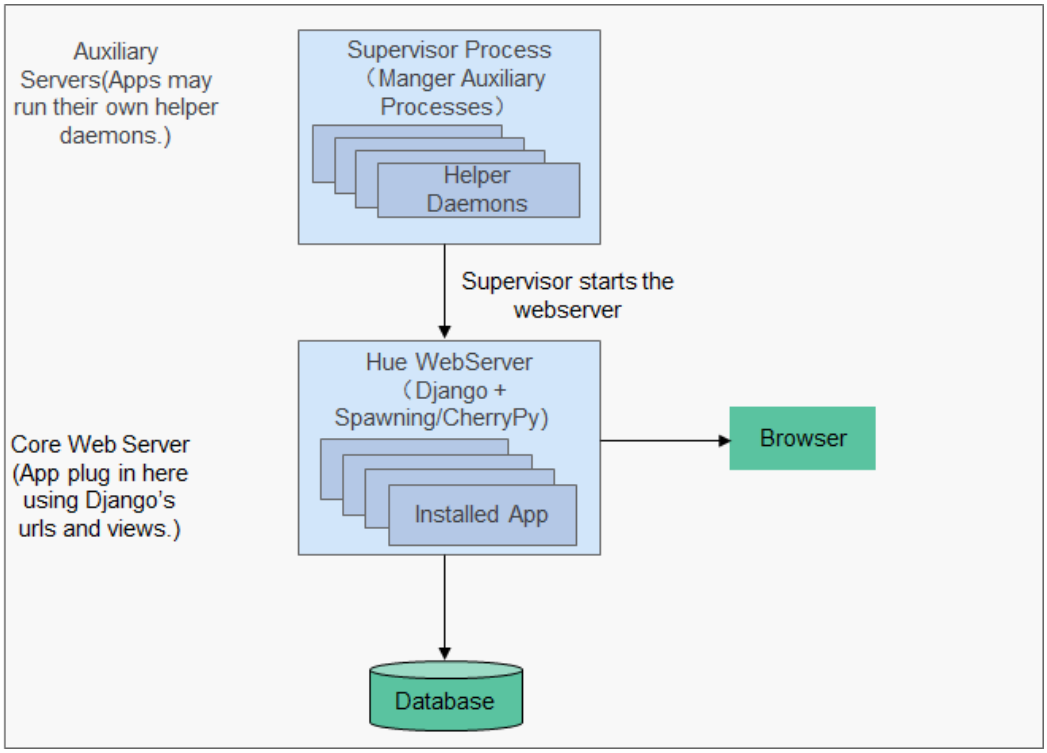


图1-64中各部分的功能说明如表1-11所示。

表 1-11 结构图说明

名称	描述
Supervisor Process	Supervisor负责WebServer上APP的进程管理：启动、停止、监控等。
Hue WebServer	通过Django Python的Web框架提供如下功能。 ● 部署APPs。 ● 提供图形化用户界面。 ● 与数据库连接，存储APPs的持久化数据。

1.4.10.2 Hue 与其他组件的关系

Hue 与 Hadoop 集群的关系

Hue与Hadoop集群的交互关系如图1-65所示。

图 1-65 Hue 与 Hadoop 集群

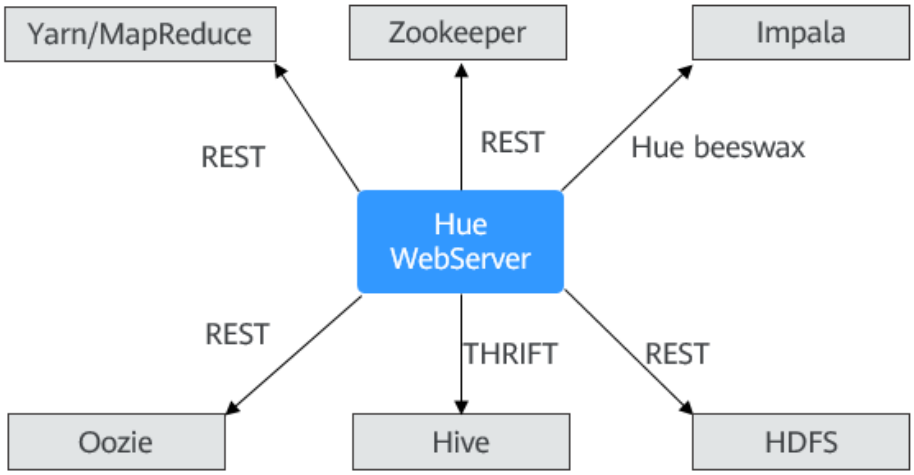


表 1-12 Hue 与其它组件的关系

名称	描述
HDFS	HDFS提供REST接口与Hue交互，用于查询、操作HDFS文件。 在Hue把用户请求从用户界面组装成接口数据，通过调用REST接口调用HDFS，通过浏览器返回结果呈现给用户。
Hive	Hive提供THRIFT接口与Hue交互，用于执行Hive SQL、查询表元数据。 在Hue界面编辑HQL语句，通THRIFT接口提交HQL语句到HIVESERVER执行，同时把执行通过浏览器呈现给用户。
Yarn/MapReduce	MapReduce提供REST与Hue交互，用于查询Yarn作业信息。 进入Hue页面，输入筛选条件参数，UI将参数发送到后台，Hue通过调用MapReduce（MR1/MR2-YARN）提供的REST接口，获取任务运行的状态，起始结束时间、运行日志等信息。
Oozie	Oozie提供REST接口与Hue交互，用于创建工作流、Coordinator、Bundle，以及它们的任务管理和监控。 在Hue前端提供图形化工作流、Coordinator、Bundle编辑器，Hue调用Oozie REST接口对工作流、Coordinator、Bundle进行创建、修改、删除、提交、监控。
ZooKeeper	ZooKeeper提供REST接口与Hue交互，用于查询ZooKeeper节点信息。 在Hue前端显示ZooKeeper节点信息，Hue调用ZooKeeper REST接口获取这些节点信息。

名称	描述
Impala	Impala提供Hue beeswax接口与Hue交互，用于执行Hive SQL、查询表元数据。 在Hue界面编辑HQL语句，通Hue beeswax接口提交HQL语句到HIVESERVER执行，同时把执行结果通过浏览器呈现给用户。

1.4.10.3 Hue 开源增强特性

Hue 开源增强特性

- 存储策略定义。HDFS文件存储在多种等级的存储介质中，有不同的副本数。本特性可以手工设置HDFS目录的存储策略，或者根据HDSF文件最近访问时间和最近修改时间，自动调整文件存储策略、修改文件副本数、移动文件所在目录、自动删除文件，以便充分利用存储的性能和容量。
- MR引擎。用户执行Hive SQL可以选择使用MR引擎执行。
- 可靠性增强。Hue自身主备部署。Hue与HDFS、Oozie、Hive、Yarn等对接时，支持Failover或负载均衡工作模式。

1.4.11 Impala

Impala直接对存储在HDFS、HBase或对象存储服务（OBS）中的Hadoop数据提供快速、交互式SQL查询。除了使用相同的统一存储平台之外，Impala还使用与Apache Hive相同的元数据，SQL语法（Hive SQL），ODBC驱动程序和用户界面（Hue中的Impala查询UI）。这为实时或面向批处理的查询提供了一个熟悉且统一的平台。作为查询大数据的工具的补充，Impala不会替代基于MapReduce构建的批处理框架，例如Hive。基于MapReduce构建的Hive和其他框架最适合长时间运行的批处理作业。

Impala主要特点如下：

- 支持Hive查询语言（HiveQL）中大多数的SQL-92功能，包括 SELECT，JOIN和聚合函数。
- HDFS，HBase 和对象存储服务（OBS）存储，包括：
  - HDFS文件格式：基于分隔符的text file，Parquet，Avro，SequenceFile和RCFile。
  - 压缩编解码器：Snappy，GZIP，Deflate，BZIP。
- 常见的数据访问接口包括：
  - JDBC驱动程序。
  - ODBC驱动程序。
  - HUE beeswax和Impala查询UI。
- impala-shell命令行接口。
- 支持Kerberos身份认证。

Impala主要应用于实时查询数据的离线分析（如日志分析，集群状态分析）、大规模的数据挖掘（用户行为分析，兴趣分区，区域展示）等场景下。

有关Impala的详细信息，请参见**https://impala.apache.org/impala-docs.html**。

1.4.12 Kafka

1.4.12.1 Kafka 基本原理

Kafka是一个分布式的、分区的、多副本的消息发布-订阅系统，它提供了类似于JMS的特性，但在设计上完全不同，它具有消息持久化、高吞吐、分布式、多客户端支持、实时等特性，适用于离线和在线的消息消费，如常规的消息收集、网站活性跟踪、聚合统计系统运营数据（监控数据）、日志收集等大量数据的互联网服务的数据收集场景。

Kafka 结构

生产者（Producer）将消息发布到Kafka主题（Topic）上，消费者（Consumer）订阅这些主题并消费这些消息。在Kafka集群上一个服务器称为一个Broker。对于每一个主题，Kafka集群保留一个用于缩放、并行化和容错性的分区（Partition）。每个分区是一个有序、不可变的消息序列，并不断追加到提交日志文件。分区的消息每个也被赋值一个称为偏移顺序（Offset）的序列化编号。

图 1-66 Kafka 结构

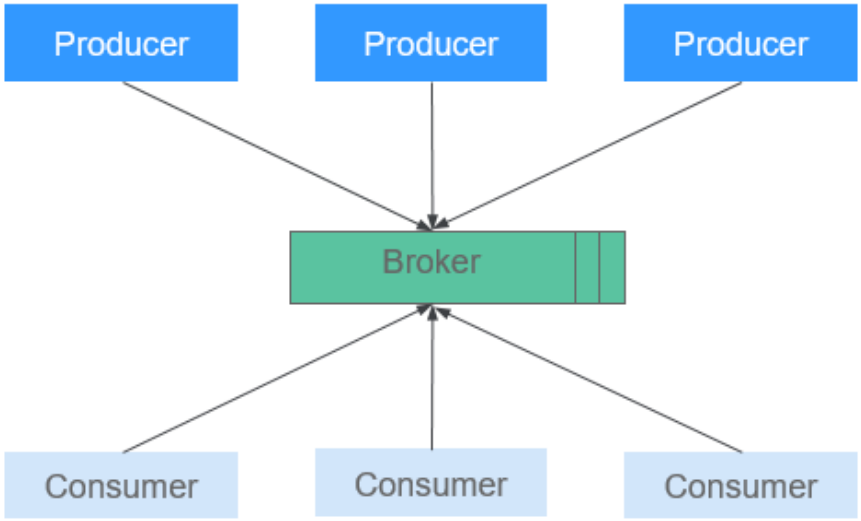


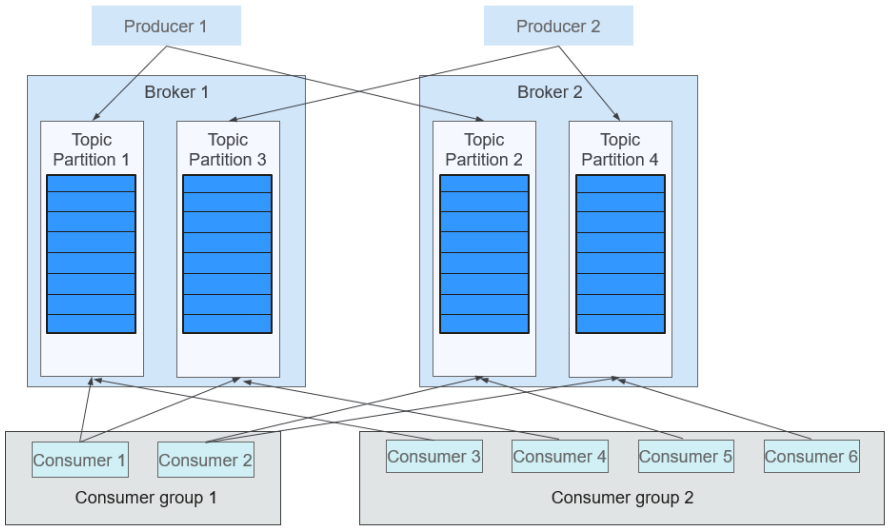
表 1-13 Kafka 结构图说明

名称	说明
Broker	在Kafka集群上一个服务器称为一个Broker。
Topic/主题	一个Topic就是一个类别或者一个可订阅的条目名称，也即一类消息。一个主题可以有多个分区，这些分区可以作为并行的一个单元。
Partition/分区	是一个有序的、不可变的消息序列，这个序列可以被连续地追加一个提交日志。在分区内的每条消息都有一个有序的ID号，这个ID号被称为偏移（Offset），这个偏移量可以唯一确定每条消息在分区内的位置。

名称	说明
Producer/生产者	向Kafka的主题发布消息。
Consumer/消费者	向Topic订阅，并且接收发布到这些Topic的消息。

各模块间关系如图1-67所示。

图 1-67 Kafka 模块间关系



消费者使用一个消费者组名称来标记自己，主题的每个消息被传递给每个订阅消费者组中的一个消费者。如果所有的消费者实例都属于同样的消费组，它们就以传统队列负载均衡方式工作。如上图中，Consumer1与Consumer2之间为负载均衡方式；Consumer3、Consumer4、Consumer5与Consumer6之间为负载均衡方式。如果消费者实例都属于不同的消费组，则消息会被广播给所有消费者。如上图中，Topic1中的消息，同时会广播到Consumer Group1与Consumer Group2中。

关于Kafka架构和详细原理介绍，请参见：**https://kafka.apache.org/24/documentation.html**。

Kafka 原理

- **消息可靠性**

  Kafka Broker收到消息后，会持久化到磁盘，同时，Topic的每个Partition有自己的Replica（备份），每个Replica分布在不同的Broker节点上，以保证当某一节点失效时，可以自动故障转移到可用消息节点。

- **高吞吐量**

  Kafka通过以下方式提供系统高吞吐量：

  - 数据磁盘持久化：消息不在内存中cache，直接写入到磁盘，充分利用磁盘的顺序读写性能。

  - Zero-copy：减少IO操作步骤。
  - 数据批量发送：提高网络利用率。
  - Topic划分为多个Partition，提高并发度，可以由多个Producer、Consumer数目之间的关系并发来读、写消息。Producer根据用户指定的算法，将消息发送到指定的Partition。

- **消息订阅-通知机制**

  消费者对感兴趣的主题进行订阅，并采取pull的方式消费数据，使得消费者可以根据其消费能力自主地控制消息拉取速度，同时，可以根据自身情况自主选择消费模式，例如批量、重复消费，从尾端开始消费等；另外，需要消费者自己负责维护其自身消息的消费记录。

- **可扩展性**

  当在Kafka集群中可通过增加Broker节点以提供更大容量时。新增的Broker会向ZooKeeper注册，而Producer及Consumer会及时从ZooKeeper感知到这些变化，并及时作出调整。

Kafka 开源特性

- 可靠性

  提供At-Least Once，At-Most Once，Exactly Once消息可靠传递。消息被处理的状态是在Consumer端维护，需要结合应用层实现Exactly Once。

- 高吞吐

  同时为发布和订阅提供高吞吐量。

- 持久化

  将消息持久化到磁盘，因此可用于批量消费，以及实时应用程序。通过将数据持久化到硬盘以及replication防止数据丢失。

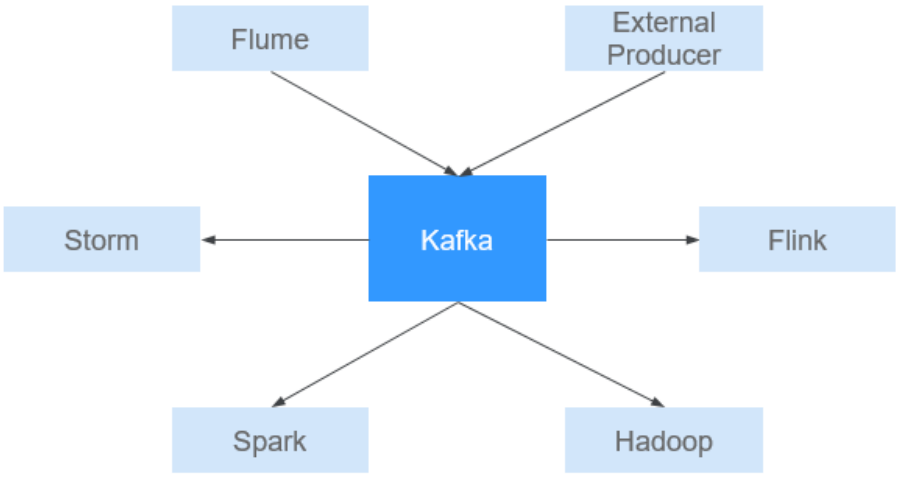
- 分布式

  分布式系统，易于向外扩展。所有的Producer、Broker和Consumer都支持部署多个形成分布式的集群。无需停机即可扩展系统。

1.4.12.2 Kafka 与其他组件的关系

Kafka作为一个消息发布-订阅系统，为整个大数据平台多个子系统之间数据的传递提供了高速数据流转方式。可以实时接受来自外部的消息，并提供给在线以及离线业务进行处理。具体的关系如下图所示：

图 1-68 与其他组件关系



1.4.12.3 Kafka 开源增强特性

Kafka 开源增强特性

- 支持监控如下Topic级别的指标：
  - Topic输入的字节流量
  - Topic输出的字节流量
  - Topic拒绝的字节流量
  - Topic每秒失败的fetch请求数
  - Topic每秒失败的Produce请求数
  - Topic每秒输入的消息条数
  - Topic每秒的fetch请求数
  - Topic每秒的produce请求数
- 支持查询Broker ID与节点IP的对应关系。在Linux客户端下，使用**kafka-broker-info.sh**查询Broker ID与节点IP的对应关系。

1.4.13 KafkaManager

KafkaManager是Apache Kafka的管理工具，提供Kafka集群界面化的Metric监控和集群管理。

通过KafkaManager进行以下操作：

- 支持管理多个Kafka集群
- 支持界面检查集群状态（主题，消费者，偏移量，分区，副本，节点）
- 支持界面执行副本的leader选举
- 使用选择生成分区分配以选择要使用的分区方案
- 支持界面执行分区重新分配（基于生成的分区方案）
- 支持界面选择配置创建主题（支持多种Kafka版本集群）

- 支持界面删除主题（仅支持0.8.2+并设置了delete.topic.enable = true）
- 支持批量生成多个主题的分区分配，并可选择要使用的分区方案
- 支持批量运行重新分配多个主题的分区
- 支持为已有主题增加分区
- 支持更新现有主题的配置
- 可以为分区级别和主题级别度量标准启用JMX查询
- 可以过滤掉zookeeper中没有ids / owner /&offsets /目录的使用者。

1.4.14 KrbServer 及 LdapServer

1.4.14.1 KrbServer 及 LdapServer 基本原理

KrbServer 及 LdapServer 简介

为了管理集群中数据与资源的访问控制权限，推荐以安全模式安装集群。在安全模式下，客户端应用程序在访问集群中的任意资源之前均需要通过身份认证，建立安全会话链接。MRS通过KrbServer为所有组件提供Kerberos认证功能，实现了可靠的认证机制。

LdapServer支持轻量目录访问协议（Lightweight Directory Access Protocol，简称为LDAP），为Kerberos认证提供用户和用户组数据保存能力。

KrbServer 及 LdapServer 结构

用户登录时安全认证功能主要依赖于Kerberos和LDAP。

图 1-69 安全认证场景架构

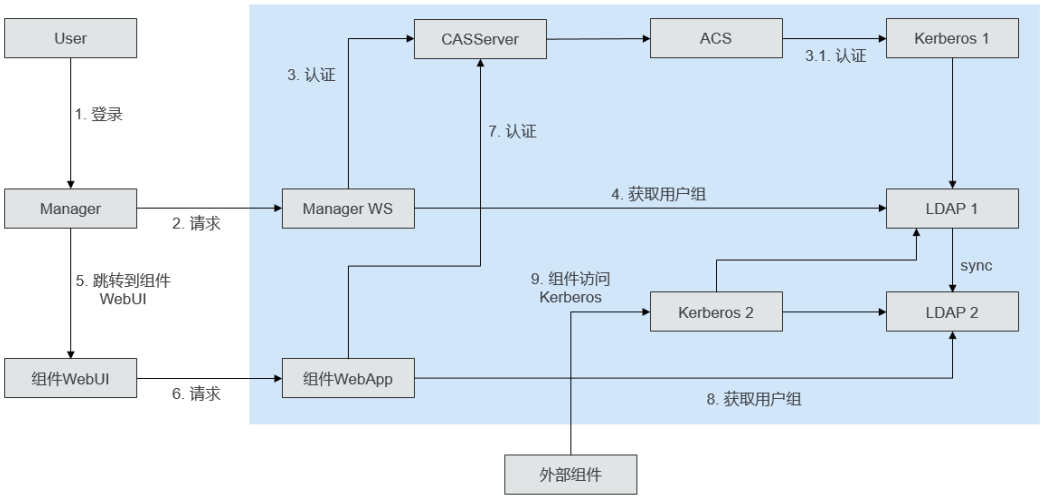


图1-69可分为三类场景：

- 登录Manager WebUI
  认证架构包含步骤1、2、3、4
- 登录组件Web UI

认证架构包含步骤5、6、7、8

- 组件间访问
  认证架构为步骤9

表 1-14 关键模块解释

名称	含义
Manager	集群Manager
Manager WS	WebBrowser
Kerberos1	部署在Manager中的KrbServer（管理平面）服务，即OMS Kerberos
Kerberos2	部署在集群中的KrbServer（业务平面）服务
LDAP1	部署在Manager中的LdapServer（管理平面）服务，即OMS LDAP
LDAP2	部署在集群中的LdapServer（业务平面）服务

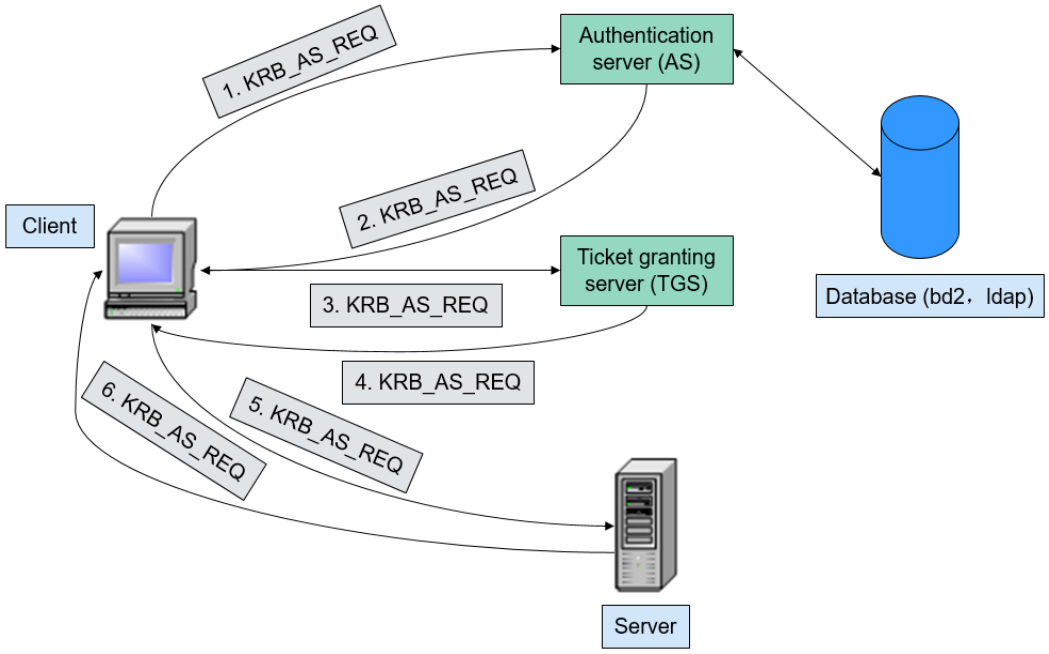
Kerberos1访问LDAP数据：以负载均衡方式访问主备LDAP1两个实例和双备LDAP2两个实例。只能在主LDAP1主实例上进行数据的写操作，可以在LDAP1或者LDAP2上进行数据的读操作。

Kerberos2访问LDAP数据：读操作可以访问LDAP1和LDAP2，数据的写操作只能在主LDAP1实例进行。

KrbServer 及 LdapServer 原理

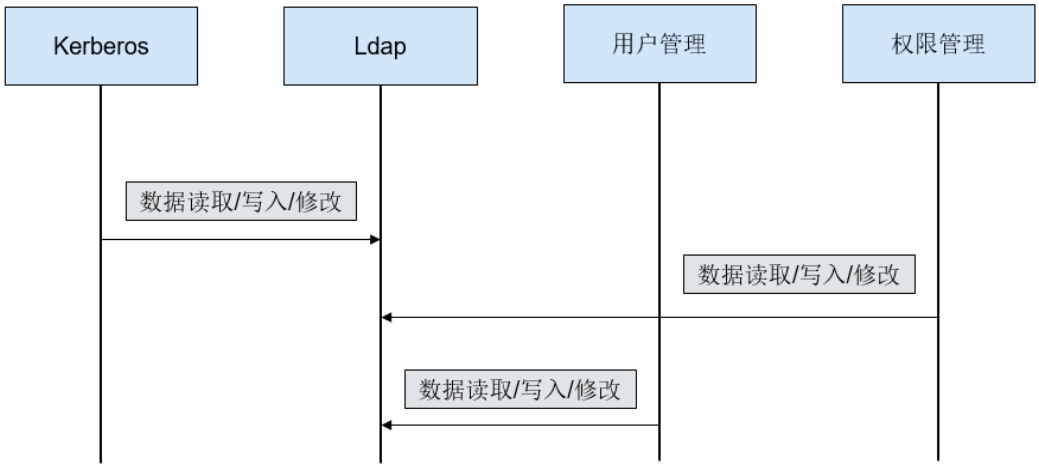
Kerberos认证

图 1-70 认证流程图



LDAP数据读写

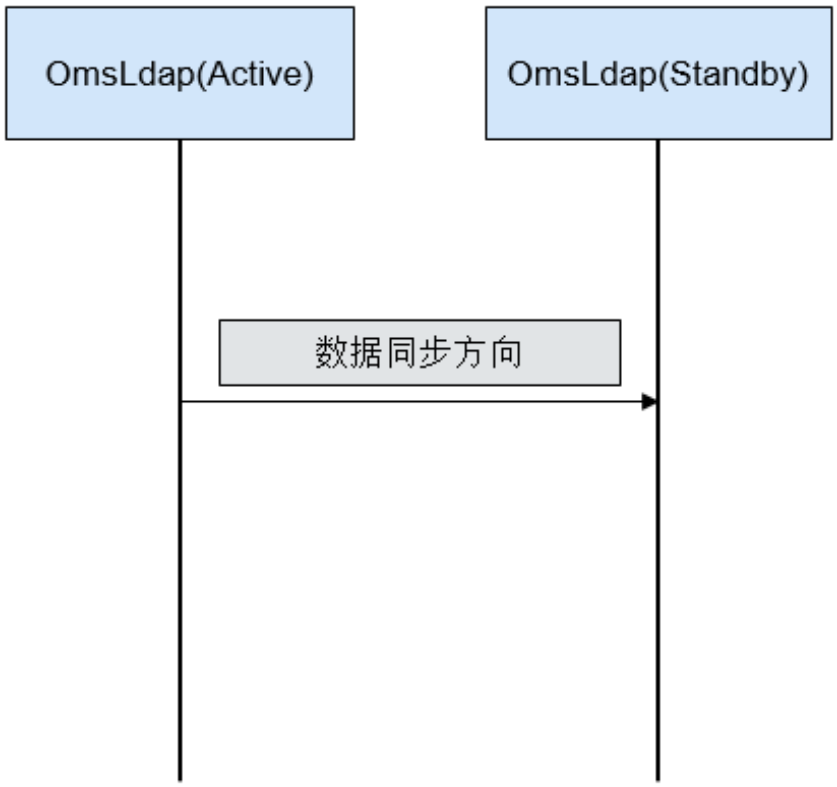
图 1-71 数据修改过程



LDAP数据同步

- 安装集群前OMS LDAP数据同步

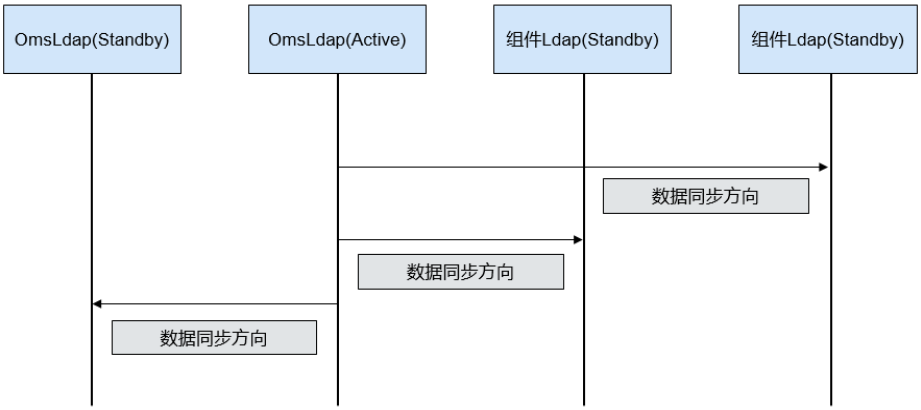
  图 1-72 OMS LDAP 数据同步

  安装集群前数据同步方向：主OMS LDAP同步到备OMS LDAP。

- 安装集群后LDAP数据同步

图 1-73 LDAP 数据同步



安装集群后数据同步方向：主OMS LDAP同步到备OMS LDAP、备组件LDAP和备组件LDAP。

1.4.14.2 KrbServer 及 LdapServer 开源增强特性

KrbServer 及 LdapServer 开源增强特性：集群内服务认证

在使用安全模式的MRS集群中，任意服务间的相互访问基于Kerberos安全架构方案。集群内某个服务（例如HDFS）在启动准备阶段的时候，会首先在Kerberos中获取该服务对应的服务名称sessionkey（即keytab，用于应用程序进行身份认证）。其他任意服务（例如YARN）需要访问HDFS并在HDFS中执行增、删、改、查数据的操作时，必须获取对应的TGT和ST，用于本次安全访问的认证。

KrbServer 及 LdapServer 开源增强特性：应用开发认证

MRS各组件提供了应用开发接口，用于客户或者上层业务产品集群使用。在应用开发过程中，安全模式的集群提供了特定的应用开发认证接口，用于应用程序的安全认证与访问。例如hadoop-common api提供的UserGroupInformation类，该类提供了多个安全认证api接口：

- setConfiguration()主要是获取对应的配置，设置全局变量等参数。
- loginUserFromKeytab()获取TGT接口。

KrbServer 及 LdapServer 开源增强特性：跨系统互信特性

MRS提供两个Manager之间的互信功能，用于实现系统之间的数据读、写等操作。

1.4.15 Kudu

Kudu是专为Apache Hadoop平台开发的列式存储管理器，具有Hadoop生态系统应用程序的共同技术特性：在通用的商用硬件上运行，可水平扩展，提供高可用性。

Kudu的设计具有以下优点：

- 能够快速处理OLAP工作负载
- 支持与MapReduce，Spark和其他Hadoop生态系统组件集成
- 与Apache Impala的紧密集成，使其成为将HDFS与Apache Parquet结合使用的更好选择

- 提供强大而灵活的一致性模型，允许您根据每个请求选择一致性要求，包括用于严格可序列化的一致性的选项
- 提供同时运行顺序读写和随机读写的良好性能
- 易于管理
- 高可用性。Master和TServer采用raft算法，该算法可确保只要副本总数的一半以上可用，tablet就可以进行读写操作。例如，如果3个副本中有2个副本或5个副本中有3个副本可用，则tablet可用。即使主tablet出现故障，也可以通过只读的副tablet提供读取服务
- 支持结构化数据模型

通过结合所有以上属性，Kudu的目标是支持在当前Hadoop存储技术上难以实现或无法实现的应用。

Kudu的应用场景有：

- 需要最终用户立即使用新到达数据的报告型应用
- 同时支持大量历史数据查询和细粒度查询的时序应用
- 使用预测模型并基于所有历史数据定期刷新预测模型来做出实时决策的应用

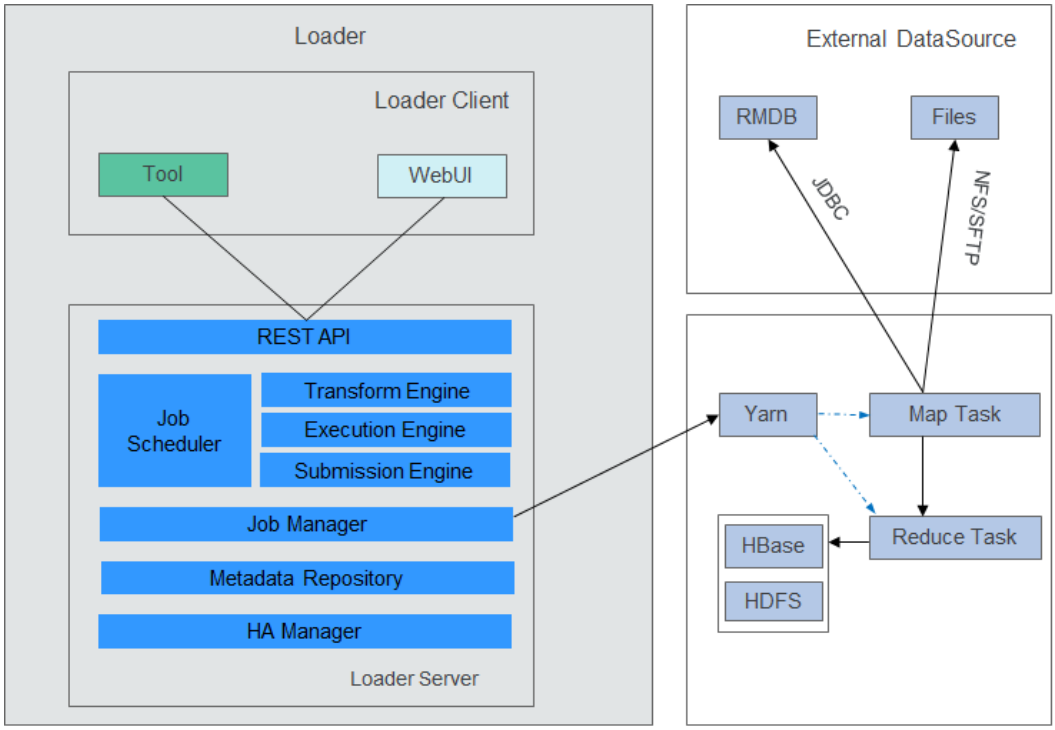
1.4.16 Loader

1.4.16.1 Loader 基本原理

Loader是在开源Sqoop组件的基础上进行了一些扩展，实现MRS与关系型数据库、文件系统之间交换“数据”、“文件”，同时也可以将数据从关系型数据库或者文件服务器导入到HDFS/HBase中，或者反过来从HDFS/HBase导出到关系型数据库或者文件服务器中。

Loader模型主要由Loader Client和Loader Server组成，如图1-74所示。

图 1-74 Loader 模型



上图中各部分的功能说明如表1-15所示。

表 1-15 Loader 模型组成

名称	描述
Loader Client	Loader的客户端，包括WebUI和CLI版本两种交互界面。
Loader Server	Loader的服务端，主要功能包括：处理客户端操作请求、管理连接器和元数据、提交MapReduce作业和监控MapReduce作业状态等。
REST API	实现RESTful（HTTP + JSON）接口，处理来自客户端的操作请求。
Job Scheduler	简单的作业调度模块，支持周期性的执行Loader作业。
Transform Engine	数据转换处理引擎，支持字段合并、字符串剪切、字符串反序等。
Execution Engine	Loader作业执行引擎，支持以MapReduce方式执行Loader作业。
Submission Engine	Loader作业提交引擎，支持将作业提交给MapReduce执行。
Job Manager	管理Loader作业，包括创建作业、查询作业、更新作业、删除作业、激活作业、去激活作业、启动作业、停止作业。
Metadata Repository	元数据仓库，存储和管理Loader的连接器、转换步骤、作业等数据。
HA Manager	管理Loader Server进程的主备状态，Loader Server包含2个节点，以主备方式部署。

Loader通过MapReduce作业实现并行的导入或者导出作业任务，不同类型的导入导出作业可能只包含Map阶段或者同时Map和Reduce阶段。

Loader同时利用MapReduce实现容错，在作业任务执行失败时，可以重新调度。

- **数据导入到HBase**

  在MapReduce作业的Map阶段中从外部数据源抽取数据。

  在MapReduce作业的Reduce阶段中，按Region的个数启动同样个数的Reduce Task，Reduce Task从Map接收数据，然后按Region生成HFile，存放在HDFS临时目录中。

  在MapReduce作业的提交阶段，将HFile从临时目录迁移到HBase目录中。

- **数据导入HDFS**

  在MapReduce作业的Map阶段中从外部数据源抽取数据，并将数据输出到HDFS临时目录下（以“*输出目录*-ldtmp”命名）。

  在MapReduce作业的提交阶段，将文件从临时目录迁移到输出目录中。

- **数据导出到关系型数据库**

  在MapReduce作业的Map阶段，从HDFS或者HBase中抽取数据，然后将数据通过JDBC接口插入到临时表（Staging Table）中。

在MapReduce作业的提交阶段，将数据从临时表迁移到正式表中。

- **数据导出到文件系统**

  在MapReduce作业的Map阶段，从HDFS或者HBase中抽取数据，然后将数据写入到文件服务器临时目录中。

  在MapReduce作业的提交阶段，将文件从临时目录迁移到正式目录。

Loader的架构和详细原理介绍，请参见：**https://sqoop.apache.org/docs/1.99.3/index.html**。

1.4.16.2 Loader 与其他组件的关系

与Loader有交互关系的组件有HDFS、HBase、Hive、Yarn、Mapreduce和ZooKeeper。Loader作为客户端使用这些组件的某些功能，如存储数据到HDFS和HBase，从HDFS和HBase表读数据，同时Loader本身也是一个Mapreduce客户端程序，完成一些数据导入导出任务。

1.4.16.3 Loader 开源增强特性

Loader 开源增强特性：数据导入导出

Loader是在开源Sqoop组件的基础上进行了一些扩展，除了包含Sqoop开源组件本身已有的功能外，还开发了如下的增强特性：

- 提供数据转化功能
- 支持图形化配置转换步骤
- 支持从SFTP/FTP服务器导入数据到HDFS/OBS
- 支持从SFTP/FTP服务器导入数据到HBase表
- 支持从SFTP/FTP服务器导入数据到Phoenix表
- 支持从SFTP/FTP服务器导入数据到Hive表
- 支持从HDFS/OBS导出数据到SFTP/FTP服务器
- 支持从HBase表导出数据到SFTP/FTP服务器
- 支持从Phoenix表导出数据到SFTP/FTP服务器
- 支持从关系型数据库导入数据到HBase表
- 支持从关系型数据库导入数据到Phoenix表
- 支持从关系型数据库导入数据到Hive表
- 支持从HBase表导出数据到关系型数据库
- 支持从Phoenix表导出数据到关系型数据库
- 支持从Oracle分区表导入数据到HDFS/OBS
- 支持从Oracle分区表导入数据到HBase表
- 支持从Oracle分区表导入数据到Phoenix表
- 支持从Oracle分区表导入数据到Hive表
- 支持从HDFS/OBS导出数据到Oracle分区表
- 支持从HBase导出数据到Oracle分区表
- 支持从Phoenix表导出数据到Oracle分区表

- 在同一个集群内，支持从HDFS导数据到HBase、Phoenix表和Hive表
- 在同一个集群内，支持从HBase和Phoenix表导数据到HDFS/OBS
- 导入数据到HBase和Phoenix表时支持使用bulkload和put list两种方式
- 支持从SFTP/FTP导入所有类型的文件到HDFS，开源只支持导入文本文件
- 支持从HDFS/OBS导出所有类型的文件到SFTP，开源只支持导出文本文件和sequence格式文件
- 导入（导出）文件时，支持对文件进行转换编码格式，支持的编码格式为jdk支持的所有格式
- 导入（导出）文件时，支持保持原来文件的目录结构和文件名不变
- 导入（导出）文件时，支持对文件进行合并，如输入文件为海量个文件，可以合并为*n*个文件（*n*值可配）
- 导入（导出）文件时，可以对文件进行过滤，过滤规则同时支持通配符和正则表达式
- 支持批量导入/导出ETL任务
- 支持ETL任务分页查询、关键字查询和分组管理
- 对外部组件提供浮动IP

1.4.17 Manager

1.4.17.1 Manager 基本原理

Manager 功能

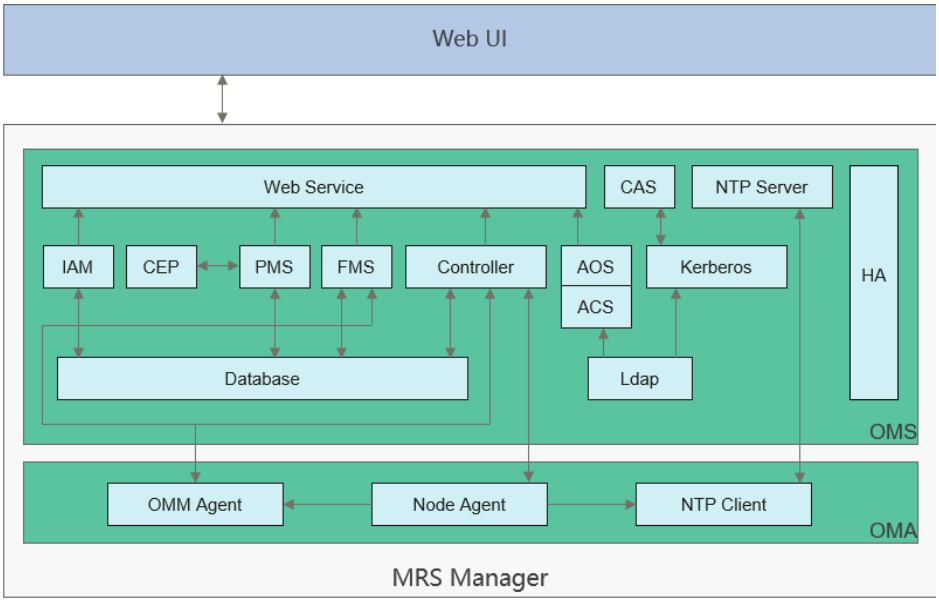
Manager是MRS的运维管理系统，为部署在集群内的服务提供统一的集群管理能力。

Manager支持大规模集群的性能监控、告警、用户管理、权限管理、审计、服务管理、健康检查、日志采集等功能。

Manager 结构

Manager的整体逻辑架构如**图1-75**所示。

图 1-75 Manager 逻辑架构



Manager由OMS和OMA组成：

- OMS：操作维护系统的管理节点，OMS一般有两个，互为主备。
- OMA：操作维护系统中的被管理节点，一般有多个。

图1-75中各模块的说明如表1-16所示：

表 1-16 业务模块说明

模块名称	描述
Web Service	是一个部署在Tomcat下的Web服务，提供Manager的https接口，用于通过浏览器访问Manager。同时还提供基于Syslog和SNMP协议的北向接入能力。
OMS	操作维护系统的管理节点，OMS节点一般有两个，互为主备。
OMA	操作维护系统中的被管理节点，一般有多个。
Controller	Controller是Manager的控制中心，负责汇聚来自集群中所有节点的信息，统一向MRS集群管理员展示，以及负责接收来自MRS集群管理员的操作指令，并且依据操作指令所影响的范围，向集群的所有相关节点同步信息。 Manager的控制进程，负责各种管理动作的执行： 1. Web Service将各种管理动作（安装、启停服务、修改配置等）下发到Controller。 2. Controller将命令分解，分解后将动作下发到每一个Node Agent。例如启动一个服务，会涉及多个角色和实例。 3. Controller负责监控每一个动作的执行情况。

模块名称	描述
Node Agent	Node Agent存在于每一个集群节点，是Manager在单个节点的使能器。 ● Node Agent代表本节点上部署的所有组件与Controller交互，实现整个集群多点到单点的汇聚。 ● Node Agent是Controller对部署在该节点上组件做一切操作的使能器，其代表着Controller的功能。 Node Agent每隔3秒向Controller发送心跳信息，不支持配置时间间隔。
IAM	负责记录审计日志。在Manager的UI上每一个非查询类操作，都有对应的审计日志。
PMS	性能监控模块，搜集每一个OMA上的性能监控数据并提供查询。
CEP	汇聚功能模块。比如将所有OMA上的磁盘已用空间汇总成一个性能指标。
FMS	告警模块，搜集每一个OMA上的告警并提供查询。
OMM Agent	OMA上面性能监控和告警的Agent，负责收集该Agent Node上的性能监控数据和告警数据。
CAS	统一认证中心，登录Web Service时需要在CAS进行登录认证，浏览器通过URL自动跳转访问CAS。
AOS	权限管理模块，管理用户和用户组的权限。
ACS	用户和用户组管理模块，管理用户及用户归属的用户组。
Kerberos	在OMS与集群中各部署一个。 ● OMS Kerberos提供单点登录及Controller与Node Agent间认证的功能。 ● 集群中Kerberos提供组件用户安全认证功能，其服务名称为KrbServer，包含两种角色实例： – KerberosServer：认证服务器，为MRS提供安全认证使用。 – KerberosAdmin：管理Kerberos用户的进程。
Ldap	在OMS与集群中各部署一个。 ● OMS Ldap为用户认证提供数据存储。 ● 集群中的Ldap作为OMS Ldap的备份，其服务名称为LdapServer，角色实例为SlapdServer。
Database	Manager的数据库，负责存储日志、告警等信息。
HA	高可用性管理模块，主备OMS通过HA进行主备管理。
NTP Server NTP Client	负责同步集群内各节点的系统时钟。

1.4.17.2 Manager 关键特性

Manager 关键特性：统一监控告警

Manager提供可视化、便捷的监控告警功能。用户可以快速获取集群关键性能指标，并评测集群健康状态，同时提供性能指标的定制化显示功能及指标转换告警方法。Manager可监控所有组件的运行情况，并在故障时实时上报告警。通过界面的联机帮助，用户可以查看性能指标和告警恢复的详细方法，进行快速排障。

Manager 关键特性：统一用户权限管理

Manager提供系统中各组件的权限集中管理功能。

Manager引入角色的概念，采用RBAC的方式对系统进行权限管理，集中呈现和管理系统中各组件零散的权限功能，并且将各个组件的权限以权限集合（即角色）的形式组织，形成统一的系统权限概念。这样一方面对普通用户屏蔽了内部的权限管理细节，另一方面对MRS集群管理员简化了权限管理的操作方法，提升了权限管理的易用性和用户体验。

Manager 关键特性：单点登录

提供Manager WebUI与组件WebUI之间的单点登录，以及MRS与第三方系统集成时的单点登录。

此功能统一了Manager系统用户和组件用户的管理及认证。整个系统使用LDAP管理用户，使用Kerberos进行认证，并在OMS和组件间各使用一套Kerberos和LDAP的管理机制，通过CAS实现单点登录（包括单点登入和单点登出）。用户只需要登录一次，即可在Manager WebUI和组件Web UI之间，甚至第三方系统之间进行任务跳转操作，无需切换用户重新登录。

说明

- 出于安全考虑，CAS Server只能保留用户使用的TGT（ticket-granting ticket）20分钟。
- 如用户20分钟内不对页面（包括Manager和组件WebUI）进行操作，页面自动锁定。

Manager 关键特性：自动健康检查与巡检

Manager为用户提供界面化的系统运行环境自动检查服务，帮助用户实现一键式系统运行健康度巡检和审计，保障系统的正常运行，降低系统运维成本。用户查看检查结果后，还可导出检查报告用于存档及问题分析。

Manager 关键特性：租户管理

Manager引入了多租户的概念，集群拥有的CPU、内存和磁盘等资源，可以整合规划为一个集合体，这个集合体就是租户。多个不同的租户统称多租户。

多租户功能支持层级式的租户模型，支持动态的添加和删除租户，实现资源的隔离，可以对租户的计算资源和存储资源进行动态配置和管理。

- 计算资源指租户Yarn任务队列资源，可以修改任务队列的配额，并查看任务队列的使用状态和使用统计。
- 存储资源目前支持HDFS存储，可以添加删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额。

Manager作为MRS的统一租户管理平台，用户可以在界面上根据业务需要，在集群中创建租户、管理租户。

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。默认情况下，新的计算资源和存储资源的全部权限将分配给租户的角色。
- 修改租户的计算资源或存储资源，对应的角色关联权限将自动更新。

Manager还提供了多实例的功能，使用户在资源控制和业务隔离的场景中可以独立使用HBase、Hive和Spark组件。多实例功能默认关闭，可以选择手动启用。

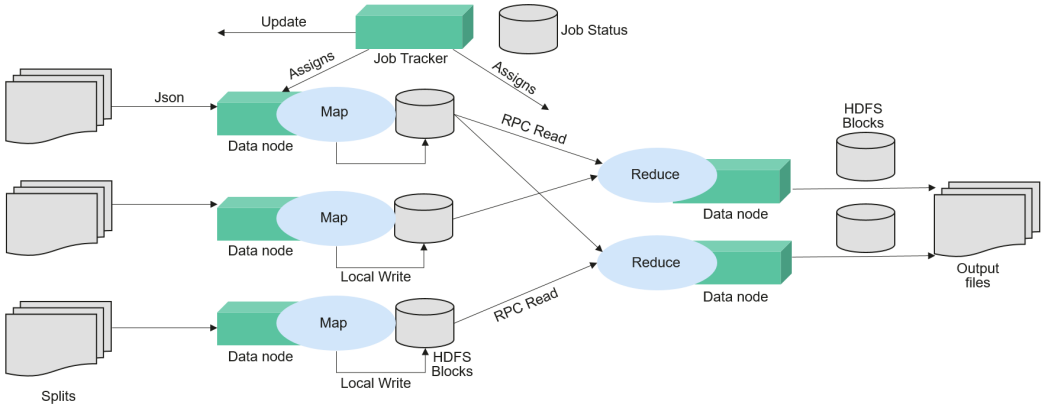
1.4.18 MapReduce

1.4.18.1 MapReduce 基本原理

MapReduce是Hadoop的核心，是Google提出的一个软件架构，用于大规模数据集（大于1TB）的并行运算。概念“Map（映射）”和“Reduce（化简）”，及他们的主要思想，都是从函数式编程语言借来的，还有从矢量编程语言借来的特性。

当前的软件实现是指定一个Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的Reduce（化简）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

图 1-76 分布式批处理引擎



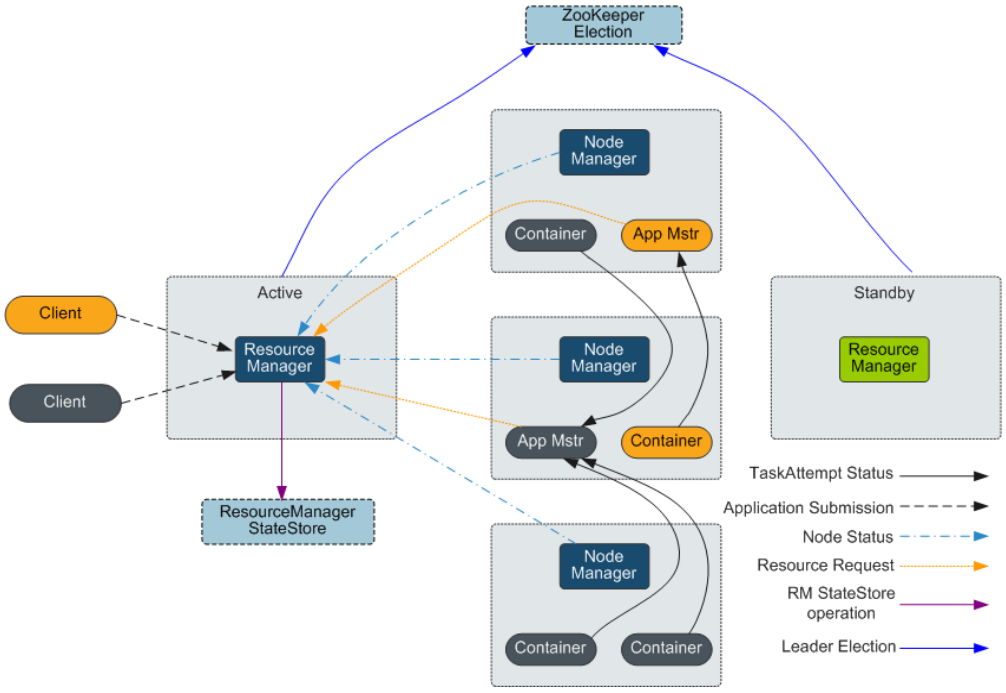
MapReduce是用于并行处理大数据集的软件框架。MapReduce的根源是函数性编程中的map和reduce函数。Map函数接受一组数据并将其转换为一个键/值对列表，输入域中的每个元素对应一个键/值对。Reduce函数接受Map函数生成的列表，然后根据它们的键缩小键/值对列表。MapReduce起到了将大事务分散到不同设备处理的能力，这样原本必须用单台较强服务器才能运行的任务，在分布式环境下也能完成。

更多信息，请参阅**MapReduce教程**。

MapReduce 结构

如**图1-77**所示，MapReduce通过实现YARN的Client和ApplicationMaster接口集成到YARN中，利用YARN申请计算所需资源。

图 1-77 Apache YARN&MapReduce 的基本架构



1.4.18.2 MapReduce 与其他组件的关系

MapReduce 和 HDFS 的关系

- HDFS是Hadoop分布式文件系统，具有高容错和高吞吐量的特性，可以部署在价格低廉的硬件上，存储应用程序的数据，适合有超大数据集的应用程序。
- 而MapReduce是一种编程模型，用于大数据集（大于1TB）的并行运算。在MapReduce程序中计算的数据可以来自多个数据源，如Local FileSystem、HDFS、数据库等。最常用的是HDFS，可以利用HDFS的高吞吐性能读取大规模的数据进行计算。同时在计算完成后，也可以将数据存储到HDFS。

MapReduce 和 YARN 的关系

MapReduce是运行在YARN之上的一个批处理的计算框架。MRv1是Hadoop 1.0中的MapReduce实现，它由编程模型（新旧编程接口）、运行时环境（由JobTracker和TaskTracker组成）和数据处理引擎（MapTask和ReduceTask）三部分组成。该框架在扩展性、容错性（JobTracker单点）和多框架支持（仅支持MapReduce一种计算框架）等方面存在不足。MRv2是Hadoop 2.0中的MapReduce实现，它在源码级重用了MRv1的编程模型和数据处理引擎实现，但运行时环境由YARN的ResourceManager和ApplicationMaster组成。其中ResourceManager是一个全新的资源管理系统，而ApplicationMaster则负责MapReduce作业的数据切分、任务划分、资源申请和任务调度与容错等工作。

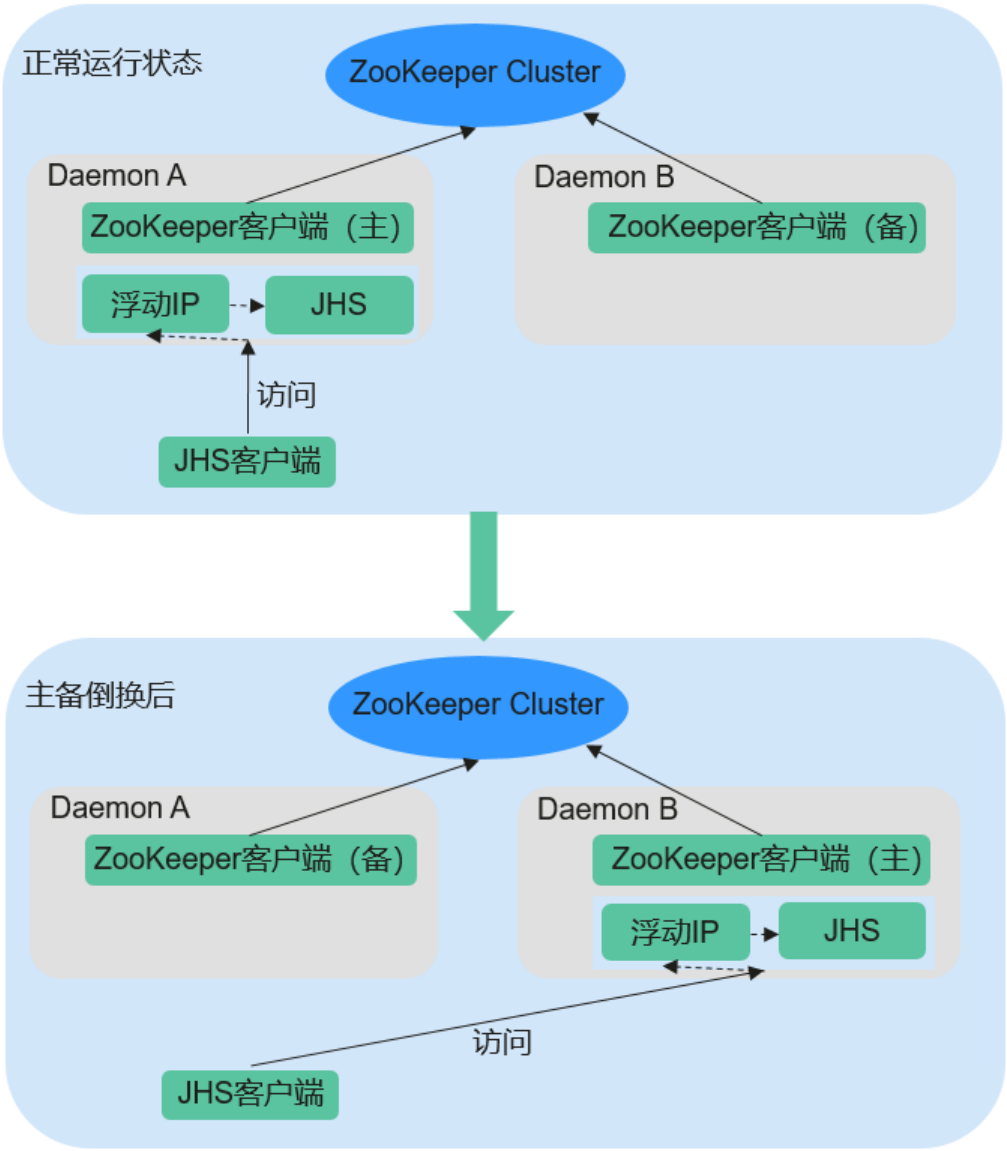
1.4.18.3 MapReduce 开源增强特性

MapReduce 开源增强特性：JobHistoryServer HA 特性

JobHistoryServer（JHS）是用于查看MapReduce历史任务信息的服务器，当前开源JHS只支持单实例服务。JobHistoryServer HA能够解决JHS单点故障时，应用访问

MapReduce接口无效，导致整体应用执行失败的场景，从而大大提升MapReduce服务的高可用性。

图 1-78 JobHistoryServer HA 主备倒换的状态转移过程



JobHistoryServer高可用性

- 采用ZooKeeper实现主备选举和倒换；
- JobHistoryServer使用浮动IP对外提供服务；
- 兼容JHS单实例，也支持HA双实例；
- 同一时刻，只有一个节点启动JHS进程，防止多个JHS操作同一文件冲突；
- 支持扩容减容、实例迁移、升级、健康检查等。

MapReduce 开源增强特性：特定场景优化 MapReduce 的 Merge/Sort 流程提升 MapReduce 性能

下图展示了MapReduce任务的工作流程。

图 1-79 MapReduce job

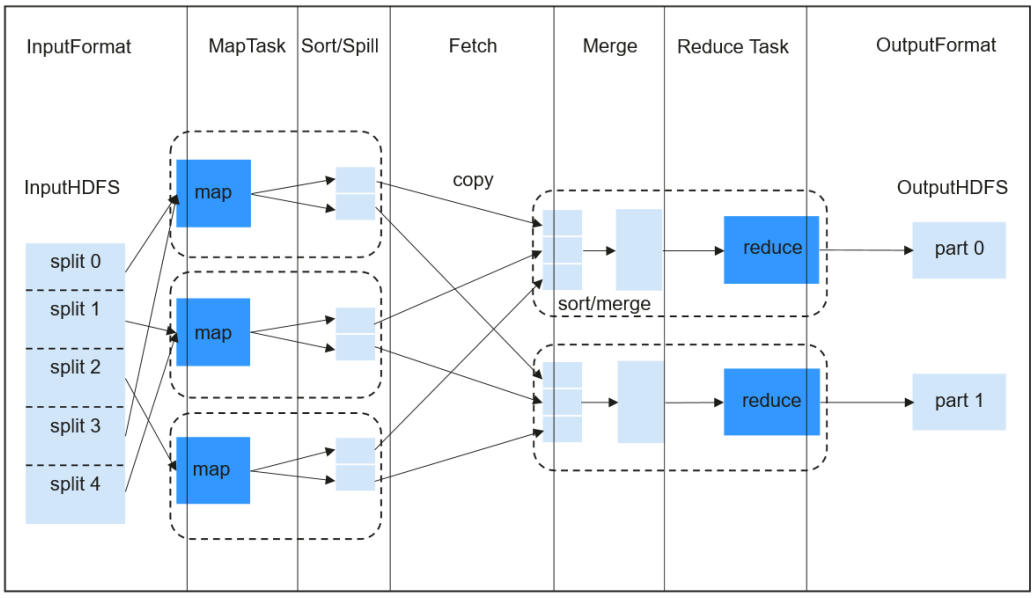
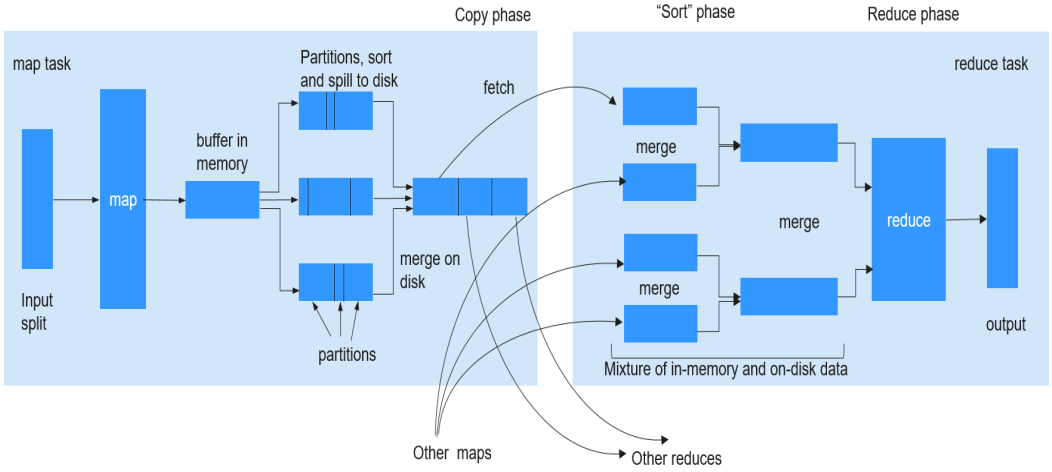


图 1-80 MapReduce job execution flow



Reduce过程分为三个不同步骤：Copy、Sort（实际应当称为Merge）及Reduce。在Copy过程中，Reducer尝试从NodeManagers获取Maps的输出并存储在内存或硬盘中。紧接着进行Shuffle过程（包含Sort及Reduce），这个过程将获取到的Maps输出进行存储并有序地合并然后提供给Reducer。当Job有大量的Maps输出需要处理的时候，Shuffle过程将变得非常耗时。对于一些特定的任务（例如hash join或hash aggregation类型的SQL任务），Shuffle过程中的排序并非必须的。但是Shuffle却默认必须进行排序，所以需要对此处进行改进。

此特性通过对MapReduce API进行增强，能自动针对此类型任务关闭Sort过程。当Sort被关闭，获取Maps输出数据以后，直接合并后输出给Reduce，避免了由于排序而浪费大量时间。这种方式极大程度地提升了大部分SQL任务的效率。

MapReduce 开源增强特性：MR History Server 优化解决日志小文件问题

运行在Yarn上的作业在执行完成后，NodeManager会通过LogAggregationService把产生的日志收集到HDFS上，并从本地文件系统中删除。日志收集到HDFS上以后由MR

HistoryServer来进行统一的日志管理。LogAggregationService在收集日志时会把container产生的本地日志合并成一个日志文件上传到HDFS，在一定程度上可以减少日志文件的数量。但在规模较大且任务繁忙的集群上，经过长时间的运行，HDFS依然会面临存储的日志文件过多的问题。

以一个20节点的计算场景为例，默认清理周期（15日）内将产生约1800万日志文件，占用NameNode近18G内存空间，同时拖慢HDFS的系统响应速度。

由于收集到HDFS上的日志文件只有读取和删除的需求，因此可以利用Hadoop Archives功能对收集的日志文件目录进行定期归档。

日志归档

在MR HistoryServer中新增AggregatedLogArchiveService模块，定期检查日志目录中的文件数。在文件数达到设定阈值时，启动归档任务进行日志归档，并在归档完成后删除原日志文件，以减少HDFS上的文件数量。

归档日志清理

由于Hadoop Archives不支持在归档文件中进行删除操作，因此日志清理时需要删除整个归档文件包。通过修改AggregatedLogDeletionService模块，获取归档日志中最新的日志生成时间，若所有日志文件均满足清理条件，则清理该归档日志包。

归档日志浏览

Hadoop Archives支持URI直接访问归档包中的文件内容，因此浏览过程中，当MR History Server发现原日志文件不存在时，直接将URI重定向到归档文件包中即可访问到已归档的日志文件。

说明

- 本功能通过调用HDFS的Hadoop Archives功能进行日志归档。由于Hadoop Archives归档任务实际上是执行一个MR应用程序，所以在每次执行日志归档任务后，会新增一条MR执行记录。
- 本功能归档的日志来源于日志收集功能，因此只有在日志收集功能开启状态下本功能才会生效。

1.4.19 Oozie

1.4.19.1 Oozie 基本原理

Oozie 简介

Oozie是一个基于工作流引擎的开源框架，它能够提供对Hadoop作业的任务调度与协调。

Oozie 结构

Oozie引擎是一个Web App应用，默认集成到Tomcat中，采用pg数据库。

基于Ext提供WEB Console，该Console仅提供对Oozie工作流的查看和监控功能。通过Oozie对外提REST方式的WS接口，Oozie client通过该接口控制（启动、停止等操作）Workflow流程，从而编排、运行Hadoop MapReduce任务，如图1-81所示。

图 1-81 Oozie 框架

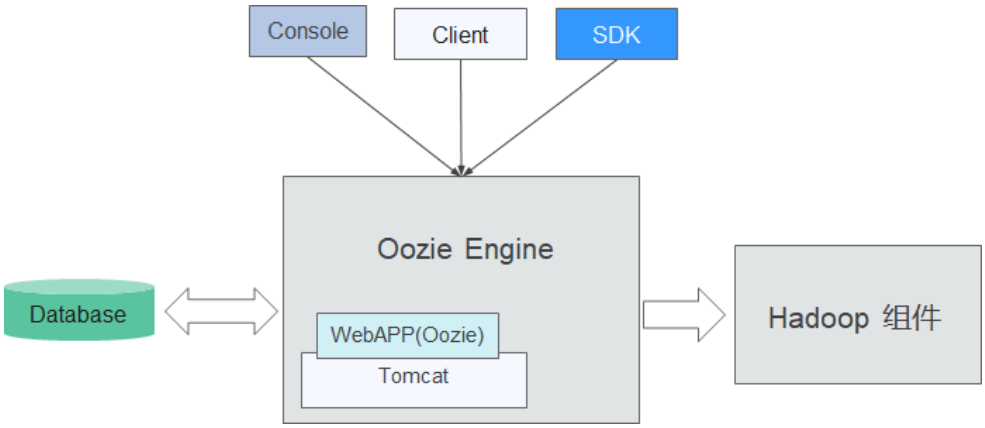


图1-81中各部分的功能说明如表1-17所示。

表 1-17 结构图说明

名称	描述
Console	提供对Oozie流程的查看和监控功能。
Client	通过接口控制workflow流程：可以执行提交流程，启动流程，运行流程，终止流程，恢复流程等操作。
SDK	软件开发工具包SDK（SoftwareDevelopmentKit）是被软件工程师用于为特定的软件包、软件框架、硬件平台、操作系统等建立应用软件的开发工具的集合。
Database	pg数据库。
WebApp（Oozie）	WebApp（Oozie）即Oozie server，可以用内置的Tomcat容器，也可以用外部的，记录的信息比如日志等放在pg数据库中。
Tomcat	Tomcat服务器是免费的开放源代码的Web应用服务器。
Hadoop组件	底层执行Oozie编排流程的各个组件，包括MapReduce、Hive等。

Oozie 原理

Oozie是一个工作流引擎服务器，用于运行MapReduce任务工作流。同时Oozie还是一个Java Web程序，运行在Tomcat容器中。

Oozie工作流通过HPDL（一种通过XML自定义处理的语言，类似JBOSS JBPM的JPDL）来构造。包含“Control Node”（可控制的工作流节点）、“Action Node”。

- “Control Node”用于控制工作流的编排，如“start”（开始）、“end”（关闭）、“error”（异常场景）、“decision”（选择）、“fork”（并行）、“join”（合并）等。
- Oozie工作流中拥有多个“Action Node”，如MapReuce、Java等。

所有的“Action Node”以有向无环图DAG（Direct Acyclic Graph）的模式部署运行。所以在“Action Node”的运行步骤上是有方向的，当上一个“Action Node”运行完成后才能运行下一个“Action Node”。一旦当前“Action Node”完成，远程服务器将回调Oozie的接口，这时Oozie又会以同样的方式执行工作流中的下一个“Action Node”，直到工作流中所有“Action Node”都完成（完成包括失败）。

Oozie工作流提供各种类型的“Action Node”用于支持不同的业务需要，如MapReduce，HDFS，SSH，Java以及Oozie子流程。

1.4.19.2 Oozie 开源增强特性

Oozie 开源增强特性：安全增强

支持Oozie权限管理，提供管理员与普通用户两种角色。

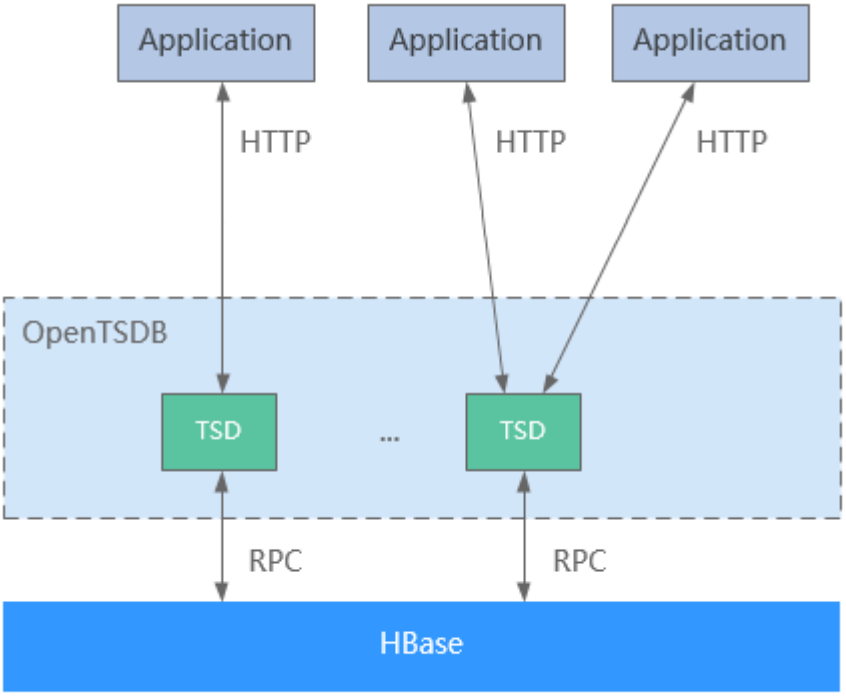
支持单点登录登出，HTTPS访问以及审计日志。

1.4.20 OpenTSDB

OpenTSDB是一个基于HBase的分布式、可伸缩的时间序列数据库。OpenTSDB的设计目标是用来采集大规模集群中的监控类信息，并可实现数据的秒级查询，解决海量监控类数据在普通数据库中查询存储的局限性。

OpenTSDB由时间序列守护进程（TSD）和一组命令行实用程序组成。与OpenTSDB的交互主要通过运行一个或多个TSD来实现。每个TSD都是独立的。没有主服务器，没有共享状态，因此您可以根据需要运行任意数量的TSD来处理您向其投入的任何负载。每个TSD使用CloudTable集群中的HBase来存储和检索时间序列数据。数据模式经过高度优化，可快速聚合相似的时间序列，从而最大限度地减少存储空间。TSD的用户不需要直接访问底层存储。您可以通过HTTP API与TSD进行通信。所有通信都发生在同一个端口上（TSD通过查看它收到的前几个字节来确定客户端的协议）。

图 1-82 OpenTSDB 架构



OpenTSDB使用场景有如下几个特点：

- 采集指标在某一时间点具有唯一值，没有复杂的结构及关系。
- 监控的指标具有随着时间不断变化的特点。
- 具有HBase的高吞吐，良好的伸缩性等特点。

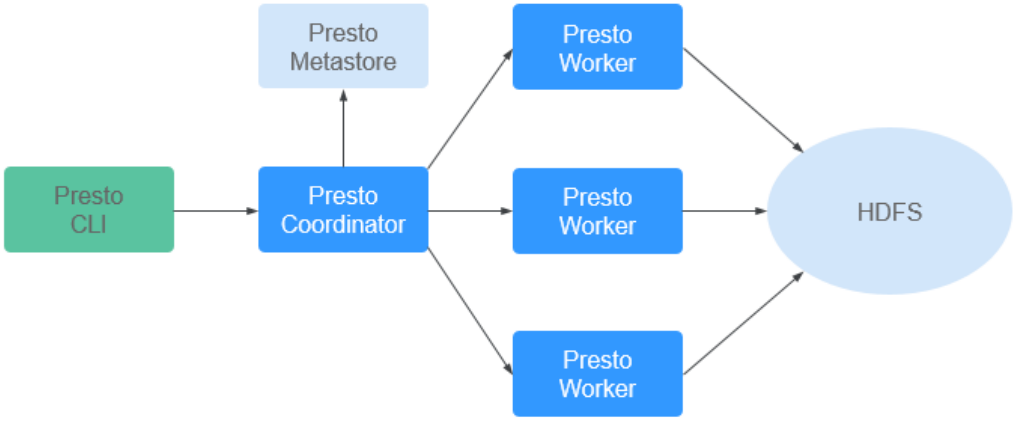
OpenTSDB提供基于HTTP的应用程序编程接口，以实现与外部系统的集成。 几乎所有OpenTSDB功能都可通过API访问，例如查询时间序列数据，管理元数据和存储数据点。详情请参见：**https://opentsdb.net/docs/build/html/api_http/index.html**。

1.4.21 Presto

Presto是一个开源的用户交互式分析查询的SQL查询引擎，用于针对各种大小的数据源进行交互式分析查询。其主要应用于海量结构化数据/半结构化数据分析、海量多维数据聚合/报表、ETL、Ad-Hoc查询等场景。

Presto允许查询的数据源包括Hadoop分布式文件系统（HDFS），Hive，HBase，Cassandra，关系数据库甚至专有数据存储。一个Presto查询可以组合不同数据源，执行跨数据源的数据分析。

图 1-83 Presto 架构



Presto分布式地运行在一个集群中，包含一个Coordinator和多个Worker进程，查询从客户端（例如CLI）提交到Coordinator，Coordinator进行SQL的解析和生成执行计划，然后分发到多个Worker进程上执行。

Presto 多实例

MRS支持为大规格的集群默认安装Presto多实例，即一个Core/Task节点上安装多个Worker实例，分别为Worker1，Worker2，Worker3…，多个Worker实例共同与Coordinator交互执行计算任务，相比较单实例，能够大大提高节点资源的利用率和计算效率。

Presto多实例仅作用于ARM架构规格，当前单节点最多支持4个实例。

1.4.22 Ranger

1.4.22.1 Ranger 基本原理

Apache Ranger提供一个集中式安全管理框架，提供统一授权和统一审计能力。它可以对整个Hadoop生态中如HDFS、Hive、HBase、Kafka、Storm等进行细粒度的数据访问控制。用户可以利用Ranger提供的前端WebUI控制台通过配置相关策略来控制用户对这些组件的访问权限 。

Ranger架构如**图1-84**所示

图 1-84 Ranger 结构

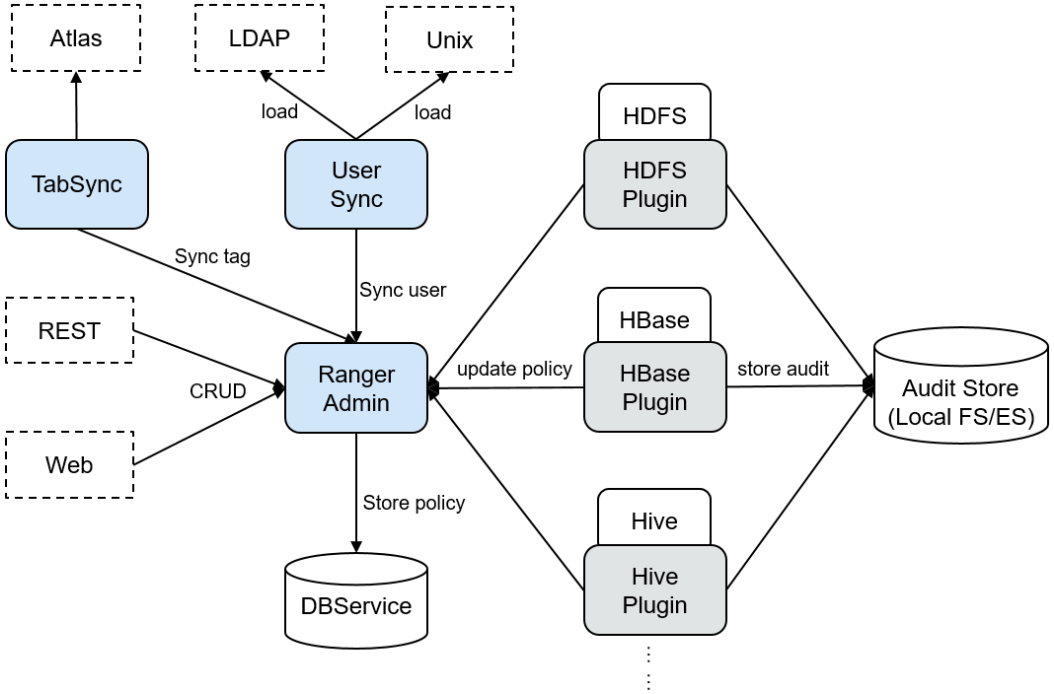


表 1-18 结构图说明

名称	描述
RangerAdmin	Ranger的管理角色，拥有策略管理、用户管理、审计管理等功能，提供WebUI和RestFul接口。
UserSync	负责周期从外部同步用户和用户组信息并写入RangerAdmin中。
TagSync	负责周期从外部Atlas服务同步标签信息并写入RangerAdmin中。

Ranger 原理

- 组件Ranger插件

  Ranger为各组件提供了基于PBAC（Policy-Based Access Control）的权限管理插件，用于替换组件自身原本的鉴权插件。Ranger插件都是由组件侧自身的鉴权接口扩展而来，用户在Ragner WebUI上对指定service设置权限策略，Ranger插件会定期从RangerAdmin处更新策略并缓存在组件本地文件，当有客户端请求需要进行鉴权时，Ranger插件会对请求中携带的用户在策略中进行匹配，随后返回接受或拒绝。

- UserSync用户同步

  UserSync周期性从LDAP/Unix中同步数据到RangerAdmin中，其中安全模式向从LDAP中同步，非安全模式从Unix中同步。同步模式默认采取增量模式，每次同步周期UserSync只会更新新增或者变更的用户和用户组，当用户或者用户组被删除时，UserSync不会同步该变更到RangerAdmin，即RangerAdmin中不会同步删除。为了提高性能，UserSync也不会同步没有所属用户的用户组到RangerAdmin中。

- 统一审计

  Ranger插件支持记录审计日志，当前审计日志存储介质支持本地文件。

- 高可靠性

  Ranger支持RangerAdmin双主，两个RangerAdmin同时提供服务，任意一个RangerAdmin故障不会影响Ranger的功能。

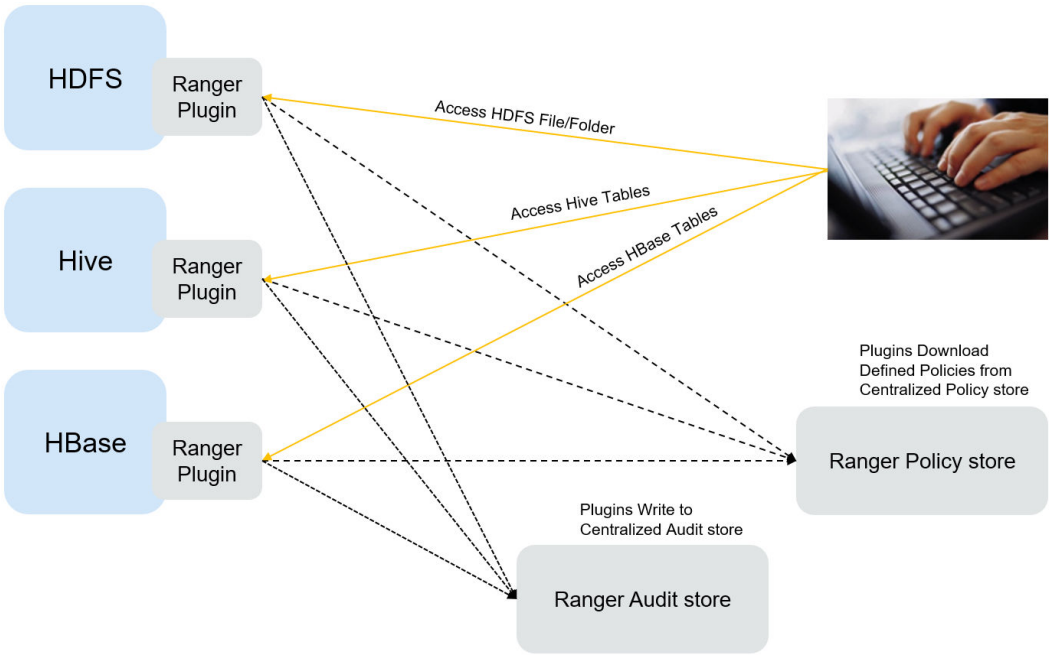
- 高性能

  Ranger提供Load-Balance能力，通过浏览器访问Ranger WebUI时Load-Balance会自动选择当前负载较小的RangerAdmin来提供服务。

1.4.22.2 Ranger 与其他组件的关系

Ranger为组件提供基于PBAC的鉴权插件，供组件服务端运行，目前支持Ranger鉴权的组件有HDFS、Yarn、Hive、HBase、Kafka、Storm和Spark2x，后续会支持更多组件。

图 1-85 Ranger 与组件的关系



1.4.23 Spark

1.4.23.1 Spark 基本原理

说明

Spark组件适用于MRS 3.x之前版本。

Spark 简介

Spark是一个开源的，并行数据处理框架，能够帮助用户简单的开发快速，统一的大数据应用，对数据进行离线处理，流式处理，交互式分析等等。

Spark提供了一个快速的计算，写入，以及交互式查询的框架。相比于Hadoop，Spark拥有明显的性能优势。Spark使用in-memory的计算方式，通过这种方式来避免一个MapReduce工作流中的多个任务对同一个数据集进行计算时的IO瓶颈。Spark利用Scala语言实现，Scala能够使得处理分布式数据集时，能够像处理本地化数据一样。除了交互式的数据分析，Spark还能够支持交互式的数据挖掘，由于Spark是基于内存的计算，很方便处理迭代计算，而数据挖掘的问题通常都是对同一份数据进行迭代计算。除此之外，Spark能够运行于安装Hadoop 2.0 Yarn的集群。之所以Spark能够在保留MapReduce容错性，数据本地化，可扩展性等特性的同时，能够保证性能的高效，并且避免繁忙的磁盘IO，主要原因是因为Spark创建了一种叫做RDD（Resilient Distributed Dataset）的内存抽象结构。

原有的分布式内存抽象，例如key-value store以及数据库，支持对于可变状态的细粒度更新，这一点要求集群需要对数据或者日志的更新进行备份来保障容错性。这样就会给数据密集型的工作流带来大量的IO开销。而对于RDD来说，它只有一套受限制的接口，仅支持粗粒度的更新，例如map，join等等。通过这种方式，Spark只需要简单的记录建立数据的转换操作的日志，而不是完整的数据集，就能够提供容错性。这种数据的转换链记录就是数据集的溯源。由于并行程序，通常是对一个大数据集应用相同的计算过程，因此之前提到的粗粒度的更新限制并没有想象中的大。事实上，Spark论文中阐述了RDD完全可以作为多种不同计算框架，例如MapReduce，Pregel等的编程模型。并且，Spark同时提供了操作允许用户显式地将数据转换过程持久化到硬盘。对于数据本地化，是通过允许用户能够基于每条记录的键值，控制数据分区实现的。（采用这种方式的一个明显好处是，能够保证两份需要进行关联的数据将会被同样的方式进行哈希）。如果内存的使用超过了物理限制，Spark将会把这些比较大的分区写入到硬盘，由此来保证可扩展性。

Spark具有如下特点：

- 快速：数据处理能力，比MapReduce快10-100倍。
- 易用：可以通过Java，Scala，Python，简单快速的编写并行的应用处理大数据量，Spark提供了超过80种的操作符来帮助用户组件并行程序。
- 普遍性：Spark提供了众多的工具。可以在一个应用中，方便的将这些工具进行组合。
- 与Hadoop集成：Spark能够直接运行于Hadoop的集群，并且能够直接读取现存的Hadoop数据。

MRS服务的Spark组件具有以下优势：

- MRS服务的Spark Streaming组件支持数据实时处理能力而非定时触发。
- MRS服务的Spark组件支持Structured Streaming，支持DataSet API来构建流式应用，提供了exactly-once的语义支持，流和流的join操作支持内连接和外连接。
- MRS服务的Spark组件支持pandas_udf，可以利用pandas_udf替代pyspark中原来的udf对数据进行处理，可以减少60%-90%的处理时长（受具体操作影响）。
- MRS服务的Spark组件支持 Graph 功能，支持图计算作业使用图进行建模。
- MRS服务的SparkSQL兼容部分Hive语法（以Hive-Test-benchmark测试集上的64个SQL语句为准）和标准SQL语法（以tpc-ds测试集上的99个SQL语句为准）。

Spark 结构

Spark的结构如**图1-86**所示，各模块的说明如**表 基本概念说明**所示。

图 1-86 Spark 结构

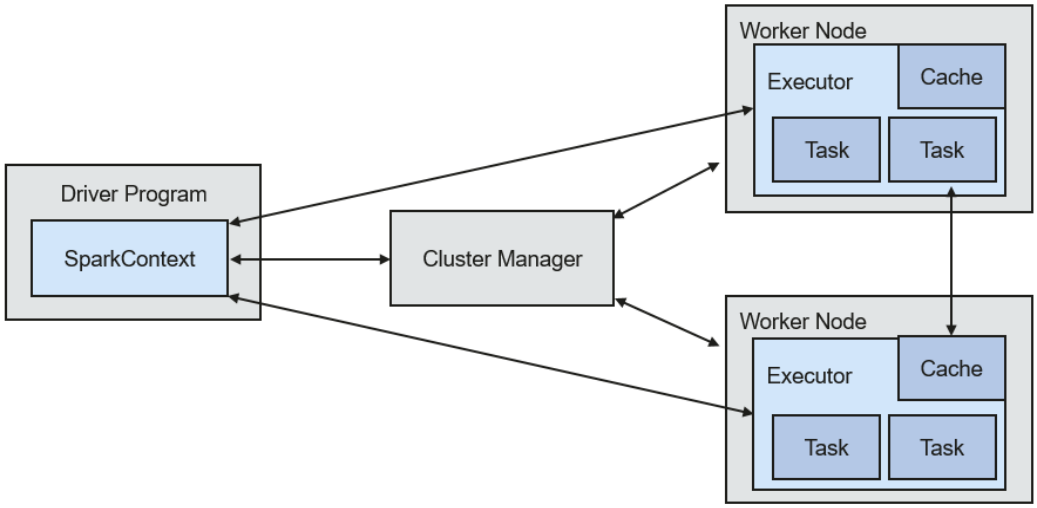


表 1-19 基本概念说明

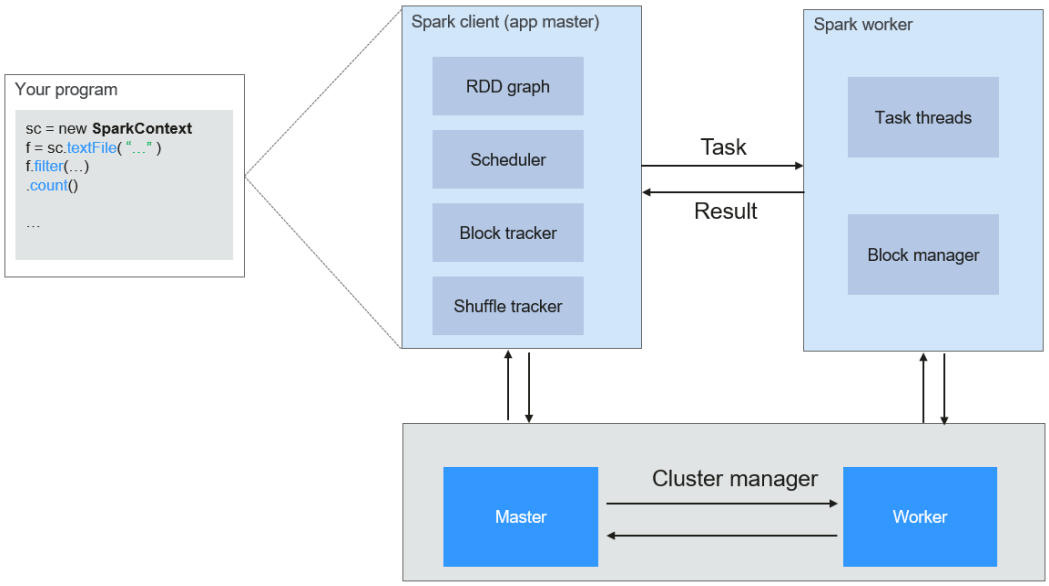
模块	说明
Cluster Manager	集群管理器，管理集群中的资源。Spark支持多种集群管理器，Spark自带的Standalone集群管理器、Mesos或YARN，系统默认采用YARN模式。
Application	Spark应用，由一个Driver Program和多个Executor组成。
Deploy Mode	部署模式，分为cluster和client模式。cluster模式下，Driver会在集群内的节点运行；而在client模式下，Driver在客户端运行（集群外）。
Driver Program	是Spark应用程序的主进程，运行Application的main()函数并创建SparkContext。负责应用程序的解析、生成Stage并调度Task到Executor上。通常SparkContext代表Driver Program。
Executor	在Work Node上启动的进程，用来执行Task，管理并处理应用中使用到的数据。一个Spark应用一般包含多个Executor，每个Executor接收Driver的命令，并执行一到多个Task。
Worker Node	集群中负责启动并管理Executor以及资源的节点。
Job	一个Action算子（比如collect算子）对应一个Job，由并行计算的多个Task组成。
Stage	每个Job由多个Stage组成，每个Stage是一个Task集合，由DAG分割而成。
Task	承载业务逻辑的运算单元，是Spark平台上可执行的最小工作单元。一个应用根据执行计划以及计算量分为多个Task。

Spark 应用运行原理

Spark的应用运行架构如**图 Spark应用运行架构**所示，运行流程如下所示：

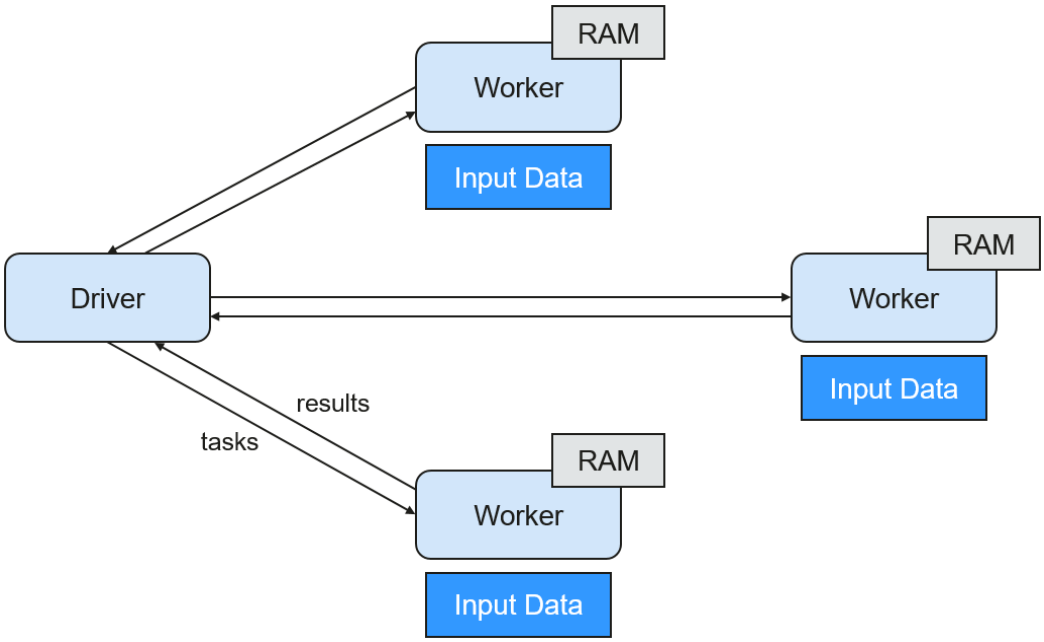
1. 应用程序（Application）是作为一个进程的集合运行在集群上的，由Driver进行协调。
2. 在运行一个应用时，Driver会去连接集群管理器（Standalone、Mesos、YARN）申请运行Executor资源，并启动ExecutorBackend。然后由集群管理器在不同的应用之间调度资源。Driver同时会启动应用程序DAG调度、Stage划分、Task生成。
3. 然后Spark会把应用的代码（传递给**SparkContext**的JAR或者Python定义的代码）发送到Executor上。
4. 所有的Task执行完成后，用户的应用程序运行结束。

图 1-87 Spark 应用运行架构



Spark采用Master和Worker的模式，如**图 Spark的Master和Worker**所示。用户在Spark客户端提交应用程序，调度器将Job分解为多个Task发送到各个Worker中执行，各个Worker将计算的结果上报给Driver（即Master），Driver聚合结果返回给客户端。

图 1-88 Spark 的 Master 和 Worker



在此结构中，有几个说明点：

- 应用之间是独立的。

  每个应用有自己的executor进程，Executor启动多个线程，并行地执行任务。无论是在调度方面，或者是executor方面。各个Driver独立调度自己的任务；不同的应用任务运行在不同的JVM上，即不同的Executor。
- 不同Spark应用之间是不共享数据的，除非把数据存储在外部的存储系统上（比如HDFS）。
- 因为Driver程序在集群上调度任务，所以Driver程序最好和worker节点比较近，比如在一个相同的局部网络内。

Spark on YARN有两种部署模式：

- yarn-cluster模式下，Spark的Driver会运行在YARN集群内的ApplicationMaster进程中，ApplicationMaster已经启动之后，提交任务的客户端退出也不会影响任务的运行。
- yarn-client模式下，Driver启动在客户端进程内，ApplicationMaster进程只用来向YARN集群申请资源。

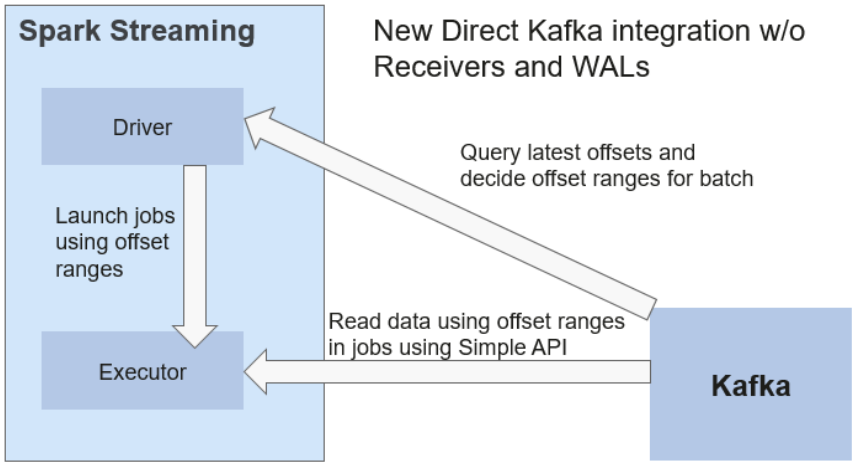
Spark Streaming 原理

Spark Streaming是一种构建在Spark上的实时计算框架，扩展了Spark处理大规模流式数据的能力。当前Spark支持两种数据处理方式：

- Direct Streaming

  Direct Streaming方式主要通过采用Direct API对数据进行处理。以Kafka Direct接口为例，与启动一个Receiver来连续不断地从Kafka中接收数据并写入到WAL中相比，Direct API简单地给出每个batch区间需要读取的偏移量位置。然后，每个batch的Job被运行，而对应偏移量的数据在Kafka中已准备好。这些偏移量信息也被可靠地存储在checkpoint文件中，应用失败重启时可以直接读取偏移量信息。

图 1-89 Direct Kafka 接口数据传输



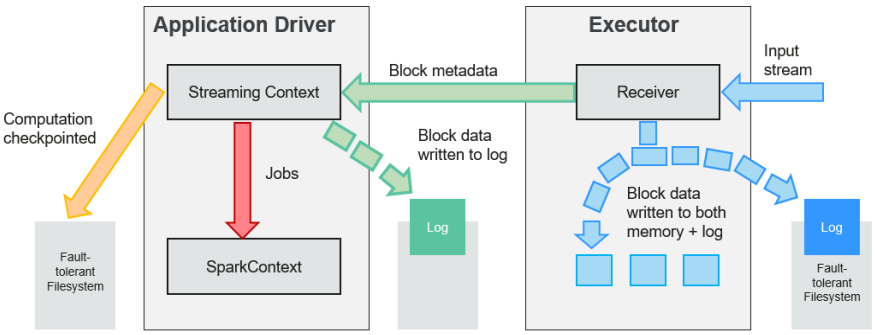
需要注意的是，Spark Streaming可以在失败后重新从Kafka中读取并处理数据段。然而，由于语义仅被处理一次，重新处理的结果和没有失败处理的结果是一致的。

因此，Direct API消除了需要使用WAL和Receivers的情况，且确保每个Kafka记录仅被接收一次，这种接收更加高效。使得Spark Streaming和Kafka可以很好地整合在一起。总体来说，这些特性使得流处理管道拥有高容错性、高效性及易用性，因此推荐使用Direct Streaming方式处理数据。

- Receiver

  在一个Spark Streaming应用开始时（也就是Driver开始时），相关的StreamingContext（所有流功能的基础）使用SparkContext启动Receiver成为长驻运行任务。这些Receiver接收并保存流数据到Spark内存中以供处理。用户传送数据的生命周期如图1-90所示：

  图 1-90 数据传输生命周期

  a. 接收数据（蓝色箭头）

     Receiver将数据流分成一系列小块，存储到Executor内存中。另外，在启用预写日志（Write-ahead Log，简称WAL）以后，数据同时还写入到容错文件系统的预写日志中。

  b. 通知Driver（绿色箭头）

     接收块中的元数据被发送到Driver的StreamingContext。这个元数据包括：

     - 定位其在Executor内存中数据位置的块Reference ID。

- 若启用了WAL，还包括块数据在日志中的偏移信息。

c. 处理数据（红色箭头）

对每个批次的数据，StreamingContext使用Block信息产生RDD及其Job。StreamingContext通过运行任务处理Executor内存中的Block来执行Job。

d. 周期性地设置检查点（橙色箭头）

为了容错的需要，StreamingContext会周期性地设置检查点，并保存到外部文件系统中。

容错性

Spark及其RDD允许无缝地处理集群中任何Worker节点的故障。鉴于Spark Streaming建立于Spark之上，因此其Worker节点也具备了同样的容错能力。然而，由于Spark Streaming的长正常运行需求，其应用程序必须也具备从Driver进程（协调各个Worker的主要应用进程）故障中恢复的能力。使Spark Driver能够容错是件很棘手的事情，因为可能是任意计算模式实现的任意用户程序。不过Spark Streaming应用程序在计算上有一个内在的结构：在每批次数据周期性地执行同样的Spark计算。这种结构允许把应用的状态（也叫做Checkpoint）周期性地保存到可靠的存储空间中，并在Driver重新启动时恢复该状态。

对于文件这样的源数据，这个Driver恢复机制足以做到零数据丢失，因为所有的数据都保存在了像HDFS这样的容错文件系统中。但对于像Kafka和Flume等其他数据源，有些接收到的数据还只缓存在内存中，尚未被处理，就有可能会丢失。这是由于Spark应用的分布操作方式引起的。当Driver进程失败时，所有在Cluster Manager中运行的Executor，连同在内存中的所有数据，也同时被终止。为了避免这种数据损失，Spark Streaming引进了WAL功能。

WAL通常被用于数据库和文件系统中，用来保证任何数据操作的持久性，即先将操作记入一个持久的日志，再对数据施加这个操作。若施加操作的过程中执行失败了，则通过读取日志并重新施加前面预定的操作，系统就得到了恢复。下面介绍了如何利用这样的概念保证接收到的数据的持久性。

Kafka数据源使用Receiver来接收数据，是Executor中的长运行任务，负责从数据源接收数据，并且在数据源支持时还负责确认收到数据的结果（收到的数据被保存在Executor的内存中，然后Driver在Executor中运行来处理任务）。

当启用了预写日志以后，所有收到的数据同时还保存到了容错文件系统的日志文件中。此时即使Spark Streaming失败，这些接收到的数据也不会丢失。另外，接收数据的正确性只在数据被预写到日志以后Receiver才会确认，已经缓存但还没有保存的数据可以在Driver重新启动之后由数据源再发送一次。这两个机制确保了零数据丢失，即所有的数据或者从日志中恢复，或者由数据源重发。

如果需要启用预写日志功能，可以通过如下动作实现：

- 通过“streamingContext.checkpoint”设置checkpoint的目录，这个目录是一个HDFS的文件路径，既用作保存流的checkpoint，又用作保存预写日志。
- 设置SparkConf的属性“spark.streaming.receiver.writeAheadLog.enable”为“true”（默认值是“false”）。

在WAL被启用以后，所有Receiver都获得了能够从可靠收到的数据中恢复的优势。建议缓存RDD时不采取多备份选项，因为用于预写日志的容错文件系统很可能也复制了数据。

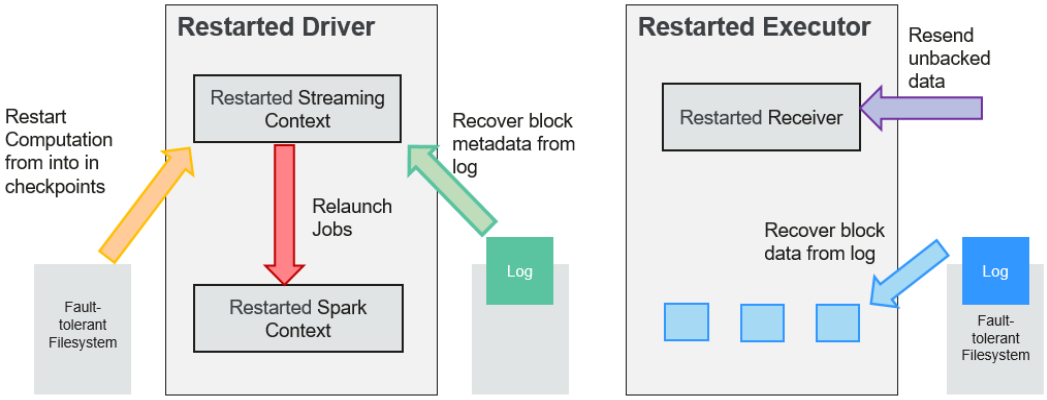
说明

在启用了预写日志以后，数据接收吞吐率会有降低。由于所有数据都被写入容错文件系统，文件系统的写入吞吐率和用于数据复制的网络带宽，可能就是潜在的瓶颈了。在此情况下，最好创建更多的Recevier增加数据接收的并行度，或使用更好的硬件以增加容错文件系统的吞吐率。

恢复流程

当一个失败的Driver重启时，按如下流程启动：

图 1-91 计算恢复流程



1. 恢复计算（橙色箭头）

   使用checkpoint信息重启Driver，重新构造SparkContext并重启Receiver。

2. 恢复元数据块（绿色箭头）

   为了保证能够继续下去所必备的全部元数据块都被恢复。

3. 未完成作业的重新形成（红色箭头）

   由于失败而没有处理完成的批处理，将使用恢复的元数据再次产生RDD和对应的作业。

4. 读取保存在日志中的块数据（蓝色箭头）

   在这些作业执行时，块数据直接从预写日志中读出。这将恢复在日志中可靠地保存的所有必要数据。

5. 重发尚未确认的数据（紫色箭头）

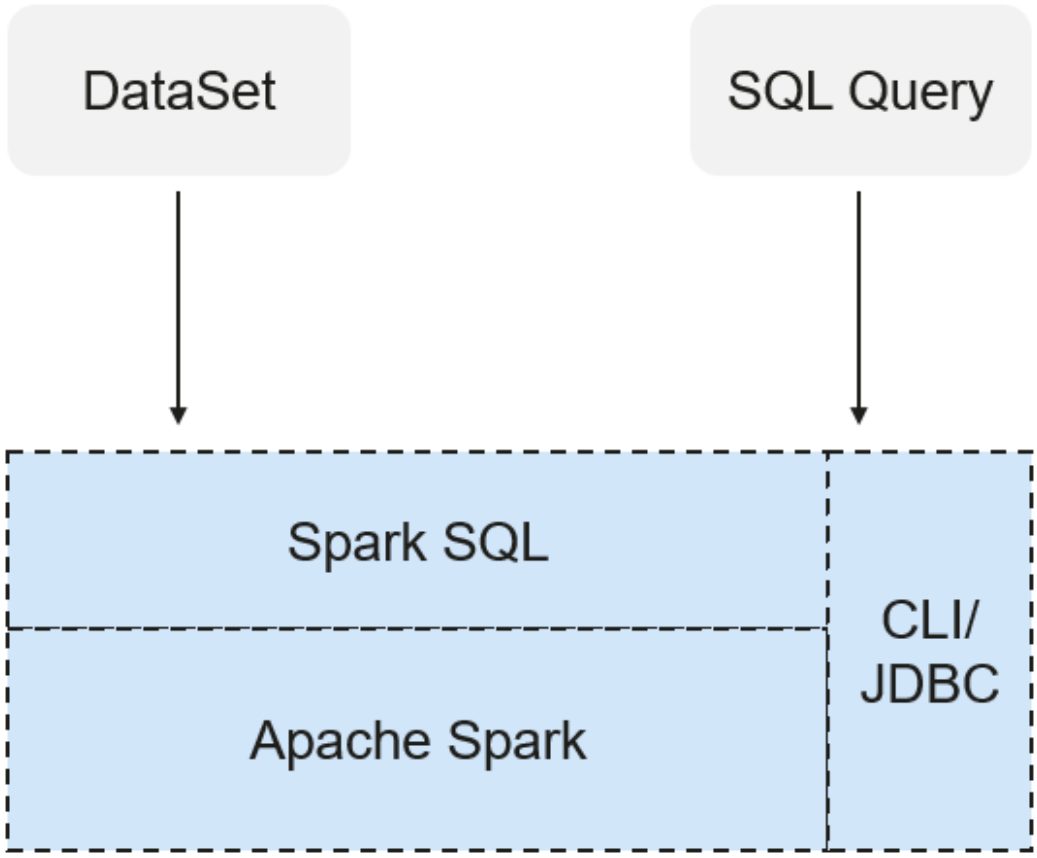
   失败时没有保存到日志中的缓存数据将由数据源再次发送。因为Receiver尚未对其确认。

因此通过预写日志和可靠的Receiver,Spark Streaming就可以保证没有输入数据会由于Driver的失败而丢失。

SparkSQL 和 DataSet 原理

SparkSQL

图 1-92 SparkSQL 和 DataSet



Spark SQL是Spark中用于结构化数据处理的模块。在Spark应用中，可以无缝的使用SQL语句亦或是DataSet API对结构化数据进行查询。

Spark SQL以及DataSet还提供了一种通用的访问多数据源的方式，可访问的数据源包括Hive、CSV、Parquet、ORC、JSON和JDBC数据源，这些不同的数据源之间也可以实现互相操作。Spark SQL复用了Hive的前端处理逻辑和元数据处理模块，使用Spark SQL可以直接对已有的Hive数据进行查询。

另外，SparkSQL还提供了诸如API、CLI、JDBC等诸多接口，对客户端提供多样接入形式。

Spark SQL Native DDL/DML

Spark 1.5版本将很多DDL/DML命令下压到Hive执行，造成了与Hive的耦合，且在一定程度上不够灵活（比如报错不符合预期、结果与预期不一致等）。

Spark 3.1.1版本实现了命令的本地化，使用Spark SQL Native DDL/DML取代Hive执行DDL/DML命令。一方面实现和Hive的解耦，另一方面可以对命令进行定制化。

DataSet

DataSet是一个由特定域的对象组成的强类型集合，可通过功能或关系操作并行转换其中的对象。 每个Dataset还有一个非类型视图，即由多个列组成的DataSet，称为DataFrame。

DataFrame是一个由多个列组成的结构化的分布式数据集合，等同于关系数据库中的一张表，或者是R/Python中的data frame。DataFrame是Spark SQL中的最基本的概

念，可以通过多种方式创建，例如结构化的数据集、Hive表、外部数据库或者是RDD。

可用于DataSet的操作分为Transformation和Action：

- Transformation操作可生成新的DataSet。
  如map、filter、select和aggregate (groupBy)。
- Action操作可触发计算及返回记结果。
  如count、show或向文件系统写数据。

通常使用以下两种方法创建一个DataSet：

- 最常见的方法是通过使用SparkSession上的read函数将Spark指向存储系统上的某些文件。
  ```
  val people = spark.read.parquet("...").as[Person]  // Scala
  DataSet<Person> people = spark.read().parquet("...").as(Encoders.bean(Person.class));//Java
  ```
- 还可通过已存在的DataSet上可用的transformation操作来创建数据集。
  例如，在已存在的DataSet上应用map操作来创建新的DataSet：
  ```
  val names = people.map(_.name)  // 使用Scala语言，且names为一个Datase
  Dataset<String> names = people.map((Person p) -> p.name, Encoders.STRING)); // Java
  ```

CLI和JDBCServer

除了API编程接口之外，Spark SQL还对外提供CLI/JDBC接口：

- spark-shell和spark-sql脚本均可以提供CLI，以便于调试。
- JDBCServer提供JDBC接口，外部可直接通过发送JDBC请求来完成结构化数据的计算和解析。

SparkSession 原理

SparkSession是Spark编程的统一API，也可看作是读取数据的统一入口。SparkSession提供了一个统一的入口点来执行以前分散在多个类中的许多操作，并且还为那些较旧的类提供了访问器方法，以实现最大的兼容性。

使用构建器模式创建SparkSession。如果存在SparkSession，构建器将自动重用现有的SparkSession；如果不存在则会创建一个SparkSession。 在I/O期间，在构建器中设置的配置项将自动同步到Spark和Hadoop。

```
import org.apache.spark.sql.SparkSession
val sparkSession = SparkSession.builder
  .master("local")
  .appName("my-spark-app")
  .config("spark.some.config.option", "config-value")
  .getOrCreate()
```

- SparkSession可以用于对数据执行SQL查询，将结果返回为DataFrame。
  ```
  sparkSession.sql("select * from person").show
  ```
- SparkSession可以用于设置运行时的配置项，这些配置项可以在SQL中使用变量替换。
  ```
  sparkSession.conf.set("spark.some.config", "abcd")
  sparkSession.conf.get("spark.some.config")
  sparkSession.sql("select ${spark.some.config}")
  ```
- SparkSession包括一个“catalog”方法，其中包含使用Metastore（即数据目录）的方法。方法返回值为数据集，可以使用相同的Dataset API来运行。
  ```
  val tables = sparkSession.catalog.listTables()
  val columns = sparkSession.catalog.listColumns("myTable")
  ```

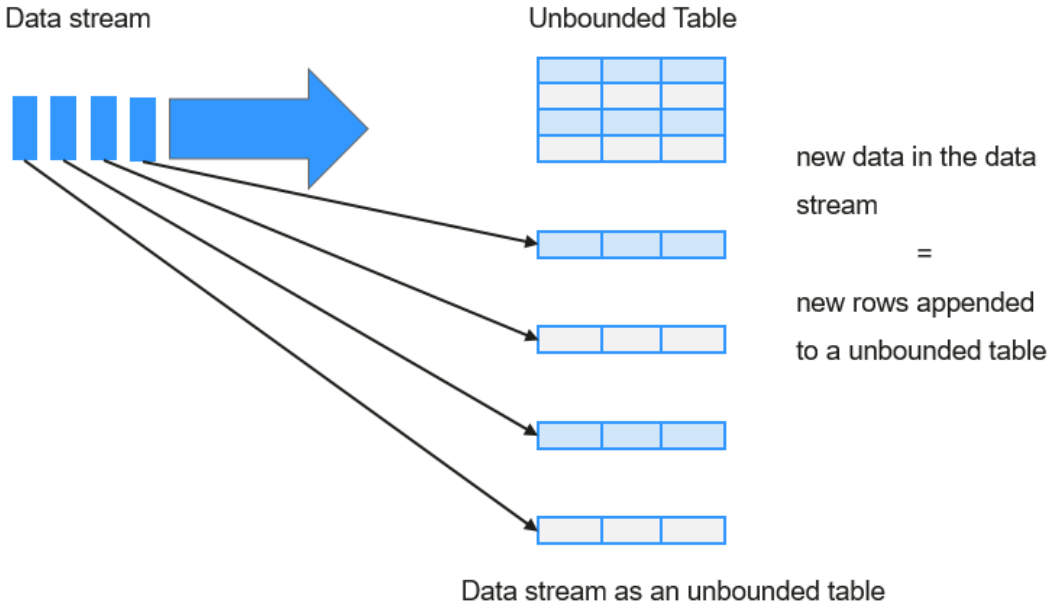
- 底层SparkContext可以通过SparkSession的SparkContext API访问。
  ```
  val sparkContext = sparkSession.sparkContext
  ```

Structured Streaming 原理

Structured Streaming是构建在Spark SQL引擎上的流式数据处理引擎，用户可以使用Scala、Java、Python或R中的Dataset/DataFrame API进行流数据聚合运算、按事件时间窗口计算、流流Join等操作。当流数据连续不断的产生时，Spark SQL将会增量的、持续不断的处理这些数据并将结果更新到结果集中。同时，系统通过checkpoint和Write Ahead Logs确保端到端的完全一次性容错保证。

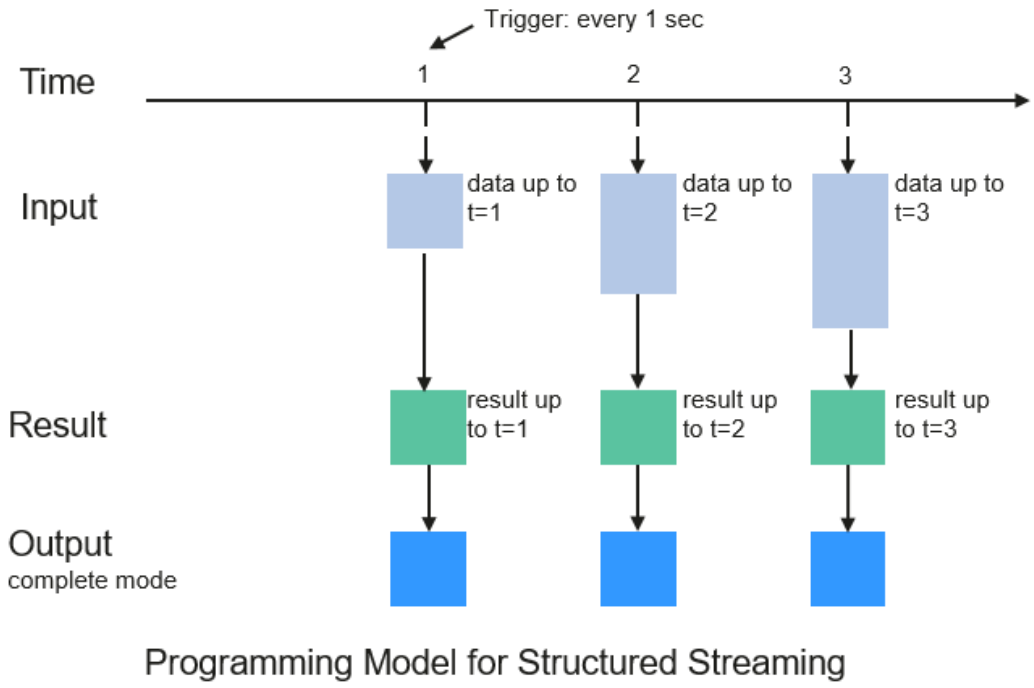
Structured Streaming的核心是将流式的数据看成一张不断增加的数据库表，这种流式的数据处理模型类似于数据块处理模型，可以把静态数据库表的一些查询操作应用在流式计算中，Spark执行标准的SQL查询，从不断增加的无边界表中获取数据。

图 1-93 Structured Streaming 无边界表



每一条查询的操作都会产生一个结果集Result Table。每一个触发间隔，当新的数据新增到表中，都会最终更新Result Table。无论何时结果集发生了更新，都能将变化的结果写入一个外部的存储系统。

图 1-94 Structured Streaming 数据处理模型



Structured Streaming在OutPut阶段可以定义不同的存储方式，有如下3种：

- Complete Mode：整个更新的结果集都会写入外部存储。整张表的写入操作将由外部存储系统的连接器完成。
- Append Mode：当时间间隔触发时，只有在Result Table中新增加的数据行会被写入外部存储。这种方式只适用于结果集中已经存在的内容不希望发生改变的情况下，如果已经存在的数据会被更新，不适合适用此种方式。
- Update Mode：当时间间隔触发时，只有在Result Table中被更新的数据才会被写入外部存储系统。注意，和Complete Mode方式的不同之处是不更新的结果集不会写入外部存储。

Spark 常见基本概念

- **RDD**

  即弹性分布数据集（Resilient Distributed Dataset），是Spark的核心概念。指的是一个只读的，可分区的分布式数据集，这个数据集的全部或部分可以缓存在内存中，在多次计算间重用。

  RDD的生成：

  - 从HDFS输入创建，或从与Hadoop兼容的其他存储系统中输入创建。
  - 从父RDD转换得到新RDD。
  - 从数据集合转换而来，通过编码实现。

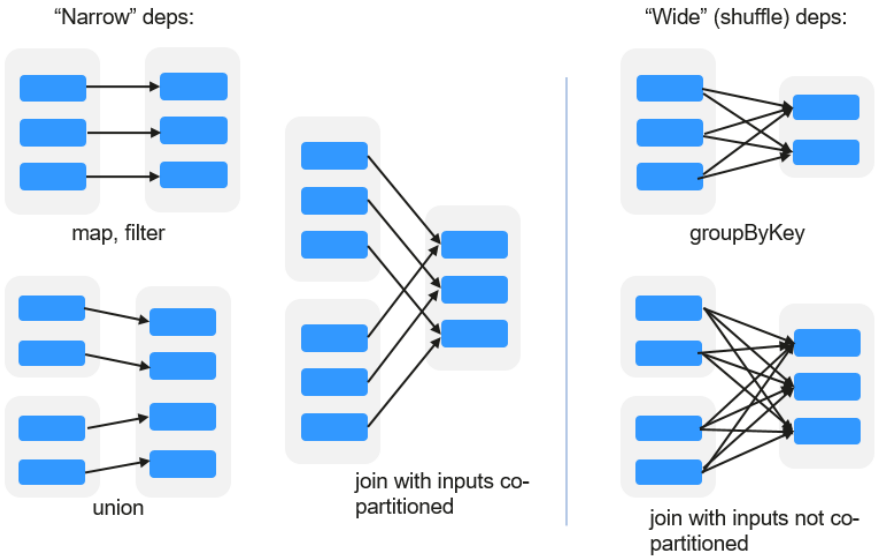
  RDD的存储：

  - 用户可以选择不同的存储级别缓存RDD以便重用（RDD有11种存储级别）。
  - 当前RDD默认是存储于内存，但当内存不足时，RDD会溢出到磁盘中。

- **Dependency（RDD的依赖）**

  RDD的依赖分别为：窄依赖和宽依赖。

图 1-95 RDD 的依赖



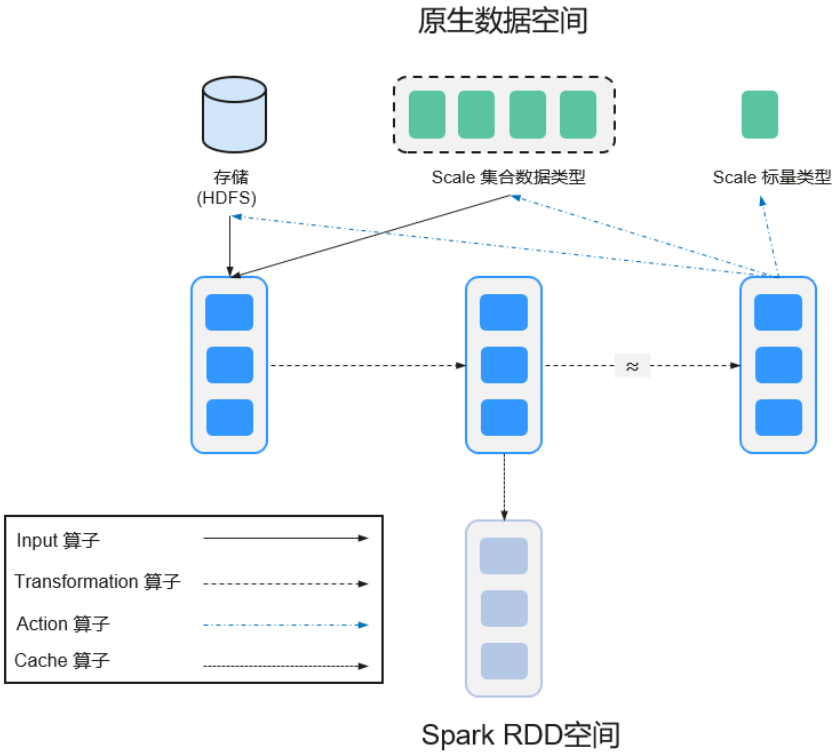
- **窄依赖：**指父RDD的每一个分区最多被一个子RDD的分区所用。
- **宽依赖：**指子RDD的分区依赖于父RDD的所有分区。

窄依赖对优化很有利。逻辑上，每个RDD的算子都是一个fork/join（此join非上文的join算子，而是指同步多个并行任务的barrier）：把计算fork到每个分区，算完后join，然后fork/join下一个RDD的算子。如果直接翻译到物理实现，是很不经济的：一是每一个RDD（即使是中间结果）都需要物化到内存或存储中，费时费空间；二是join作为全局的barrier，是很昂贵的，会被最慢的那个节点拖死。如果子RDD的分区到父RDD的分区是窄依赖，就可以实施经典的fusion优化，把两个fork/join合为一个；如果连续的变换算子序列都是窄依赖，就可以把很多个fork/join并为一个，不但减少了大量的全局barrier，而且无需物化很多中间结果RDD，这将极大地提升性能。Spark把这个叫做流水线（pipeline）优化。

- **Transformation和Action（RDD的操作）**

  对RDD的操作包含Transformation（返回值还是一个RDD）和Action（返回值不是一个RDD）两种。RDD的操作流程如图1-96所示。其中Transformation操作是Lazy的，也就是说从一个RDD转换生成另一个RDD的操作不是马上执行，Spark在遇到Transformations操作时只会记录需要这样的操作，并不会去执行，需要等到有Actions操作的时候才会真正启动计算过程进行计算。Actions操作会返回结果或把RDD数据写到存储系统中。Actions是触发Spark启动计算的动因。

图 1-96 RDD 操作示例



RDD看起来与Scala集合类型没有太大差别，但数据和运行模型大相迥异。

```
val file = sc.textFile("hdfs://...")
val errors = file.filter(_.contains("ERROR"))
errors.cache()
errors.count()
```

a. textFile算子从HDFS读取日志文件，返回file（作为RDD）。

b. filter算子筛出带“ERROR”的行，赋给errors（新RDD）。filter算子是一个Transformation操作。

c. cache算子缓存下来以备未来使用。

d. count算子返回errors的行数。count算子是一个Action操作。

Transformation操作可以分为如下几种类型：

- 视RDD的元素为简单元素。

  输入输出一对一，且结果RDD的分区结构不变，主要是map。

  输入输出一对多，且结果RDD的分区结构不变，如flatMap（map后由一个元素变为一个包含多个元素的序列，然后展平为一个个的元素）。

  输入输出一对一，但结果RDD的分区结构发生了变化，如union（两个RDD合为一个，分区数变为两个RDD分区数之和）、coalesce（分区减少）。

  从输入中选择部分元素的算子，如filter、distinct（去除重复元素）、subtract（本RDD有、其他RDD无的元素留下来）和sample（采样）。

- 视RDD的元素为Key-Value对。

  对单个RDD做一对一运算，如mapValues（保持源RDD的分区方式，这与map不同）；

  对单个RDD重排，如sort、partitionBy（实现一致性的分区划分，这个对数据本地性优化很重要）；

对单个RDD基于key进行重组和reduce，如groupByKey、reduceByKey；

对两个RDD基于key进行join和重组，如join、cogroup。

说明

后三种操作都涉及重排，称为shuffle类操作。

Action操作可以分为如下几种：

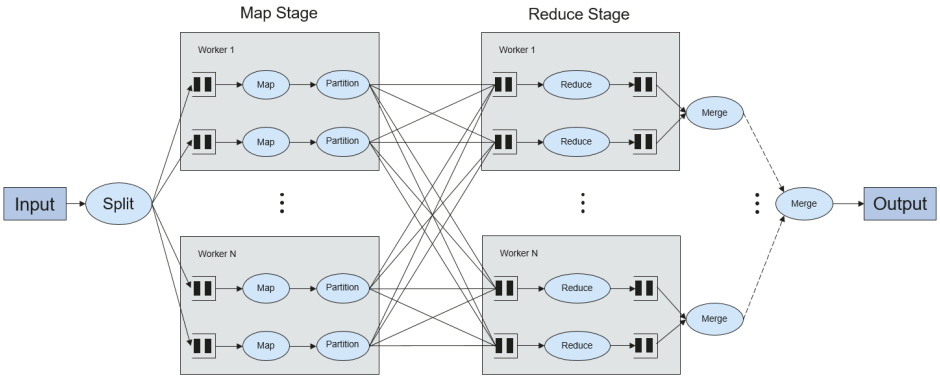
- 生成标量，如count（返回RDD中元素的个数）、reduce、fold/aggregate（返回几个标量）、take（返回前几个元素）。
- 生成Scala集合类型，如collect（把RDD中的所有元素倒入Scala集合类型）、lookup（查找对应key的所有值）。
- 写入存储，如与前文textFile对应的saveAsTextFile。
- 还有一个检查点算子checkpoint。当Lineage特别长时（这在图计算中时常发生），出错时重新执行整个序列要很长时间，可以主动调用checkpoint把当前数据写入稳定存储，作为检查点。

- **Shuffle**

  Shuffle是MapReduce框架中的一个特定的phase，介于Map phase和Reduce phase之间，当Map的输出结果要被Reduce使用时，每一条输出结果需要按key哈希，并且分发到对应的Reducer上去，这个过程就是shuffle。由于shuffle涉及到了磁盘的读写和网络的传输，因此shuffle性能的高低直接影响到了整个程序的运行效率。

  下图清晰地描述了MapReduce算法的整个流程。

  图 1-97 算法流程



  概念上shuffle就是一个沟通数据连接的桥梁，实际上shuffle这一部分是如何实现的呢，下面就以Spark为例讲解shuffle在Spark中的实现。

  Shuffle操作将一个Spark的Job分成多个Stage，前面的stages会包括一个或多个ShuffleMapTasks，最后一个stage会包括一个或多个ResultTask。

- **Spark Application的结构**

  Spark Application的结构可分为两部分：初始化SparkContext和主体程序。

  - 初始化SparkContext：构建Spark Application的运行环境。

    构建SparkContext对象，如：

    ```
    new SparkContext(master, appName, [SparkHome], [jars])
    ```

    参数介绍：

    master：连接字符串，连接方式有local、yarn-cluster、yarn-client等。

appName：构建的Application名称。

SparkHome：集群中安装Spark的目录。

jars：应用程序代码和依赖包。

– 主体程序：处理数据

- **Spark shell命令**

  Spark基本shell命令，支持提交Spark应用。命令为：

```
./bin/spark-submit \
  --class <main-class> \
  --master <master-url> \
  ... # other options
  <application-jar> \
  [application-arguments]
```

  参数解释：

  --class：Spark应用的类名。

  --master：Spark用于所连接的master，如yarn-client，yarn-cluster等。

  application-jar：Spark应用的jar包的路径。

  application-arguments：提交Spark应用的所需要的参数（可以为空）。

- **Spark JobHistory Server**

  用于监控正在运行的或者历史的Spark作业在Spark框架各个阶段的细节以及提供日志显示，帮助用户更细粒度地去开发、配置和调优作业。

1.4.23.2 Spark HA 方案介绍

Spark 多主实例 HA 原理与实现方案

基于社区已有的JDBCServer基础上，采用多主实例模式实现了其高可用性方案。集群中支持同时共存多个JDBCServer服务，通过客户端可以随机连接其中的任意一个服务进行业务操作。即使集群中一个或多个JDBCServer服务停止工作，也不影响用户通过同一个客户端接口连接其他正常的JDBCServer服务。

多主实例模式相比主备模式的HA方案，优势主要体现在对以下两种场景的改进。

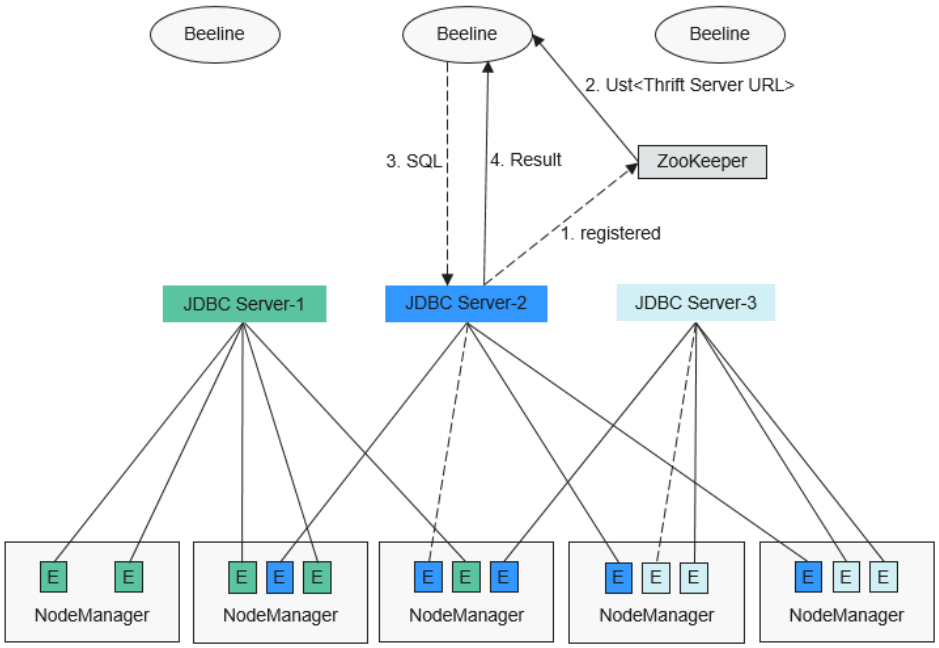
- 主备模式下，当发生主备切换时，会存在一段时间内服务不可用，该时间JDBCServer无法控制，取决于Yarn服务的资源情况。
- Spark中通过类似于HiveServer2的Thrift JDBC提供服务，用户通过Beeline以及JDBC接口访问。因此JDBCServer集群的处理能力取决于主Server的单点能力，可扩展性不够。

采用多主实例模式的HA方案，不仅可以规避主备切换服务中断的问题，实现服务不中断或少中断，还可以通过横向扩展集群来提高并发能力。

- **实现方案**

  多主实例模式的HA方案原理如下图所示。

图 1-98 Spark JDBCServer HA



1. JDBCServer在启动时，向ZooKeeper注册自身消息，在指定目录中写入节点，节点包含了该实例对应的IP，端口，版本号和序列号等信息。
2. 客户端连接JDBCServer时，需要指定Namespace，即访问ZooKeeper哪个目录下的JDBCServer实例。在连接的时候，会从Namespace下随机选择一个实例连接。
3. 客户端成功连接JDBCServer服务后，向JDBCServer服务发送SQL语句。
4. JDBCServer服务执行客户端发送的SQL语句后，将结果返回给客户端。

在HA方案中，每个JDBCServer实例都是独立且等同的，当其中一个实例在升级或者业务中断时，其他的实例也能接受客户端的连接请求。

多主实例方案遵循以下规则：

- 当一个实例异常退出时，其他实例不会接管此实例上的会话，也不会接管此实例上运行的业务。
- 当JDBCServer进程停止时，删除在ZooKeeper上的相应节点。
- 由于客户端选择服务端的策略是随机的，可能会出现会话随机分配不均匀的情况，进而可能引起实例间的负载不均衡。
- 实例进入维护模式（即进入此模式后不再接受新的客户端连接）后，当达到退服超时时间，仍在此实例上运行的业务有可能会发生失败。
- **URL连接介绍**
  - 多主实例模式

    多主实例模式的客户端读取ZooKeeper节点中的内容，连接对应的JDBCServer服务。连接字符串为：

    - 安全模式下：

      Kinit认证方式下的JDBCURL如下所示：

      ```
      jdbc:hive2://
      <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.<系统域名>@<系统域名>;
      ```

说明

- 其中“<zkNode_IP>:<zkNode_Port>”是ZooKeeper的URL，多个URL以逗号隔开。

  例如：
  “192.168.81.37:2181,192.168.195.232:2181,192.168.169.84:2181”。
- 其中“sparkthriftserver2x”是ZooKeeper上的目录，表示客户端从该目录下随机选择JDBCServer实例进行连接。

示例：安全模式下通过Beeline客户端连接时执行以下命令：

sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://** *<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/**;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.*<系统域名>*@*<系统域名>*;"**

Keytab认证方式下的JDBCURL如下所示：

```
jdbc:hive2://
<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.<系统域名>@<系统域名>;user.principal=<principal_name>;user.keytab=<path_to_keytab>
```

其中<principal_name>表示用户使用的Kerberos用户的principal，如“test@*<系统域名>*”。<path_to_keytab>表示<principal_name>对应的keytab文件路径，如“/opt/auth/test/user.keytab”。

- 普通模式下：

  ```
  jdbc:hive2://
  <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;
  ```

  示例：普通模式下通过Beeline客户端连接时执行以下命令：

  sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://** *<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/**;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;"**

– 非多主实例模式

非多主实例模式的客户端连接的是某个指定JDBCServer节点。该模式的连接字符串相比多主实例模式的去掉关于Zookeeper的参数项“serviceDiscoveryMode”和“zooKeeperNamespace”。

示例：安全模式下通过Beeline客户端连接非多主实例模式时执行以下命令：

sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://** *<server_IP>:<server_Port>*/**;user.principal=spark/hadoop.*<系统域名>*@*<系统域名>*;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.*<系统域名>*@*<系统域名>*;"**

说明

- 其中“<server_IP>:<server_Port>”是指定JDBCServer节点的URL。
- “CLIENT_HOME”是指客户端路径。

多主实例模式与非多主实例模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。

Spark 多租户 HA 方案实现

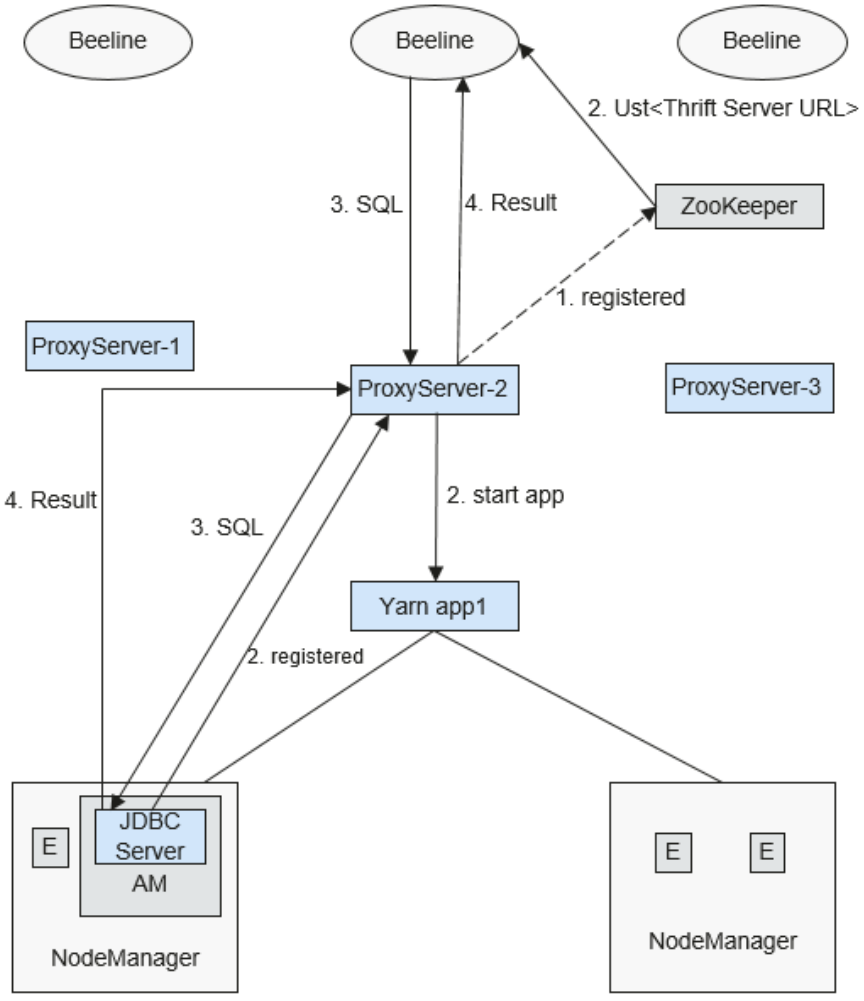
JDBCServer多主实例方案中，JDBCServer实现使用YARN-Client模式，但YARN资源队列只有一个，为了解决这种资源局限的问题，引入了多租户模式。

多租户模式是将JDBCServer和租户绑定，每一个租户对应一个或多个JDBCServer，而一个JDBCServer只给一个租户提供服务。不同的租户可以配置不同的YARN队列，从而达到资源隔离，且JDBCServer根据需求动态启动，可避免浪费资源。

- **实现方案**

  多租户模式的HA方案原理如图1-99所示。

  图 1-99 Spark JDBCServer 多租户



  a. ProxyServer在启动时，向ZooKeeper注册自身消息，在指定目录中写入节点信息，节点信息包含了该实例对应的IP，端口，版本号和序列号等信息。

  说明

  多租户模式下，JDBCServer实例是指ProxyServer（JDBCServer代理）。

  b. 客户端连接ProxyServer时，需要指定Namespace，即访问ZooKeeper哪个目录下的ProxyServer实例。在连接的时候，会从Namespace下随机选择一个实例连接，详细URL参见**URL连接介绍**。

c. 客户端成功连接ProxyServer服务，ProxyServer服务首先确认是否有该租户的JDBCServer存在，如果有，直接将Beeline连上真正的JDBCServer；如果没有，则以YARN-Cluster模式启动一个新的JDBCServer。JDBCServer启动成功后，ProxyServer会获取JDBCServer的地址，并将Beeline连上JDBCServer。

d. 客户端发送SQL语句给ProxyServer，ProxyServer将语句转交给真正连上的JDBCServer处理。最后JDBCServer服务将结果返回给ProxyServer，ProxyServer再将结果返回给客户端。

在HA方案中，每个ProxyServer服务（即实例）都是独立且等同的，当其中一个实例在升级或者业务中断时，其他的实例也能接受客户端的连接请求。

- **URL连接介绍**
  - 多租户模式

    多租户模式的客户端读取ZooKeeper节点中的内容，连接对应的ProxyServer服务。连接字符串为：

    - 安全模式下：

      Kinit认证方式下的客户端URL如下所示：

      ```
      jdbc:hive2://
      <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.<系统域名>@<系统域名>;
      ```

      说明

      - 其中“<zkNode_IP>:<zkNode_Port>”是ZooKeeper的URL，多个URL以逗号隔开。

        例如：
        “192.168.81.37:2181,192.168.195.232:2181,192.168.169.84:2181”。
      - 其中sparkthriftserver2x是ZooKeeper上的目录，表示客户端从该目录下随机选择JDBCServer实例进行连接。

      示例：安全模式下通过Beeline客户端连接时执行以下命令：

      sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://** *<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/**;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.*<系统域名>*@*<系统域名>*;"**

      Keytab认证方式下的URL如下所示：

      ```
      jdbc:hive2://
      <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.<系统域名>@<系统域名>;user.principal=<principal_name>;user.keytab=<path_to_keytab>
      ```

      其中<principal_name>表示用户使用的Kerberos用户的principal，如“test@*<系统域名>*”。<path_to_keytab>表示<principal_name>对应的keytab文件路径，如“/opt/auth/test/user.keytab”。

    - 普通模式下：

      ```
      jdbc:hive2://
      <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;
      ```

      示例：普通模式下通过Beeline客户端连接时执行以下命令：

      sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://** *<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkN*

ode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;"

- 非多租户模式

  非多租户模式的客户端连接的是某个指定JDBCServer节点。该模式的连接字符串相比多主实例模式的去掉关于ZooKeeper的参数项“serviceDiscoveryMode”和“zooKeeperNamespace”。

  示例：安全模式下通过Beeline客户端连接非多租户模式时执行以下命令：

  sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<server_IP>:<server_Port>*/;user.principal=spark/hadoop.*<系统域名>*@*<系统域名>*;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.*<系统域名>*@*<系统域名>*;"

  说明

  - 其中“<server_IP>:<server_Port>”是指定JDBCServer节点的URL。
  - “CLIENT_HOME”是指客户端路径。

  多租户模式与非多租户模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。

  指定租户

  一般情况下，某用户提交的客户端会连接到该用户默认所属租户的JDBCServer上，若需要连接客户端到指定租户的JDBCServer上，可以通过添加**--hiveconf mapreduce.job.queuename**进行指定。

  通过Beeline连接的命令示例如下（aaa为租户名称）：

  beeline --hiveconf mapreduce.job.queuename=aaa -u 'jdbc:hive2://192.168.39.30:2181,192.168.40.210:2181,192.168.215.97:2181;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark/hadoop.*<系统域名>*@*<系统域名>*;'

1.4.23.3 Spark 与 HDFS 和 YARN 的关系

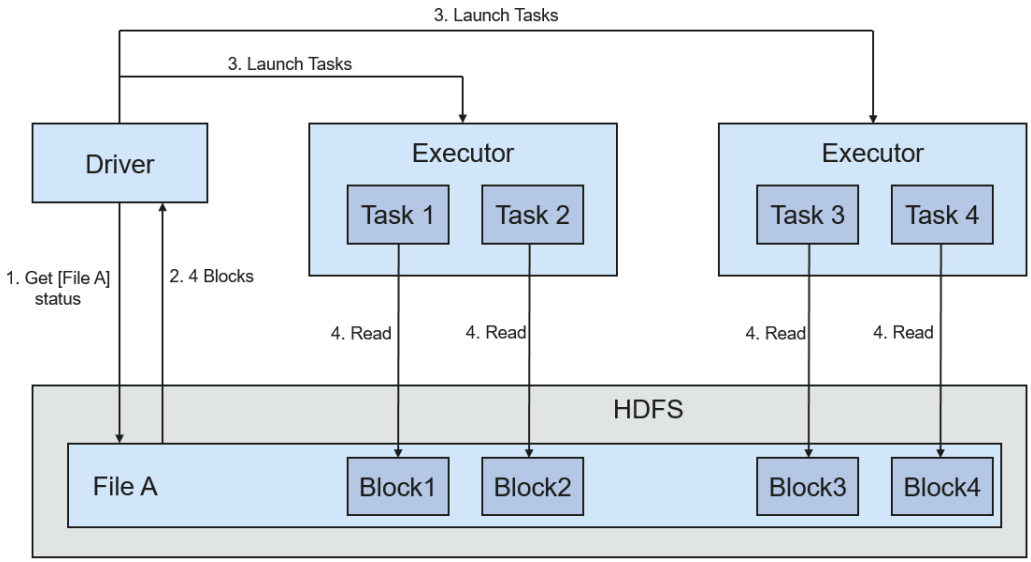
Spark 和 HDFS 的关系

通常，Spark中计算的数据可以来自多个数据源，如Local File、HDFS等。最常用的是HDFS，用户可以一次读取大规模的数据进行并行计算。在计算完成后，也可以将数据存储到HDFS。

分解来看，Spark分成控制端(Driver)和执行端（Executor）。控制端负责任务调度，执行端负责任务执行。

读取文件的过程如图 读取文件过程所示。

图 1-100 读取文件过程

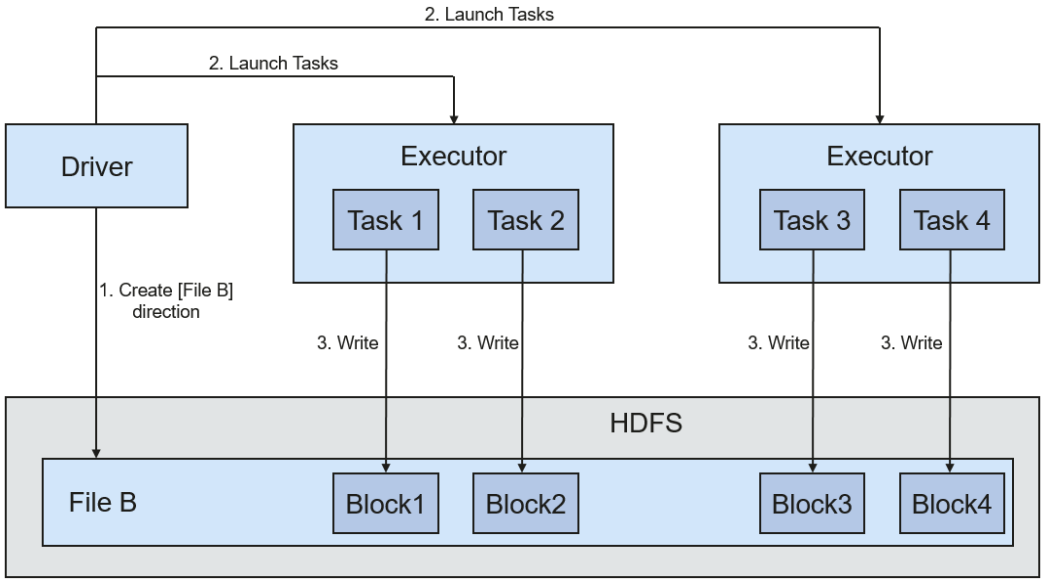


读取文件步骤的详细描述如下所示：

1. Driver与HDFS交互获取File A的文件信息。
2. HDFS返回该文件具体的Block信息。
3. Driver根据具体的Block数据量，决定一个并行度，创建多个Task去读取这些文件Block。
4. 在Executor端执行Task并读取具体的Block，作为RDD(弹性分布数据集)的一部分。

写入文件的过程如图 写入文件过程所示。

图 1-101 写入文件过程



HDFS文件写入的详细步骤如下所示：

1. Driver创建要写入文件的目录。
2. 根据RDD分区分块情况，计算出写数据的Task数，并下发这些任务到Executor。
3. Executor执行这些Task，将具体RDD的数据写入到步骤1创建的目录下。

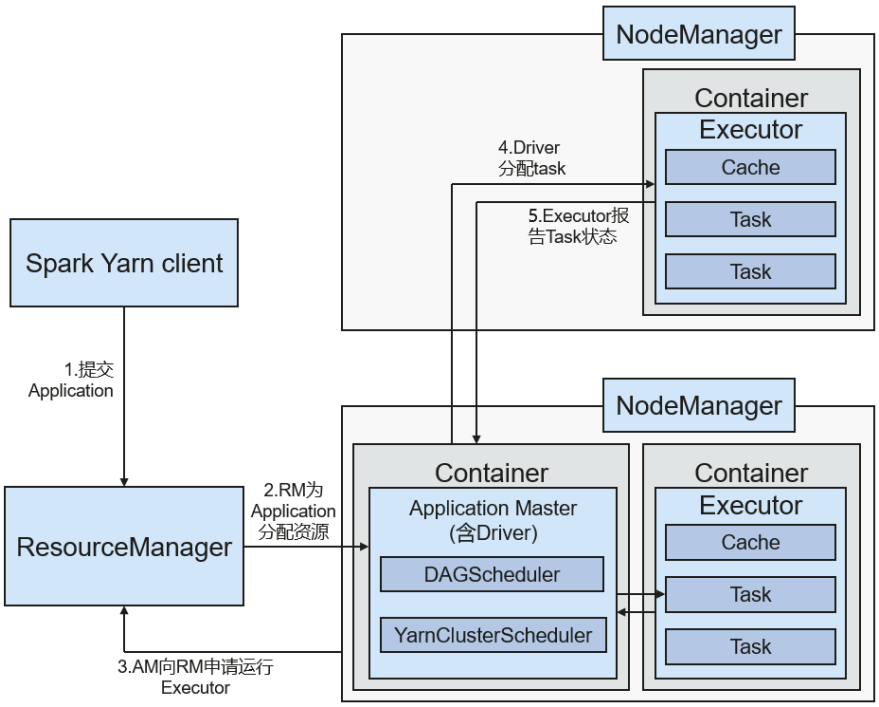
Spark 和 YARN 的关系

Spark的计算调度方式，可以通过YARN的模式实现。Spark共享YARN集群提供丰富的计算资源，将任务分布式的运行起来。Spark on YARN分两种模式：YARN Cluster和YARN Client。

- YARN Cluster模式

  运行框架如图 Spark on yarn-cluster运行框架所示。

  图 1-102 Spark on yarn-cluster 运行框架



  Spark on yarn-cluster实现流程：

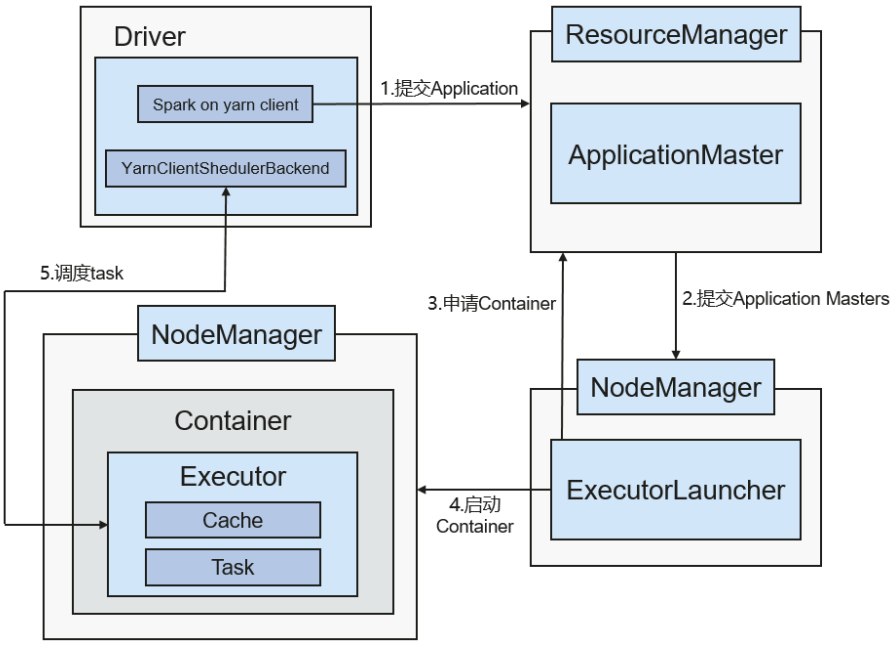
  a. 首先由客户端生成Application信息，提交给ResourceManager。
  b. ResourceManager为Spark Application分配第一个Container(ApplicationMaster)，并在该Container上启动Driver。
  c. ApplicationMaster向ResourceManager申请资源以运行Container。

     ResourceManager分配Container给ApplicationMaster，ApplicationMaster和相关的NodeManager通讯，在获得的Container上启动Executor，Executor启动后，开始向Driver注册并申请Task。
  d. Driver分配Task给Executor执行。
  e. Executor执行Task并向Driver汇报运行状况。

- YARN Client模式

  运行框架如图 Spark on yarn-client运行框架所示。

图 1-103 Spark on yarn-client 运行框架



Spark on yarn-client实现流程：

说明

在yarn-client模式下，Driver部署在Client端，在Client端启动。yarn-client模式下，不兼容老版本的客户端。推荐使用yarn-cluster模式。

a. 客户端向ResourceManager发送Spark应用提交请求，ResourceManager为其返回应答，该应答中包含多种信息(如ApplicationId、可用资源使用上限和下限等)。Client端将启动ApplicationMaster所需的所有信息打包，提交给ResourceManager上。

b. ResourceManager收到请求后，会为ApplicationMaster寻找合适的节点，并在该节点上启动它。ApplicationMaster是Yarn中的角色，在Spark中进程名字是ExecutorLauncher。

c. 根据每个任务的资源需求，ApplicationMaster可向ResourceManager申请一系列用于运行任务的Container。

d. 当ApplicationMaster（从ResourceManager端）收到新分配的Container列表后，会向对应的NodeManager发送信息以启动Container。

   ResourceManager分配Container给ApplicationMaster，ApplicationMaster和相关的NodeManager通讯，在获得的Container上启动Executor，Executor启动后，开始向Driver注册并申请Task。

   说明

   正在运行的container不会被挂起释放资源。

e. Driver分配Task给Executor执行。Executor执行Task并向Driver汇报运行状况。

1.4.23.4 Spark 开源增强特性：跨源复杂数据的 SQL 查询优化

场景描述

出于管理和信息收集的需要，企业内部会存储海量数据，包括数目众多的各种数据库、数据仓库等，此时会面临以下困境：数据源种类繁多，数据集结构化混合，相关数据存放分散等，这就导致了跨源复杂查询因传输效率低，耗时长。

当前开源Spark在跨源查询时，只能对简单的filter进行下推，因此造成大量不必要的数据传输，影响SQL引擎性能。针对下推能力进行增强，当前对aggregate、复杂projection、复杂predicate均可以下推到数据源，尽量减少不必要数据的传输，提升查询性能。

目前仅支持JDBC数据源的查询下推，支持的下推模块有aggregate、projection、predicate、aggregate over inner join、aggregate over union all等。为应对不同应用场景的特殊需求，对所有下推模块设计开关功能，用户可以自行配置是否应用上述查询下推的增强。

表 1-20 跨源查询增加特性对比

模块	增强前	增强后
aggregate	不支持aggregate下推	• 支持的聚合函数为：sum, avg, max, min, count 例如：select count(*) from table • 支持聚合函数内部表达式 例如：select sum(a+b) from table • 支持聚合函数运算，例如：select avg(a) + max(b) from table • 支持having下推 例如：select sum(a) from table where a>0 group by b having sum(a)>10 • 支持部分函数下推 支持对abs()、month()、length()等数学、时间、字符串函数进行下推。并且，除了以上内置函数，用户还可以通过SET命令新增数据源支持的函数。 例如：select sum(abs(a)) from table • 支持aggregate之后的limit、order by下推（由于Oracle不支持limit，所以Oracle中limit、order by不会下推） 例如：select sum(a) from table where a>0 group by b order by sum(a) limit 5

模块	增强前	增强后
projection	仅支持简单projection下推，例如：select a, b from table	• 支持复杂表达式下推。 例如：select (a+b)*c from table • 支持部分函数下推，详细参见表下方的说明。 例如：select length(a)+abs(b) from table • 支持projection之后的limit、order by下推。 例如：select a, b+c from table order by a limit 3
predicate	仅支持运算符左边为列名右边为值的简单filter，例如 select * from table where a>0 or b in (“aaa”, “bbb”)	• 支持复杂表达数下推 例如：select * from table where a +b>c*d or a/c in (1, 2, 3) • 支持部分函数下推，详细参见表下方的说明。 例如：select * from table where length(a)>5
aggregate over inner join	需要将两个表中相关的数据全部加载到Spark，先进行join操作，再进行aggregate操作	支持以下几种： • 支持的聚合函数为：sum, avg, max, min,count • 所有aggregate只能来自同一个表，group by可以来自一个表或者两个表，只支持inner join。 不支持的情形有： • 不支持aggregate同时来自join左表和右表的下推。 • 不支持aggregate内包含运算，如：sum(a+b)。 • 不支持aggregate运算，如：sum(a) +min(b)。
aggregate over union all	需要将两个表中相关的数据全部加载到Spark，先进行union操作，再进行aggregate操作	支持情况： 支持的聚合函数为：sum, avg, max, min,count 不支持的情况： • 不支持aggregate内包含运算，如：sum(a+b)。 • 不支持aggregate运算，如：sum(a) +min(b)。

注意事项

- 外部数据源是Hive的场景，通过Spark建的外表无法进行查询。
- 数据源只支持MySQL和Mppdb。

1.4.24 Spark2x

1.4.24.1 Spark2x 基本原理

说明

Spark2x组件适用于MRS 3.x及后续版本。

简介

Spark是基于内存的分布式计算框架。在迭代计算的场景下，数据处理过程中的数据可以存储在内存中，提供了比MapReduce高10到100倍的计算能力。Spark可以使用HDFS作为底层存储，使用户能够快速地从MapReduce切换到Spark计算平台上去。Spark提供一站式数据分析能力，包括小批量流式处理、离线批处理、SQL查询、数据挖掘等，用户可以在同一个应用中无缝结合使用这些能力。Spark2x的开源新特性请参考Spark2x开源新特性。

Spark的特点如下：

- 通过分布式内存计算和DAG（无回路有向图）执行引擎提升数据处理能力，比MapReduce性能高10倍到100倍。
- 提供多种语言开发接口（Scala/Java/Python），并且提供几十种高度抽象算子，可以很方便构建分布式的数据处理应用。
- 结合SQL、Streaming等形成数据处理栈，提供一站式数据处理能力。
- 完美契合Hadoop生态环境，Spark应用可以运行在Standalone、Mesos或者YARN上，能够接入HDFS、HBase、Hive等多种数据源，支持MapReduce程序平滑转接。

结构

Spark的架构如图1-104所示，各模块的说明如表1-21所示。

图 1-104 Spark 架构

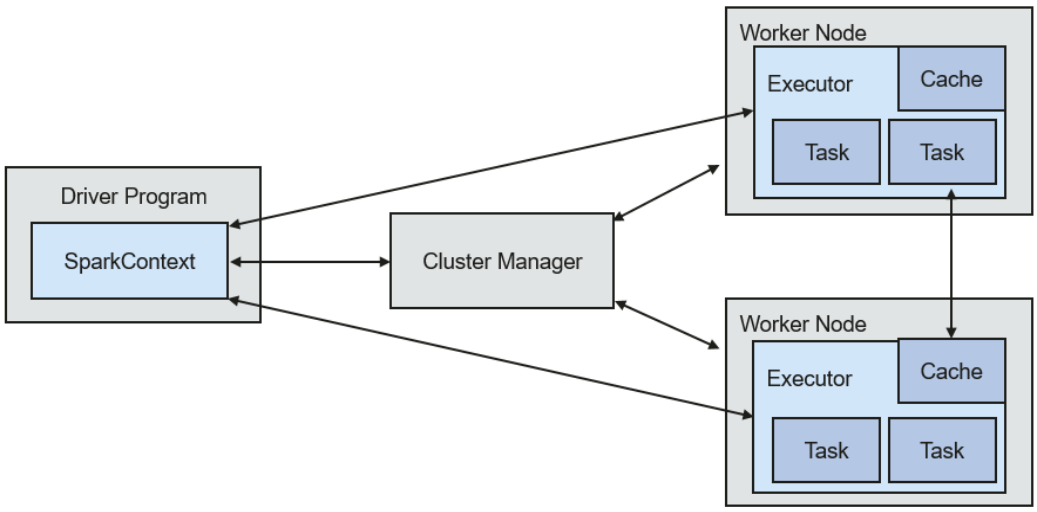


表 1-21 基本概念说明

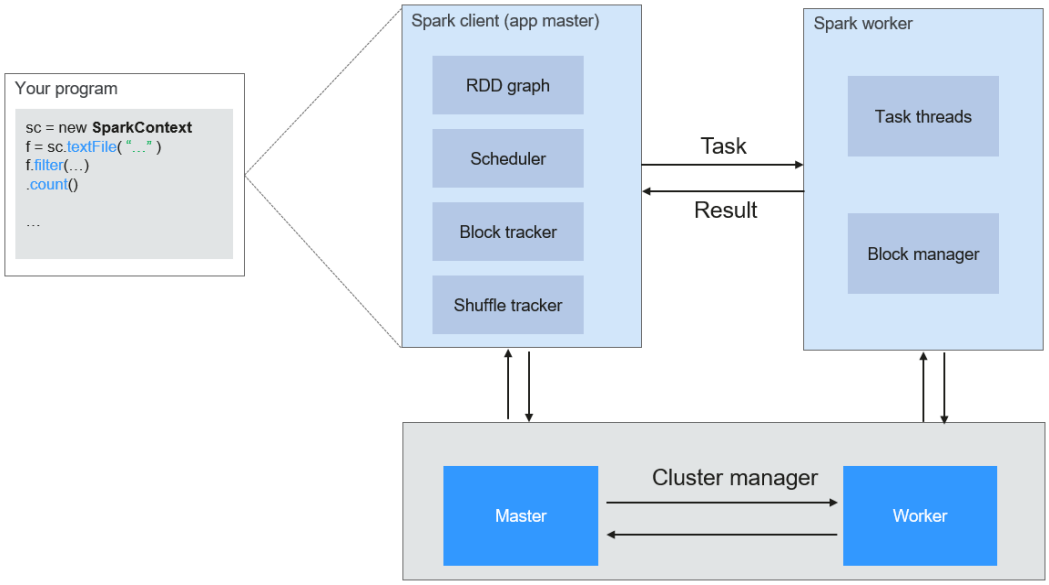
模块	说明
Cluster Manager	集群管理器，管理集群中的资源。Spark支持多种集群管理器，Spark自带的Standalone集群管理器、Mesos或YARN。Spark集群默认采用YARN模式。
Application	Spark应用，由一个Driver Program和多个Executor组成。
Deploy Mode	部署模式，分为cluster和client模式。cluster模式下，Driver会在集群内的节点运行；而在client模式下，Driver在客户端运行（集群外）。
Driver Program	是Spark应用程序的主进程，运行Application的main()函数并创建SparkContext。负责应用程序的解析、生成Stage并调度Task到Executor上。通常SparkContext代表Driver Program。
Executor	在Work Node上启动的进程，用来执行Task，管理并处理应用中使用到的数据。一个Spark应用一般包含多个Executor，每个Executor接收Driver的命令，并执行一到多个Task。
Worker Node	集群中负责启动并管理Executor以及资源的节点。
Job	一个Action算子（比如collect算子）对应一个Job，由并行计算的多个Task组成。
Stage	每个Job由多个Stage组成，每个Stage是一个Task集合，由DAG分割而成。
Task	承载业务逻辑的运算单元，是Spark平台上可执行的最小工作单元。一个应用根据执行计划以及计算量分为多个Task。

Spark 原理

Spark的应用运行架构如**图1-105**所示，运行流程如下所示：

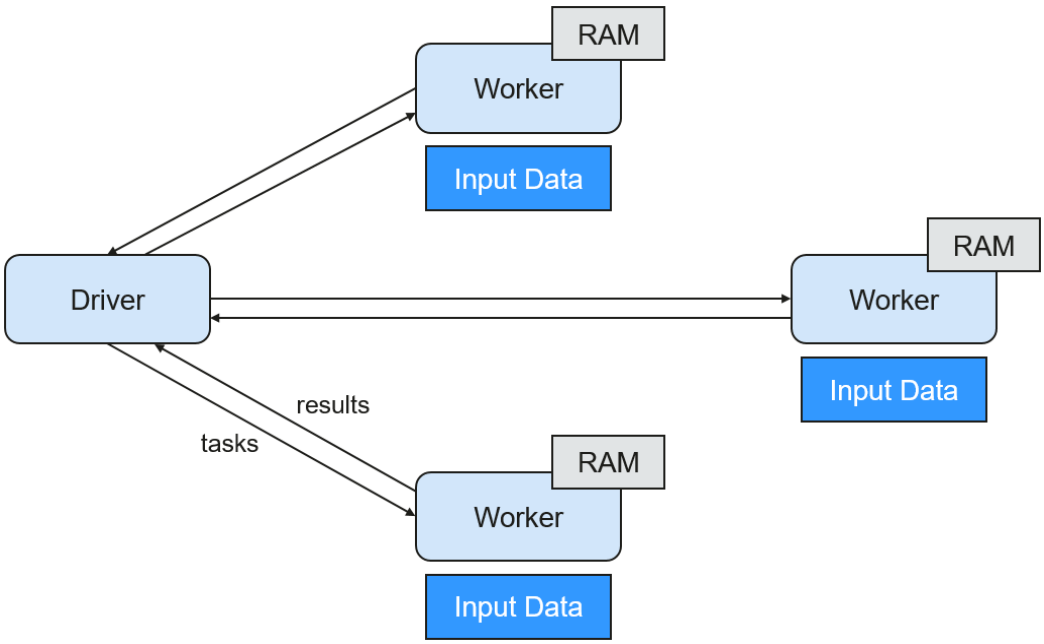
1. 应用程序（Application）是作为一个进程的集合运行在集群上的，由Driver进行协调。
2. 在运行一个应用时，Driver会去连接集群管理器（Standalone、Mesos、YARN）申请运行Executor资源，并启动ExecutorBackend。然后由集群管理器在不同的应用之间调度资源。Driver同时会启动应用程序DAG调度、Stage划分、Task生成。
3. 然后Spark会把应用的代码（传递给**SparkContext**的JAR或者Python定义的代码）发送到Executor上。
4. 所有的Task执行完成后，用户的应用程序运行结束。

图 1-105 Spark 应用运行架构



Spark采用Master和worker的模式，如图1-106所示。用户在Spark客户端提交应用程序，调度器将Job分解为多个Task发送到各个Worker中执行，各个Worker将计算的结果上报给Driver（即Master），Driver聚合结果返回给客户端。

图 1-106 Spark 的 Master 和 Worker



在此结构中，有几个说明点：

- 应用之间是独立的。

  每个应用有自己的executor进程，Executor启动多个线程，并行地执行任务。无论是在调度方面，或者是executor方面。各个Driver独立调度自己的任务；不同的应用任务运行在不同的JVM上，即不同的Executor。

- 不同Spark应用之间是不共享数据的，除非把数据存储在外部的存储系统上（比如HDFS）。
- 因为Driver程序在集群上调度任务，所以Driver程序最好和worker节点比较近，比如在一个相同的局部网络内。

Spark on YARN有两种部署模式：

- YARN-Cluster模式下，Spark的Driver会运行在YARN集群内的ApplicationMaster进程中，ApplicationMaster已经启动之后，提交任务的客户端退出也不会影响任务的运行。
- YRAN-Client模式下，Driver启动在客户端进程内，ApplicationMaster进程只用来向YARN集群申请资源。

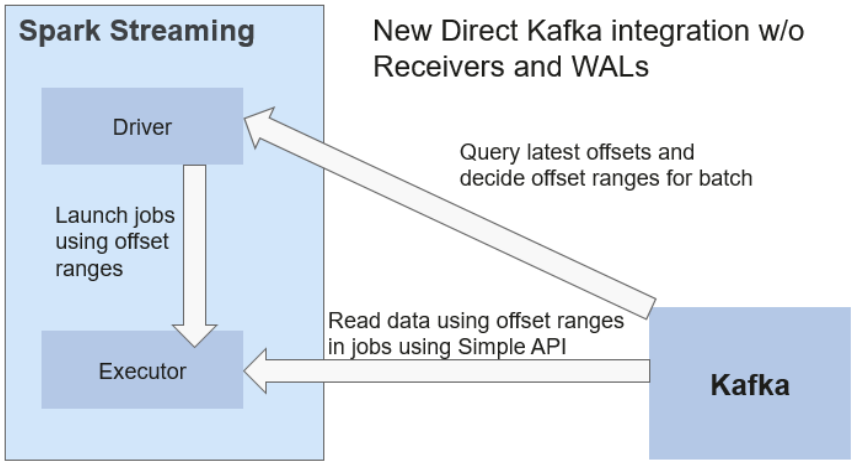
Spark Streaming 原理

Spark Streaming是一种构建在Spark上的实时计算框架，扩展了Spark处理大规模流式数据的能力。当前Spark支持两种数据处理方式：Direct Streaming和Receiver方式。

Direct Streaming计算流程

Direct Streaming方式主要通过采用Direct API对数据进行处理。以Kafka Direct接口为例，与启动一个Receiver来连续不断地从Kafka中接收数据并写入到WAL中相比，Direct API简单地给出每个batch区间需要读取的偏移量位置。然后，每个batch的Job被运行，而对应偏移量的数据在Kafka中已准备好。这些偏移量信息也被可靠地存储在checkpoint文件中，应用失败重启时可以直接读取偏移量信息。

图 1-107 Direct Kafka 接口数据传输



需要注意的是，Spark Streaming可以在失败后重新从Kafka中读取并处理数据段。然而，由于语义仅被处理一次，重新处理的结果和没有失败处理的结果是一致的。

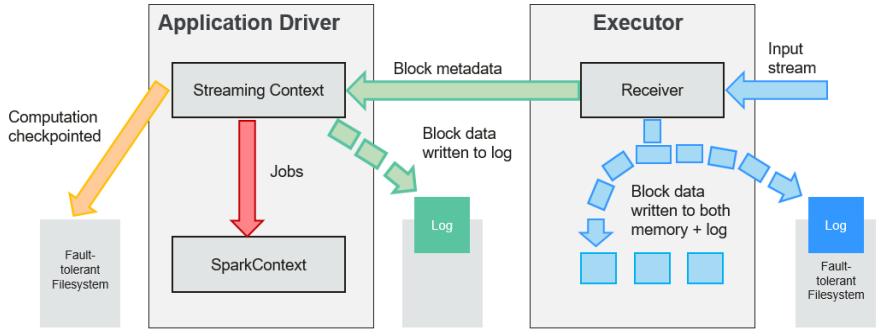
因此，Direct API消除了需要使用WAL和Receivers的情况，且确保每个Kafka记录仅被接收一次，这种接收更加高效。使得Spark Streaming和Kafka可以很好地整合在一起。总体来说，这些特性使得流处理管道拥有高容错性、高效性及易用性，因此推荐使用Direct Streaming方式处理数据。

Receiver计算流程

在一个Spark Streaming应用开始时（也就是Driver开始时），相关的StreamingContext（所有流功能的基础）使用SparkContext启动Receiver成为长驻运

行任务。这些Receiver接收并保存流数据到Spark内存中以供处理。用户传送数据的生命周期如图1-108所示：

图 1-108 数据传输生命周期



1. 接收数据（蓝色箭头）

   Receiver将数据流分成一系列小块，存储到Executor内存中。另外，在启用预写日志（Write-ahead Log，简称WAL）以后，数据同时还写入到容错文件系统的预写日志中。

2. 通知Driver（绿色箭头）

   接收块中的元数据（Metadata）被发送到Driver的StreamingContext。这个元数据包括：

   - 定位其在Executor内存中数据位置的块Reference ID。
   - 若启用了WAL，还包括块数据在日志中的偏移信息。

3. 处理数据（红色箭头）

   对每个批次的数据，StreamingContext使用Block信息产生RDD及其Job。StreamingContext通过运行任务处理Executor内存中的Block来执行Job。

4. 周期性地设置检查点（橙色箭头）

5. 为了容错的需要，StreamingContext会周期性地设置检查点，并保存到外部文件系统中。

容错性

Spark及其RDD允许无缝地处理集群中任何Worker节点的故障。鉴于Spark Streaming建立于Spark之上，因此其Worker节点也具备了同样的容错能力。然而，由于Spark Streaming的长正常运行需求，其应用程序必须也具备从Driver进程（协调各个Worker的主要应用进程）故障中恢复的能力。使Spark Driver能够容错是件很棘手的事情，因为可能是任意计算模式实现的任意用户程序。不过Spark Streaming应用程序在计算上有一个内在的结构：在每批次数据周期性地执行同样的Spark计算。这种结构允许把应用的状态（亦称Checkpoint）周期性地保存到可靠的存储空间中，并在Driver重新启动时恢复该状态。

对于文件这样的源数据，这个Driver恢复机制足以做到零数据丢失，因为所有的数据都保存在了像HDFS这样的容错文件系统中。但对于像Kafka和Flume等其他数据源，有些接收到的数据还只缓存在内存中，尚未被处理，就有可能会丢失。这是由于Spark应用的分布操作方式引起的。当Driver进程失败时，所有在Cluster Manager中运行的Executor，连同在内存中的所有数据，也同时被终止。为了避免这种数据损失，Spark Streaming引进了WAL功能。

WAL通常被用于数据库和文件系统中，用来保证任何数据操作的持久性，即先将操作记入一个持久的日志，再对数据施加这个操作。若施加操作的过程中执行失败了，则

通过读取日志并重新施加前面预定的操作，系统就得到了恢复。下面介绍了如何利用这样的概念保证接收到的数据的持久性。

Kafka数据源使用Receiver来接收数据，是Executor中的长运行任务，负责从数据源接收数据，并且在数据源支持时还负责确认收到数据的结果（收到的数据被保存在Executor的内存中，然后Driver在Executor中运行来处理任务）。

当启用了预写日志以后，所有收到的数据同时还保存到了容错文件系统的日志文件中。此时即使Spark Streaming失败，这些接收到的数据也不会丢失。另外，接收数据的正确性只在数据被预写到日志以后Receiver才会确认，已经缓存但还没有保存的数据可以在Driver重新启动之后由数据源再发送一次。这两个机制确保了零数据丢失，即所有的数据或者从日志中恢复，或者由数据源重发。

如果需要启用预写日志功能，可以通过如下动作实现：

- 通过“streamingContext.checkpoint”(path-to-directory)设置checkpoint的目录，这个目录是一个HDFS的文件路径，既用作保存流的checkpoint，又用作保存预写日志。
- 设置SparkConf的属性“spark.streaming.receiver.writeAheadLog.enable”为“true”（默认值是“false”）。

在WAL被启用以后，所有Receiver都获得了能够从可靠收到的数据中恢复的优势。建议缓存RDD时不采取多备份选项，因为用于预写日志的容错文件系统很可能也复制了数据。

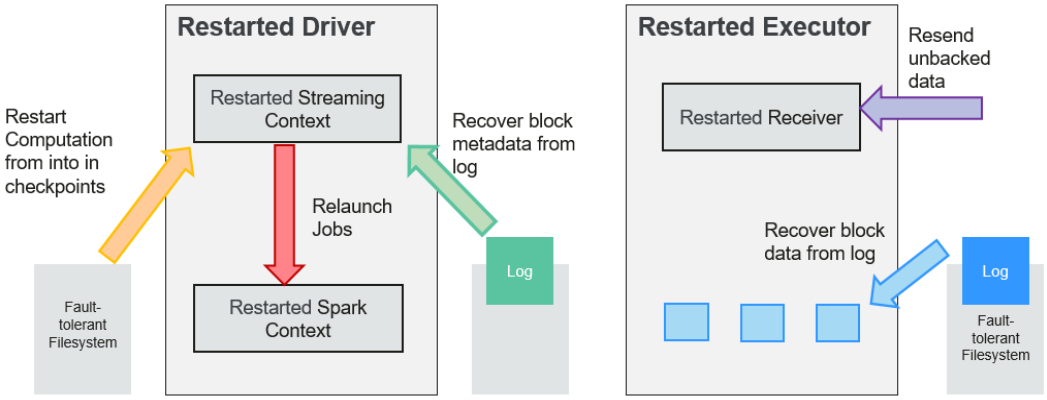
说明

在启用了预写日志以后，数据接收吞吐率会有降低。由于所有数据都被写入容错文件系统，文件系统的写入吞吐率和用于数据复制的网络带宽，可能就是潜在的瓶颈了。在此情况下，最好创建更多的Recevier增加数据接收的并行度，或使用更好的硬件以增加容错文件系统的吞吐率。

恢复流程

当一个失败的Driver重启时，按如下流程启动：

图 1-109 计算恢复流程



1. 恢复计算（橙色箭头）
   使用checkpoint信息重启Driver，重新构造SparkContext并重启Receiver。
2. 恢复元数据块（绿色箭头）
   为了保证能够继续下去所必备的全部元数据块都被恢复。

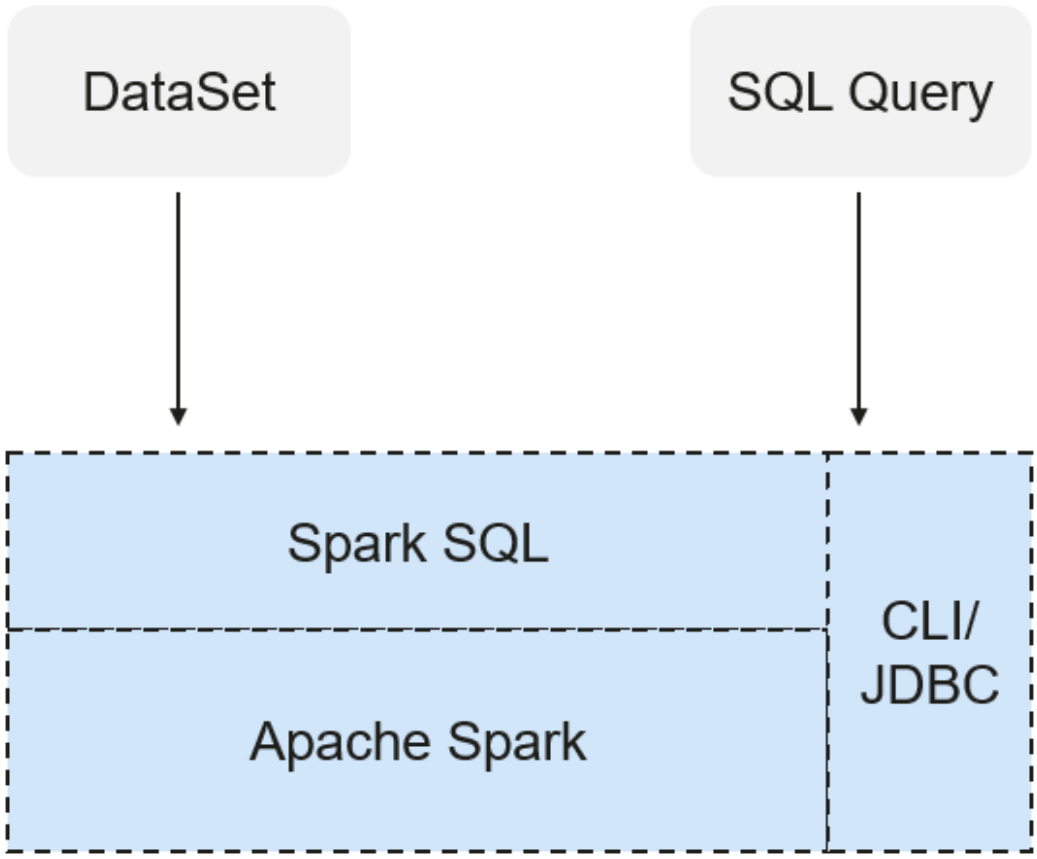
3. 未完成作业的重新形成（红色箭头）
   由于失败而没有处理完成的批处理，将使用恢复的元数据再次产生RDD和对应的作业。
4. 读取保存在日志中的块数据（蓝色箭头）
   在这些作业执行时，块数据直接从预写日志中读出。这将恢复在日志中可靠地保存的所有必要数据。
5. 重发尚未确认的数据（紫色箭头）
   失败时没有保存到日志中的缓存数据将由数据源再次发送。因为Receiver尚未对其确认。

因此通过预写日志和可靠的Receiver，Spark Streaming就可以保证没有输入数据会由于Driver的失败而丢失。

SparkSQL 和 DataSet 原理

SparkSQL

图 1-110 SparkSQL 和 DataSet



Spark SQL是Spark中用于结构化数据处理的模块。在Spark应用中，可以无缝的使用SQL语句亦或是DataSet API对结构化数据进行查询。

Spark SQL以及DataSet还提供了一种通用的访问多数据源的方式，可访问的数据源包括Hive、CSV、Parquet、ORC、JSON和JDBC数据源，这些不同的数据源之间也可以实现互相操作。Spark SQL复用了Hive的前端处理逻辑和元数据处理模块，使用Spark SQL可以直接对已有的Hive数据进行查询。

另外，SparkSQL还提供了诸如API、CLI、JDBC等诸多接口，对客户端提供多样接入形式。

Spark SQL Native DDL/DML

Spark1.5将很多DDL/DML命令下压到Hive执行，造成了与Hive的耦合，且在一定程度上不够灵活（比如报错不符合预期、结果与预期不一致等）。

Spark2x实现了命令的本地化，使用Spark SQL Native DDL/DML取代Hive执行DDL/DML命令。一方面实现和Hive的解耦，另一方面可以对命令进行定制化。

DataSet

DataSet是一个由特定域的对象组成的强类型集合，可通过功能或关系操作并行转换其中的对象。 每个Dataset还有一个非类型视图，即由多个列组成的DataSet，称为DataFrame。

DataFrame是一个由多个列组成的结构化的分布式数据集合，等同于关系数据库中的一张表，或者是R/Python中的data frame。DataFrame是Spark SQL中的最基本的概念，可以通过多种方式创建，例如结构化的数据集、Hive表、外部数据库或者是RDD。

可用于DataSet的操作分为Transformation和Action。

- Transformation操作可生成新的DataSet。
  如map、filter、select和aggregate (groupBy)。
- Action操作可触发计算及返回记结果。
  如count、show或向文件系统写数据。

通常使用两种方法创建一个DataSet：

- 最常见的方法是通过使用SparkSession上的read函数将Spark指向存储系统上的某些文件。
  ```
  val people = spark.read.parquet("...").as[Person]  // Scala
  DataSet<Person> people = spark.read().parquet("...").as(Encoders.bean(Person.class));//Java
  ```
- 还可通过已存在的DataSet上可用的transformation操作来创建数据集。 例如，在已存在的DataSet上应用map操作来创建新的DataSet：
  ```
  val names = people.map(_.name)  // 使用Scala语言，且names为一个Datase
  Dataset<String> names = people.map((Person p) -> p.name, Encoders.STRING)); // Java
  ```

CLI和JDBCServer

除了API编程接口之外，Spark SQL还对外提供CLI/JDBC接口：

- spark-shell和spark-sql脚本均可以提供CLI，以便于调试。
- JDBCServer提供JDBC接口，外部可直接通过发送JDBC请求来完成结构化数据的计算和解析。

SparkSession 原理

SparkSession是Spark2x编程的统一API，也可看作是读取数据的统一入口。SparkSession提供了一个统一的入口点来执行以前分散在多个类中的许多操作，并且还为那些较旧的类提供了访问器方法，以实现最大的兼容性。

使用构建器模式创建SparkSession。如果存在SparkSession，构建器将自动重用现有的SparkSession；如果不存在则会创建一个SparkSession。 在I/O期间，在构建器中设置的配置项将自动同步到Spark和Hadoop。

```
import org.apache.spark.sql.SparkSession
val sparkSession = SparkSession.builder
  .master("local")
  .appName("my-spark-app")
  .config("spark.some.config.option", "config-value")
  .getOrCreate()
```

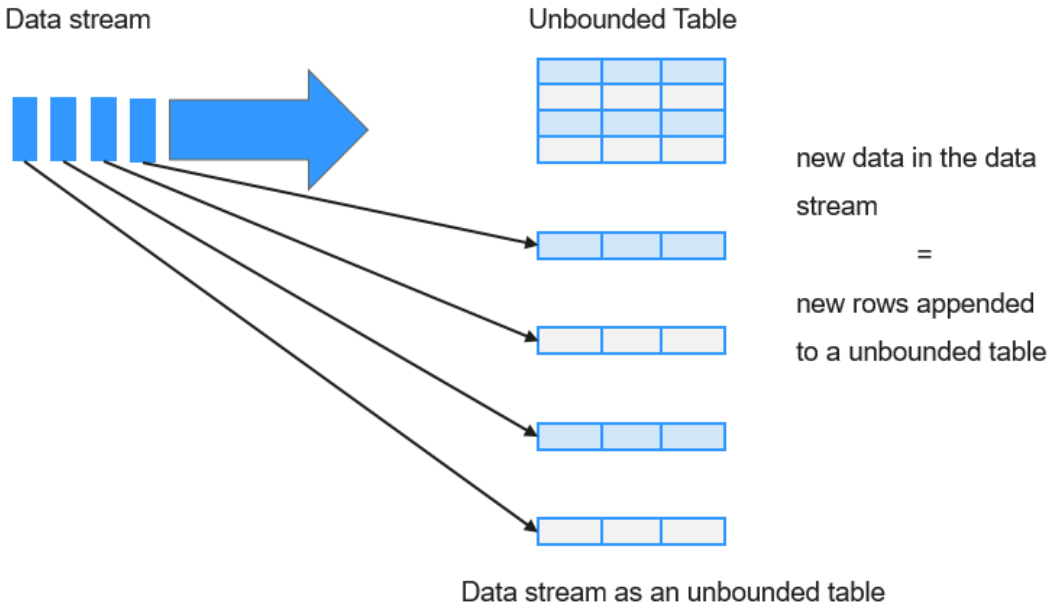
- SparkSession可以用于对数据执行SQL查询，将结果返回为DataFrame。
  ```
  sparkSession.sql("select * from person").show
  ```
- SparkSession可以用于设置运行时的配置项，这些配置项可以在SQL中使用变量替换。
  ```
  sparkSession.conf.set("spark.some.config", "abcd")
  sparkSession.conf.get("spark.some.config")
  sparkSession.sql("select ${spark.some.config}")
  ```
- SparkSession包括一个“catalog”方法，其中包含使用Metastore（即数据目录）的方法。方法返回值为数据集，可以使用相同的Dataset API来运行。
  ```
  val tables = sparkSession.catalog.listTables()
  val columns = sparkSession.catalog.listColumns("myTable")
  ```
- 底层SparkContext可以通过SparkSession的SparkContext API访问。
  ```
  val sparkContext = sparkSession.sparkContext
  ```

Structured Streaming 原理

Structured Streaming是构建在Spark SQL引擎上的流式数据处理引擎，用户可以使用Scala、Java、Python或R中的Dataset/DataFrame API进行流数据聚合运算、按事件时间窗口计算、流流Join等操作。当流数据连续不断的产生时，Spark SQL将会增量的、持续不断的处理这些数据并将结果更新到结果集中。同时，系统通过checkpoint和Write Ahead Logs确保端到端的完全一次性容错保证。

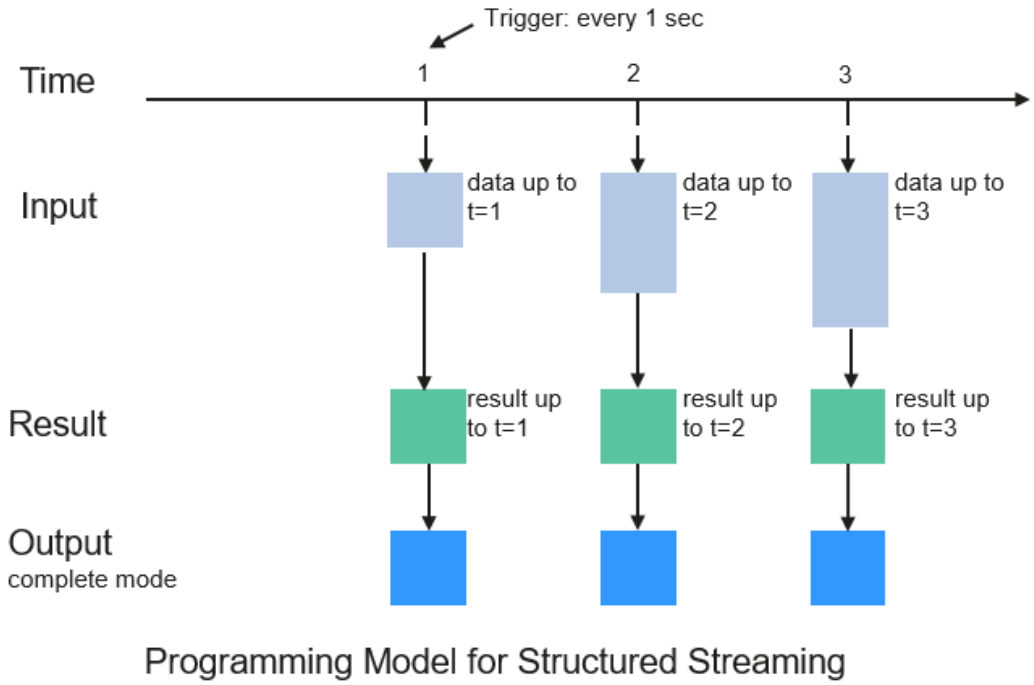
Structured Streaming的核心是将流式的数据看成一张不断增加的数据库表，这种流式的数据处理模型类似于数据块处理模型，可以把静态数据库表的一些查询操作应用在流式计算中，Spark执行标准的SQL查询，从不断增加的无边界表中获取数据。

图 1-111 Structured Streaming 无边界表



每一条查询的操作都会产生一个结果集Result Table。每一个触发间隔，当新的数据新增到表中，都会最终更新Result Table。无论何时结果集发生了更新，都能将变化的结果写入一个外部的存储系统。

图 1-112 Structured Streaming 数据处理模型



Structured Streaming在OutPut阶段可以定义不同的存储方式，有如下3种：

- Complete Mode：整个更新的结果集都会写入外部存储。整张表的写入操作将由外部存储系统的连接器完成。
- Append Mode：当时间间隔触发时，只有在Result Table中新增加的数据行会被写入外部存储。这种方式只适用于结果集中已经存在的内容不希望发生改变的情况下，如果已经存在的数据会被更新，不适合适用此种方式。
- Update Mode：当时间间隔触发时，只有在Result Table中被更新的数据才会被写入外部存储系统。注意，和Complete Mode方式的不同之处是不更新的结果集不会写入外部存储。

基本概念

- **RDD**

  即弹性分布数据集（Resilient Distributed Dataset），是Spark的核心概念。指的是一个只读的，可分区的分布式数据集，这个数据集的全部或部分可以缓存在内存中，在多次计算间重用。

  RDD的生成：

  - 从HDFS输入创建，或从与Hadoop兼容的其他存储系统中输入创建。
  - 从父RDD转换得到新RDD。
  - 从数据集合转换而来，通过编码实现。

  RDD的存储：

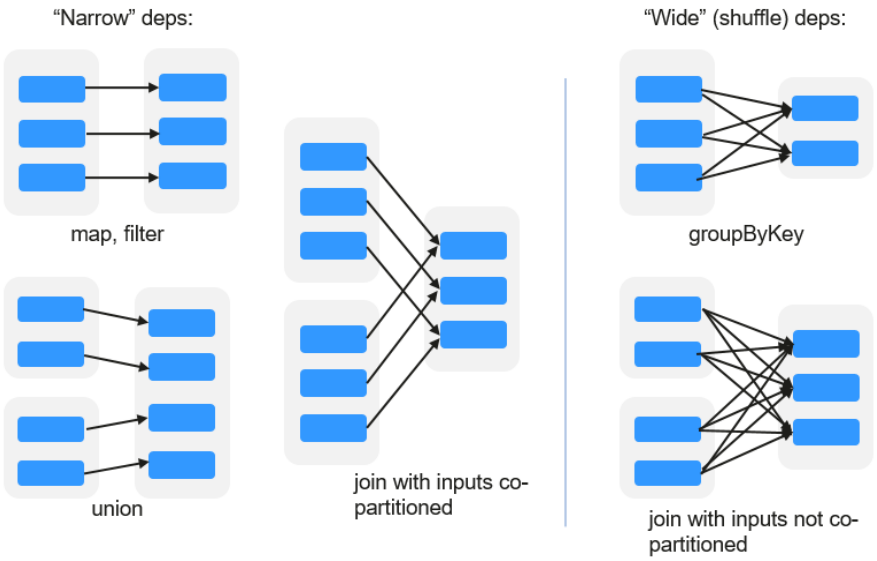
  - 用户可以选择不同的存储级别缓存RDD以便重用（RDD有11种存储级别）。

- 当前RDD默认是存储于内存，但当内存不足时，RDD会溢出到磁盘中。

- **Dependency（RDD的依赖）**

  RDD的依赖分别为：窄依赖和宽依赖。

  图 1-113 RDD 的依赖

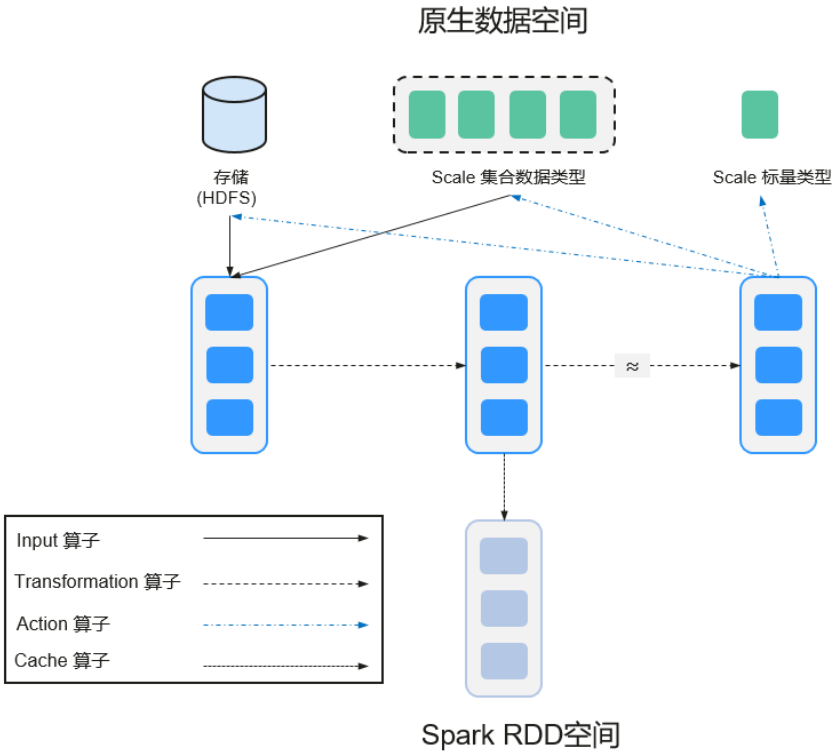
  - **窄依赖：**指父RDD的每一个分区最多被一个子RDD的分区所用。
  - **宽依赖：**指子RDD的分区依赖于父RDD的所有分区。

  窄依赖对优化很有利。逻辑上，每个RDD的算子都是一个fork/join（此join非上文的join算子，而是指同步多个并行任务的barrier）：把计算fork到每个分区，算完后join，然后fork/join下一个RDD的算子。如果直接翻译到物理实现，是很不经济的：一是每一个RDD（即使是中间结果）都需要物化到内存或存储中，费时费空间；二是join作为全局的barrier，是很昂贵的，会被最慢的那个节点拖死。如果子RDD的分区到父RDD的分区是窄依赖，就可以实施经典的fusion优化，把两个fork/join合为一个；如果连续的变换算子序列都是窄依赖，就可以把很多个fork/join并为一个，不但减少了大量的全局barrier，而且无需物化很多中间结果RDD，这将极大地提升性能。Spark把这个叫做流水线（pipeline）优化。

- **Transformation和Action（RDD的操作）**

  对RDD的操作包含Transformation（返回值还是一个RDD）和Action（返回值不是一个RDD）两种。RDD的操作流程如图1-114所示。其中Transformation操作是Lazy的，也就是说从一个RDD转换生成另一个RDD的操作不是马上执行，Spark在遇到Transformations操作时只会记录需要这样的操作，并不会去执行，需要等到有Actions操作的时候才会真正启动计算过程进行计算。Actions操作会返回结果或把RDD数据写到存储系统中。Actions是触发Spark启动计算的动因。

图 1-114 RDD 操作示例



RDD看起来与Scala集合类型没有太大差别，但数据和运行模型大相迥异。

```
val file = sc.textFile("hdfs://...")
val errors = file.filter(_.contains("ERROR"))
errors.cache()
errors.count()
```

a. textFile算子从HDFS读取日志文件，返回file（作为RDD）。

b. filter算子筛出带“ERROR”的行，赋给errors（新RDD）。filter算子是一个Transformation操作。

c. cache算子缓存下来以备未来使用。

d. count算子返回errors的行数。count算子是一个Action操作。

Transformation操作可以分为如下几种类型：

- 视RDD的元素为简单元素。

  输入输出一对一，且结果RDD的分区结构不变，主要是map。

  输入输出一对多，且结果RDD的分区结构不变，如flatMap（map后由一个元素变为一个包含多个元素的序列，然后展平为一个个的元素）。

  输入输出一对一，但结果RDD的分区结构发生了变化，如union（两个RDD合为一个，分区数变为两个RDD分区数之和）、coalesce（分区减少）。

  从输入中选择部分元素的算子，如filter、distinct（去除重复元素）、subtract（本RDD有、其他RDD无的元素留下来）和sample（采样）。

- 视RDD的元素为Key-Value对。

  对单个RDD做一对一运算，如mapValues（保持源RDD的分区方式，这与map不同）；

  对单个RDD重排，如sort、partitionBy（实现一致性的分区划分，这个对数据本地性优化很重要）；

对单个RDD基于key进行重组和reduce，如groupByKey、reduceByKey；

对两个RDD基于key进行join和重组，如join、cogroup。

说明

后三种操作都涉及重排，称为shuffle类操作。

Action操作可以分为如下几种：

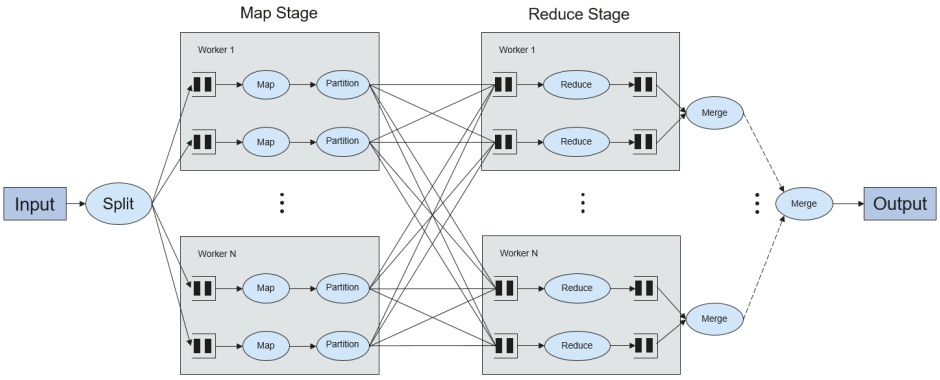
- 生成标量，如count（返回RDD中元素的个数）、reduce、fold/aggregate（返回几个标量）、take（返回前几个元素）。
- 生成Scala集合类型，如collect（把RDD中的所有元素倒入Scala集合类型）、lookup（查找对应key的所有值）。
- 写入存储，如与前文textFile对应的saveAsTextFile。
- 还有一个检查点算子checkpoint。当Lineage特别长时（这在图计算中时常发生），出错时重新执行整个序列要很长时间，可以主动调用checkpoint把当前数据写入稳定存储，作为检查点。

- **Shuffle**

  Shuffle是MapReduce框架中的一个特定的phase，介于Map phase和Reduce phase之间，当Map的输出结果要被Reduce使用时，每一条输出结果需要按key哈希，并且分发到对应的Reducer上去，这个过程就是shuffle。由于shuffle涉及到了磁盘的读写和网络的传输，因此shuffle性能的高低直接影响到了整个程序的运行效率。

  下图清晰地描述了MapReduce算法的整个流程。

  图 1-115 算法流程



  概念上shuffle就是一个沟通数据连接的桥梁，实际上shuffle这一部分是如何实现的呢，下面就以Spark为例讲解shuffle在Spark中的实现。

  Shuffle操作将一个Spark的Job分成多个Stage，前面的stages会包括一个或多个ShuffleMapTasks，最后一个stage会包括一个或多个ResultTask。

- **Spark Application的结构**

  Spark Application的结构可分为两部分：初始化SparkContext和主体程序。

  - 初始化SparkContext：构建Spark Application的运行环境。

    构建SparkContext对象，如：

    ```
    new SparkContext(master, appName, [SparkHome], [jars])
    ```

    参数介绍：

    master：连接字符串，连接方式有local、yarn-cluster、yarn-client等。

appName：构建的Application名称。

SparkHome：集群中安装Spark的目录。

jars：应用程序代码和依赖包。

– 主体程序：处理数据

- **Spark shell命令**

  Spark基本shell命令，支持提交Spark应用。命令为：

```
./bin/spark-submit \
  --class <main-class> \
  --master <master-url> \
  ... # other options
  <application-jar> \
  [application-arguments]
```

  参数解释：

  --class：Spark应用的类名。

  --master：Spark用于所连接的master，如yarn-client，yarn-cluster等。

  application-jar：Spark应用的jar包的路径。

  application-arguments：提交Spark应用的所需要的参数（可以为空）。

- **Spark JobHistory Server**

  用于监控正在运行的或者历史的Spark作业在Spark框架各个阶段的细节以及提供日志显示，帮助用户更细粒度地去开发、配置和调优作业。

1.4.24.2 Spark2x HA 方案介绍

1.4.24.2.1 Spark2x 多主实例

背景介绍

基于社区已有的JDBCServer基础上，采用多主实例模式实现了其高可用性方案。集群中支持同时共存多个JDBCServer服务，通过客户端可以随机连接其中的任意一个服务进行业务操作。即使集群中一个或多个JDBCServer服务停止工作，也不影响用户通过同一个客户端接口连接其他正常的JDBCServer服务。

多主实例模式相比主备模式的HA方案，优势主要体现在对以下两种场景的改进。

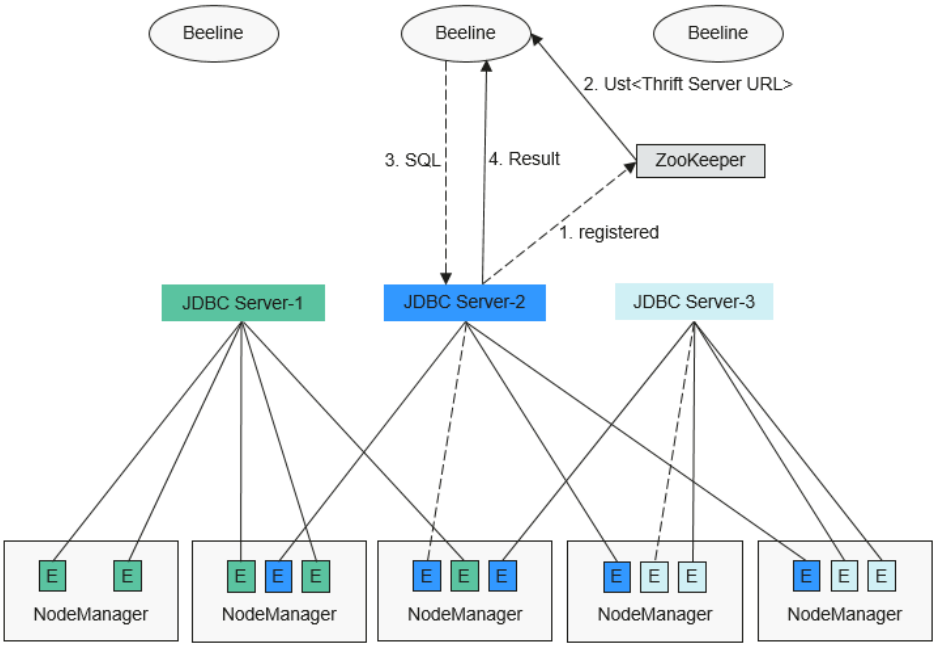
- 主备模式下，当发生主备切换时，会存在一段时间内服务不可用，该时间JDBCServer无法控制，取决于Yarn服务的资源情况。
- Spark中通过类似于HiveServer2的Thrift JDBC提供服务，用户通过Beeline以及JDBC接口访问。因此JDBCServer集群的处理能力取决于主Server的单点能力，可扩展性不够。

采用多主实例模式的HA方案，不仅可以规避主备切换服务中断的问题，实现服务不中断或少中断，还可以通过横向扩展集群来提高并发能力。

实现方案

多主实例模式的HA方案原理如下图所示。

图 1-116 Spark JDBCServer HA



1. JDBCServer在启动时，向ZooKeeper注册自身消息，在指定目录中写入节点，节点包含了该实例对应的IP，端口，版本号和序列号等信息（多节点信息之间以逗号隔开）。

   示例如下：

   ```
   [serverUri=192.168.169.84:22550
   ;version=8.1.2;sequence=0000001244,serverUri=192.168.195.232:22550 ;version=8.1.2;sequence=00000
   01242,serverUri=192.168.81.37:22550 ;version=8.1.2;sequence=0000001243]
   ```

2. 客户端连接JDBCServer时，需要指定Namespace，即访问ZooKeeper哪个目录下的JDBCServer实例。在连接的时候，会从Namespace下随机选择一个实例连接，详细URL参见**URL连接介绍**。
3. 客户端成功连接JDBCServer服务后，向JDBCServer服务发送SQL语句。
4. JDBCServer服务执行客户端发送的SQL语句后，将结果返回给客户端。

在HA方案中，每个JDBCServer服务（即实例）都是独立且等同的，当其中一个实例在升级或者业务中断时，其他的实例也能接受客户端的连接请求。

多主实例方案遵循以下规则：

- 当一个实例异常退出时，其他实例不会接管此实例上的会话，也不会接管此实例上运行的业务。
- 当JDBCServer进程停止时，删除在ZooKeeper上的相应节点。
- 由于客户端选择服务端的策略是随机的，可能会出现会话随机分配不均匀的情况，进而可能引起实例间的负载不均衡。
- 实例进入维护模式（即进入此模式后不再接受新的客户端连接）后，当达到退服超时时间，仍在此实例上运行的业务有可能会发生失败。

URL 连接介绍

多主实例模式

多主实例模式的客户端读取ZooKeeper节点中的内容，连接对应的JDBCServer服务。连接字符串为：

- 安全模式下：
  - Kinit认证方式下的JDBCURL如下所示：
    jdbc:hive2://*<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=**sparkthriftserver2x**;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;

    说明

    - 其中“<zkNode_IP>:<zkNode_Port>”是ZooKeeper的URL，多个URL以逗号隔开。

      例如：“192.168.81.37:2181,192.168.195.232:2181,192.168.169.84:2181”。
    - 其中“sparkthriftserver2x”是ZooKeeper上的目录，表示客户端从该目录下随机选择JDBCServer实例进行连接。

    示例：安全模式下通过Beeline客户端连接时执行以下命令：

    sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>***/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**
  - Keytab认证方式下的JDBCURL如下所示：
    jdbc:hive2://<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;user.principal=<principal_name>;user.keytab=<path_to_keytab>

    其中<principal_name>表示用户使用的Kerberos用户的principal，如“test@*<系统域名>*”。<path_to_keytab>表示<principal_name>对应的keytab文件路径，如“/opt/auth/test/user.keytab”。
- 普通模式下：
  jdbc:hive2://*<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=**sparkthriftserver2x**;

  示例：普通模式下通过Beeline客户端连接时执行以下命令：

  sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>***/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;"**

非多主实例模式

非多主实例模式的客户端连接的是某个指定JDBCServer节点。该模式的连接字符串相比多主实例模式的去掉关于Zookeeper的参数项“serviceDiscoveryMode”和“zooKeeperNamespace”。

示例：安全模式下通过Beeline客户端连接非多主实例模式时执行以下命令：

sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<server_IP>:<server_Port>***/;user.principal=spark2x/hadoop.<系统域名>@<系统域名>;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**

说明

- 其中“<server_IP>:<server_Port>”是指定JDBCServer节点的URL。
- “CLIENT_HOME”是指客户端路径。

多主实例模式与非多主实例模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。

1.4.24.2.2 Spark2x 多租户

背景介绍

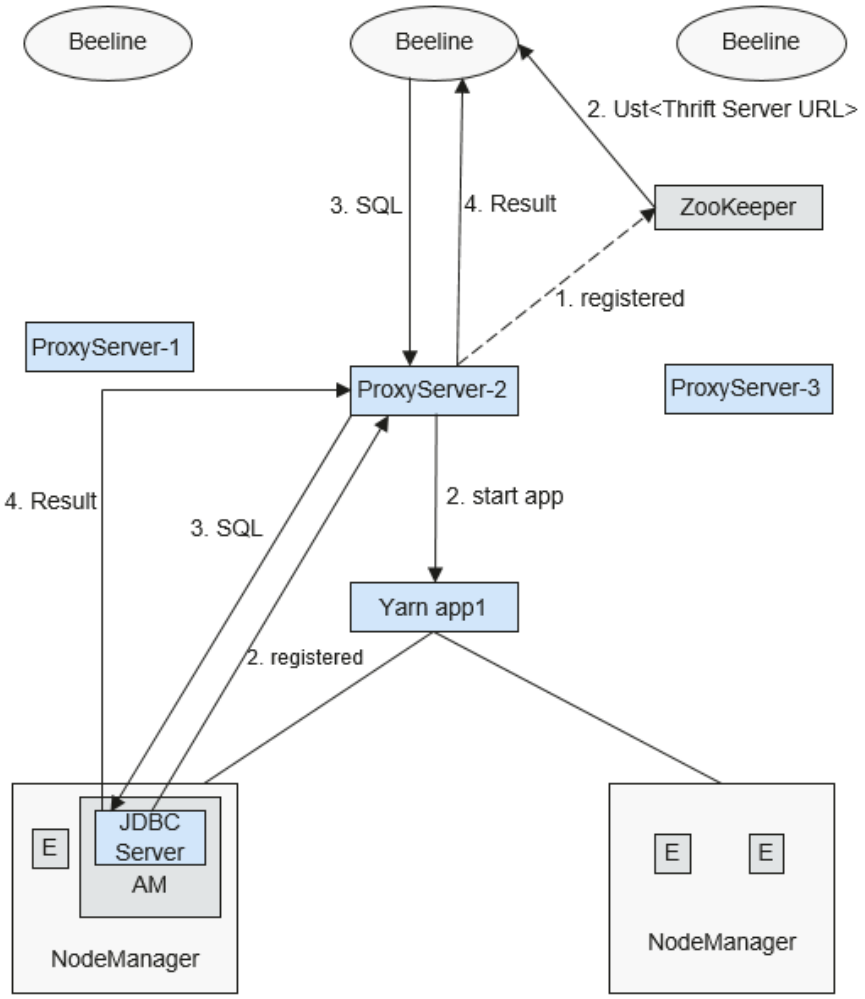
JDBCServer多主实例方案中，JDBCServer实现使用YARN-Client模式，但YARN资源队列只有一个，为了解决这种资源局限的问题，引入了多租户模式。

多租户模式是将JDBCServer和租户绑定，每一个租户对应一个或多个JDBCServer，而一个JDBCServer只给一个租户提供服务。不同的租户可以配置不同的YARN队列，从而达到资源隔离，且JDBCServer根据需求动态启动，可避免浪费资源。

实现方案

多租户模式的HA方案原理如图1-117所示。

图 1-117 Spark JDBCServer 多租户



1. ProxyServer在启动时，向ZooKeeper注册自身消息，在指定目录中写入节点信息，节点信息包含了该实例对应的IP，端口，版本号和序列号等信息（多节点信息之间以逗号隔开）。

   说明

   多租户模式下，MRS页面上的JDBCServer实例是指ProxyServer（JDBCServer代理）。

   示例如下：

   ```
   serverUri=192.168.169.84:22550
   ;version=8.1.2;sequence=0000001244,serverUri=192.168.195.232:22550
   ;version=8.1.2;sequence=0000001242,serverUri=192.168.81.37:22550
   ;version=8.1.2;sequence=0000001243,
   ```

2. 客户端连接ProxyServer时，需要指定Namespace，即访问ZooKeeper哪个目录下的ProxyServer实例。在连接的时候，会从Namespace下随机选择一个实例连接，详细URL参见**URL连接介绍**。
3. 客户端成功连接ProxyServer服务，ProxyServer服务首先确认是否有该租户的JDBCServer存在，如果有，直接将Beeline连上真正的JDBCServer；如果没有，则以YARN-Cluster模式启动一个新的JDBCServer。JDBCServer启动成功后，ProxyServer会获取JDBCServer的地址，并将Beeline连上JDBCServer。
4. 客户端发送SQL语句给ProxyServer，ProxyServer将语句转交给真正连上的JDBCServer处理。最后JDBCServer服务将结果返回给ProxyServer，ProxyServer再将结果返回给客户端。

在HA方案中，每个ProxyServer服务（即实例）都是独立且等同的，当其中一个实例在升级或者业务中断时，其他的实例也能接受客户端的连接请求。

URL 连接介绍

多租户模式

多租户模式的客户端读取ZooKeeper节点中的内容，连接对应的ProxyServer服务。连接字符串为：

- 安全模式下：
  - Kinit认证方式下的客户端URL如下所示：

    ```
    jdbc:hive2://
    <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;
    ```

    说明

    - 其中“<zkNode_IP>:<zkNode_Port>”是ZooKeeper的URL，多个URL以逗号隔开。

      例如：“192.168.81.37:2181,192.168.195.232:2181,192.168.169.84:2181”。
    - 其中sparkthriftserver2x是ZooKeeper上的目录，表示客户端从该目录下随机选择JDBCServer实例进行连接。

    示例：安全模式下通过Beeline客户端连接时执行以下命令：

    sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://** *<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/**;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;"**

  - Keytab认证方式下的URL如下所示：

```
jdbc:hive2://
<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;user.principal=<principal_name>;user.keytab=<path_to_keytab>
```

其中<principal_name>表示用户使用的Kerberos用户的principal，如“test@*<系统域名>*”。<path_to_keytab>表示<principal_name>对应的keytab文件路径，如“/opt/auth/test/user.keytab”。

- 普通模式下：

  ```
  jdbc:hive2://
  <zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>|;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;
  ```

  示例：普通模式下通过Beeline客户端连接时执行以下命令：

  sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/**;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;"**

非多租户模式

非多租户模式的客户端连接的是某个指定JDBCServer节点。该模式的连接字符串相比多主实例模式的去掉关于ZooKeeper的参数项“serviceDiscoveryMode”和“zooKeeperNamespace”。

示例：安全模式下通过Beeline客户端连接非多租户模式时执行以下命令：

sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<server_IP>:<server_Port>*/**;user.principal=spark2x/hadoop.<系统域名>@<系统域名>;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**

说明

- 其中“<server_IP>:<server_Port>”是指定JDBCServer节点的URL。
- “CLIENT_HOME”是指客户端路径。

多租户模式与非多租户模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同

指定租户

一般情况下，某用户提交的客户端会连接到该用户默认所属租户的JDBCServer上，若需要连接客户端到指定租户的JDBCServer上，可以通过添加**--hiveconf mapreduce.job.queuename**进行指定。

通过Beeline连接的命令示例如下（aaa为租户名称）：

beeline --hiveconf mapreduce.job.queuename=aaa -u 'jdbc:hive2://192.168.39.30:2181,192.168.40.210:2181,192.168.215.97:2181;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;'

1.4.24.3 Spark2x 与组件的关系

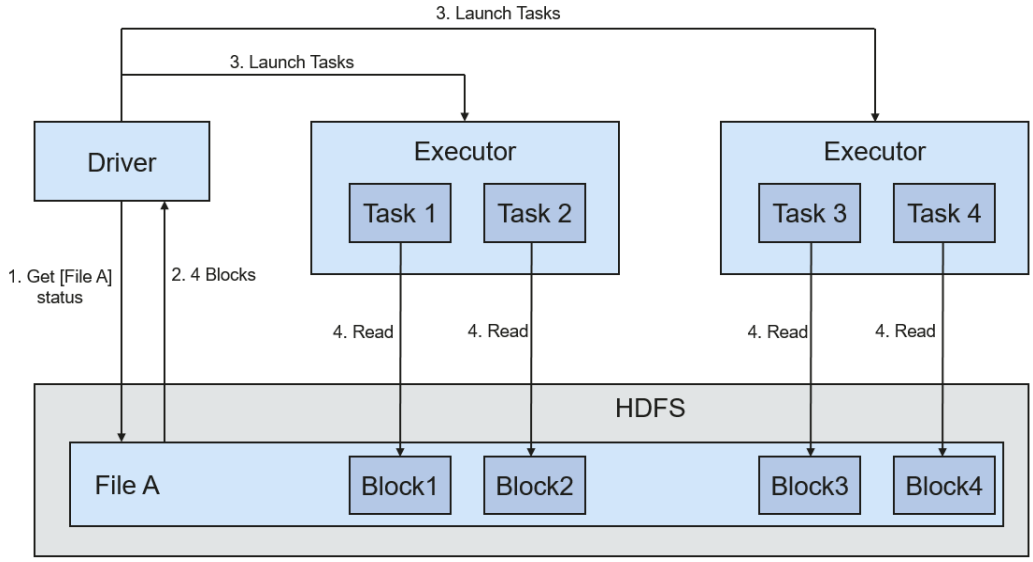
Spark 和 HDFS 的配合关系

通常，Spark中计算的数据可以来自多个数据源，如Local File、HDFS等。最常用的是HDFS，用户可以一次读取大规模的数据进行并行计算。在计算完成后，也可以将数据存储到HDFS。

分解来看，Spark分成控制端(Driver)和执行端（ Executor ）。控制端负责任务调度，执行端负责任务执行。

读取文件的过程如图1-118所示。

图 1-118 读取文件过程

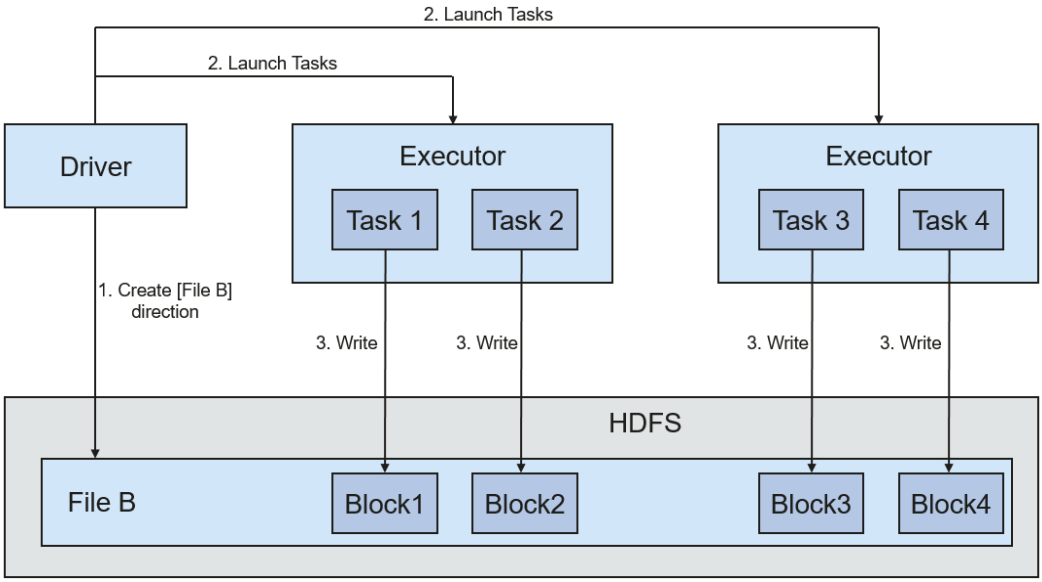


读取文件步骤的详细描述如下所示：

1. Driver与HDFS交互获取File A的文件信息。
2. HDFS返回该文件具体的Block信息。
3. Driver根据具体的Block数据量，决定一个并行度，创建多个Task去读取这些文件Block。
4. 在Executor端执行Task并读取具体的Block，作为RDD(弹性分布数据集)的一部分。

写入文件的过程如图1-119所示。

图 1-119 写入文件过程



HDFS文件写入的详细步骤如下所示：

1. Driver创建要写入文件的目录。
2. 根据RDD分区分块情况，计算出写数据的Task数，并下发这些任务到Executor。
3. Executor执行这些Task，将具体RDD的数据写入到步骤1创建的目录下。

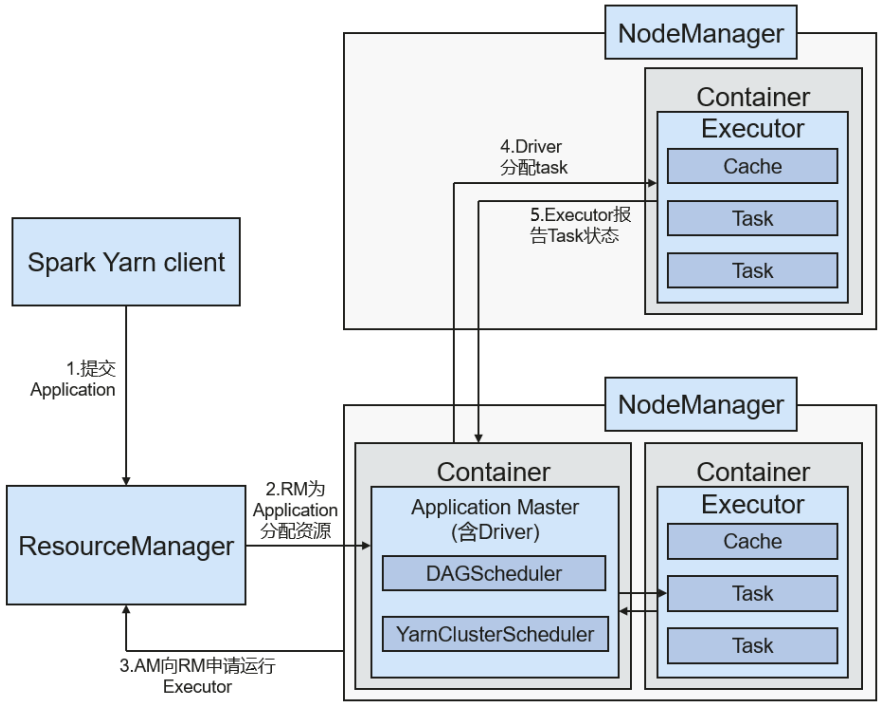
Spark 和 YARN 的配合关系

Spark的计算调度方式，可以通过YARN的模式实现。Spark共享YARN集群提供丰富的计算资源，将任务分布式的运行起来。Spark on YARN分两种模式：YARN Cluster和YARN Client。

- YARN Cluster模式

  运行框架如图1-120所示。

图 1-120 Spark on yarn-cluster 运行框架



Spark on YARN-Cluster实现流程：

a. 首先由客户端生成Application信息，提交给ResourceManager。

b. ResourceManager为Spark Application分配第一个Container(ApplicationMaster)，并在该Container上启动Driver。

c. ApplicationMaster向ResourceManager申请资源以运行Container。

ResourceManager分配Container给ApplicationMaster，ApplicationMaster和相关的NodeManager通讯，在获得的Container上启动Executor，Executor启动后，开始向Driver注册并申请Task。

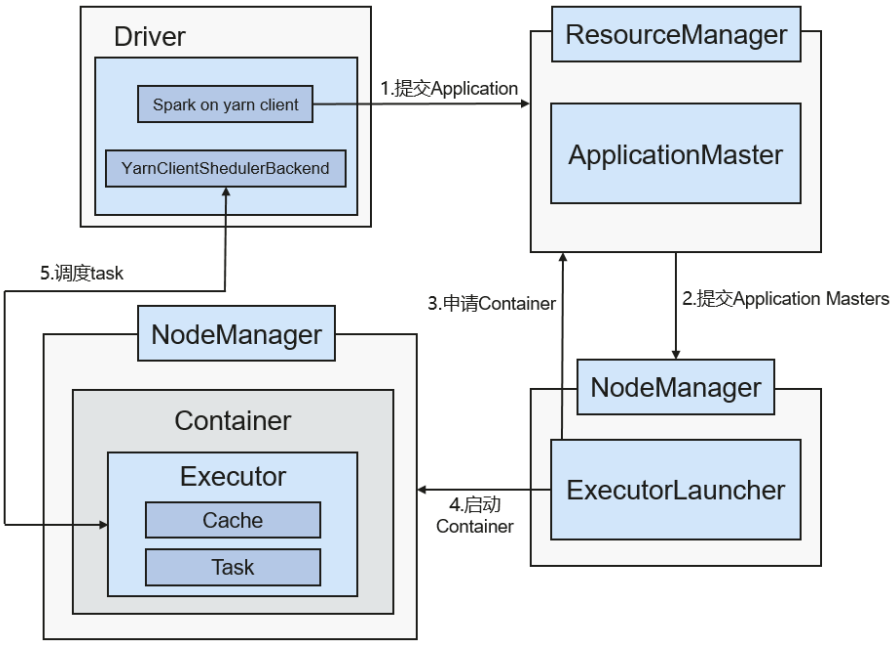
d. Driver分配Task给Executor执行。

e. Executor执行Task并向Driver汇报运行状况。

- YARN Client模式

  运行框架如图1-121所示。

图 1-121 Spark on yarn-client 运行框架



Spark on YARN-Client实现流程：

说明

在YARN-Client模式下，Driver部署在Client端，在Client端启动。YARN-Client模式下，不兼容老版本的客户端。推荐使用YARN-Cluster模式。

a. 客户端向ResourceManager发送Spark应用提交请求，Client端将启动ApplicationMaster所需的所有信息打包，提交给ResourceManager上，ResourceManager为其返回应答，该应答中包含多种信息(如ApplicationId、可用资源使用上限和下限等)。ResourceManager收到请求后，会为ApplicationMaster寻找合适的节点，并在该节点上启动它。ApplicationMaster是Yarn中的角色，在Spark中进程名字是ExecutorLauncher。

b. 根据每个任务的资源需求，ApplicationMaster可向ResourceManager申请一系列用于运行任务的Container。

c. 当ApplicationMaster（从ResourceManager端）收到新分配的Container列表后，会向对应的NodeManager发送信息以启动Container。

   ResourceManager分配Container给ApplicationMaster，ApplicationMaster和相关的NodeManager通讯，在获得的Container上启动Executor，Executor启动后，开始向Driver注册并申请Task。

   说明

   正在运行的Container不会被挂起释放资源。

d. Driver分配Task给Executor执行。Executor执行Task并向Driver汇报运行状况。

1.4.24.4 Spark2x 开源新特性

概述

Spark2x版本相对于Spark 1.5版本新增了一些开源特性。具体特性或相关概念如下：

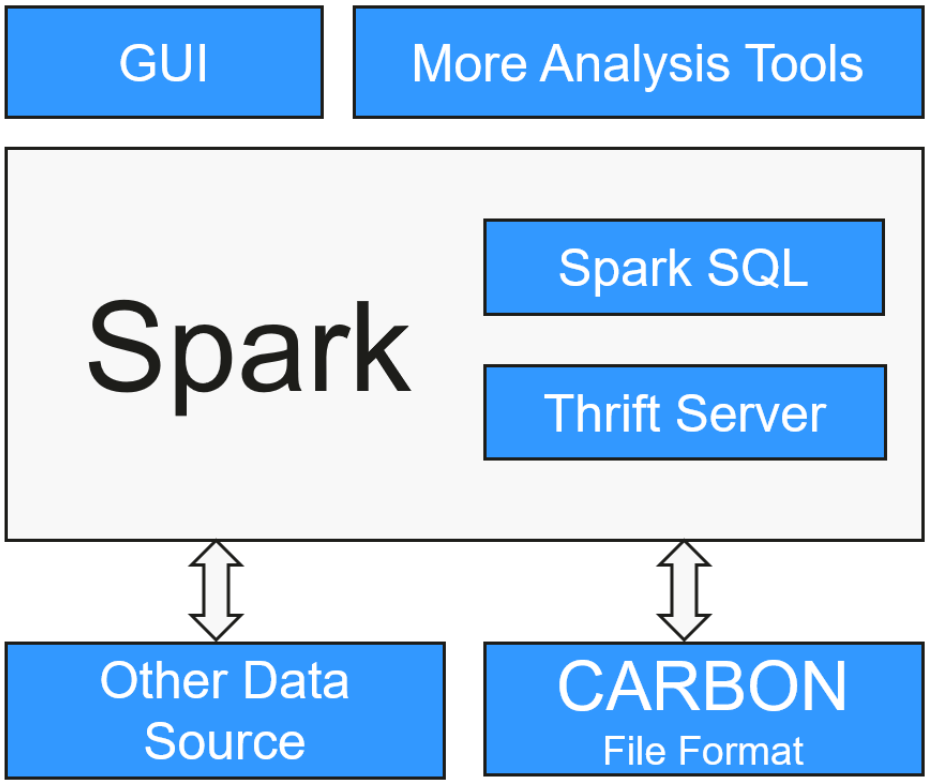
- DataSet，详见**SparkSQL和DataSet原理**。
- Spark SQL Native DDL/DML，详见**SparkSQL和DataSet原理**。
- SparkSession，详见**SparkSession原理**。
- Structured Streaming，详见**Structured Streaming原理**。
- 小文件优化。
- 聚合算法优化。
- Datasource表优化。
- 合并CBO优化。

1.4.24.5 Spark2x 开源增强特性

1.4.24.5.1 CarbonData 简介

CarbonData是一种新型的Apache Hadoop本地文件格式，使用先进的列式存储、索引、压缩和编码技术，以提高计算效率，有助于加速超过PB数量级的数据查询，可用于更快的交互查询。同时，CarbonData也是一种将数据源与Spark集成的高性能分析引擎。

图 1-122 CarbonData 基本架构



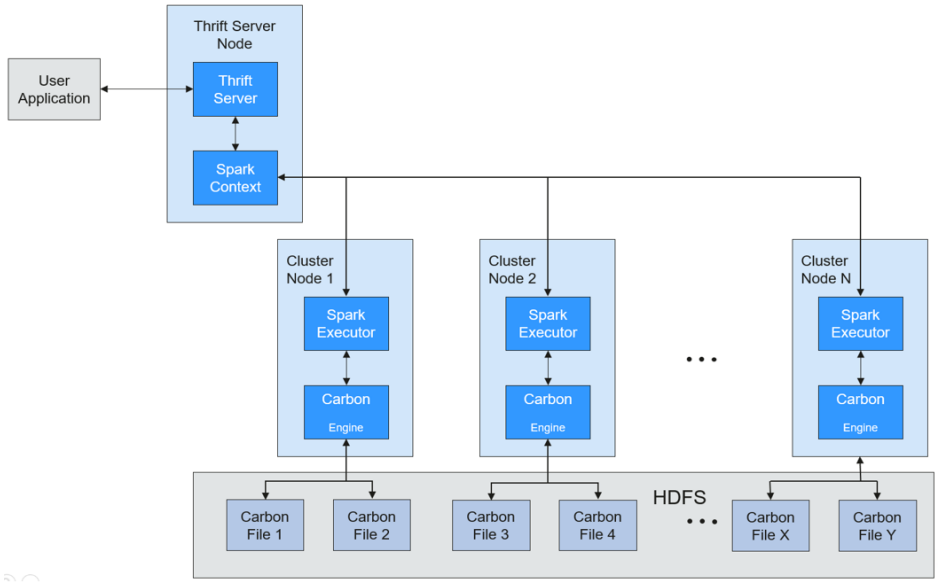
使用CarbonData的目的是对大数据即席查询提供超快速响应。从根本上说，CarbonData是一个OLAP引擎，采用类似于RDBMS中的表来存储数据。用户可将大量（10TB以上）的数据导入以CarbonData格式创建的表中，CarbonData将以压缩的多维索引列格式自动组织和存储数据。数据被加载到CarbonData后，就可以执行即席查询，CarbonData将对数据查询提供秒级响应。

CarbonData将数据源集成到Spark生态系统，用户可使用Spark SQL执行数据查询和分析。也可以使用Spark提供的第三方工具JDBCServer连接到Spark SQL。

CarbonData 结构

CarbonData作为Spark内部数据源运行，不需要额外启动集群节点中的其他进程，CarbonData Engine在Spark Executor进程之中运行。

图 1-123 CarbonData 结构



存储在CarbonData Table中的数据被分成若干个CarbonData数据文件，每一次数据查询时，CarbonData Engine模块负责执行数据集的读取、过滤等实际任务。CarbonData Engine作为Spark Executor进程的一部分运行，负责处理数据文件块的一个子集。

Table数据集数据存储在HDFS中。同一Spark集群内的节点可以作为HDFS的数据节点。

CarbonData 特性

- SQL功能：CarbonData与Spark SQL完全兼容，支持所有可以直接在Spark SQL上运行的SQL查询操作。
- 简单的Table数据集定义：CarbonData支持易于使用的DDL(数据定义语言)语句来定义和创建数据集。CarbonData DDL十分灵活、易于使用，并且足够强大，可以定义复杂类型的Table。
- 便捷的数据管理：CarbonData为数据加载和维护提供多种数据管理功能。CarbonData支持加载历史数据以及增量加载新数据。加载的数据可以基于加载时间进行删除，也可以撤销特定的数据加载操作。

- CarbonData文件格式是HDFS中的列式存储格式。该格式具有许多新型列存储文件的特性，例如，分割表，数据压缩等。CarbonData具有以下独有的特点：
  - 伴随索引的数据存储：由于在查询中设置了过滤器，可以显著加快查询性能，减少I/O扫描次数和CPU资源占用。CarbonData索引由多个级别的索引组成，处理框架可以利用这个索引来减少需要安排和处理的任务，也可以通过在任务扫描中以更精细的单元（称为blocklet）进行skip扫描来代替对整个文件的扫描。
  - 可选择的数据编码：通过支持高效的数据压缩和全局编码方案，可基于压缩/编码数据进行查询，在将结果返回给用户之前，才将编码转化为实际数据，这被称为“延迟物化”。
  - 支持一种数据格式应用于多种用例场景：例如，交互式OLAP-style查询，顺序访问（big scan），随机访问（narrow scan）。

CarbonData 关键技术和优势

- 快速查询响应：高性能查询是CarbonData关键技术的优势之一。CarbonData查询速度大约是Spark SQL查询的10倍。CarbonData使用的专用数据格式围绕高性能查询进行设计，其中包括多种索引技术、全局字典编码和多次的Push down优化，从而对TB级数据查询进行最快响应。
- 高效率数据压缩：CarbonData使用轻量级压缩和重量级压缩的组合压缩算法压缩数据，可以减少60%~80%数据存储空间，很大程度上节省硬件存储成本。

CarbonData 索引缓存服务器

为了解决日益增长的数据量给driver带来的压力与出现的各种问题，现引入单独的索引缓存服务器，将索引从Carbon查询的Spark应用侧剥离。所有的索引内容全部由索引缓存服务器管理，Spark应用通过RPC方式获取需要的索引数据。这样，释放了大量的业务侧的内存，使得业务不会受集群规模影响而性能或者功能出现问题。

1.4.24.5.2 跨源复杂数据的 SQL 查询优化

场景描述

出于管理和信息收集的需要，企业内部会存储海量数据，包括数目众多的各种数据库、数据仓库等，此时会面临以下困境：数据源种类繁多，数据集结构化混合，相关数据存放分散等，这就导致了跨源复杂查询因传输效率低，耗时长。

当前开源Spark在跨源查询时，只能对简单的filter进行下推，因此造成大量不必要的数据传输，影响SQL引擎性能。针对下推能力进行增强，当前对aggregate、复杂projection、复杂predicate均可以下推到数据源，尽量减少不必要数据的传输，提升查询性能。

目前仅支持JDBC数据源的查询下推，支持的下推模块有aggregate、projection、predicate、aggregate over inner join、aggregate over union all等。为应对不同应用场景的特殊需求，对所有下推模块设计开关功能，用户可以自行配置是否应用上述查询下推的增强。

表 1-22 跨源查询增加特性对比

模块	增强前	增强后
aggregate	不支持aggregate下推	• 支持的聚合函数为：sum, avg, max, min, count 例如：select count(*) from table • 支持聚合函数内部表达式 例如：select sum(a+b) from table • 支持聚合函数运算，例如：select avg(a) + max(b) from table • 支持having下推 例如：select sum(a) from table where a>0 group by b having sum(a)>10 • 支持部分函数下推 支持对abs()、month()、length()等数学、时间、字符串函数进行下推。并且，除了以上内置函数，用户还可以通过SET命令新增数据源支持的函数。 例如：select sum(abs(a)) from table • 支持aggregate之后的limit、order by下推（由于Oracle不支持limit，所以Oracle中limit、order by不会下推） 例如：select sum(a) from table where a>0 group by b order by sum(a) limit 5
projection	仅支持简单projection下推，例如：select a, b from table	• 支持复杂表达式下推。 例如：select (a+b)*c from table • 支持部分函数下推，详细参见表下方的说明。 例如：select length(a)+abs(b) from table • 支持projection之后的limit、order by下推。 例如：select a, b+c from table order by a limit 3
predicate	仅支持运算符左边为列名右边为值的简单filter，例如 select * from table where a>0 or b in（“aaa”，“bbb”）	• 支持复杂表达数下推 例如：select * from table where a+b>c*d or a/c in (1, 2, 3) • 支持部分函数下推，详细参见表下方的说明。 例如：select * from table where length(a)>5

模块	增强前	增强后
aggregate over inner join	需要将两个表中相关的数据全部加载到Spark，先进行join操作，再进行aggregate操作	支持以下几种： ● 支持的聚合函数为：sum, avg, max, min,count ● 所有aggregate只能来自同一个表，group by可以来自一个表或者两个表，只支持inner join。 不支持的情形有： ● 不支持aggregate同时来自join左表和右表的下推。 ● 不支持aggregate内包含运算，如：sum(a+b)。 ● 不支持aggregate运算，如：sum(a)+min(b)。
aggregate over union all	需要将两个表中相关的数据全部加载到Spark，先进行union操作，再进行aggregate操作	支持情况： 支持的聚合函数为：sum, avg, max, min,count 不支持的情况： ● 不支持aggregate内包含运算，如：sum(a+b)。 ● 不支持aggregate运算，如：sum(a)+min(b)。

注意事项

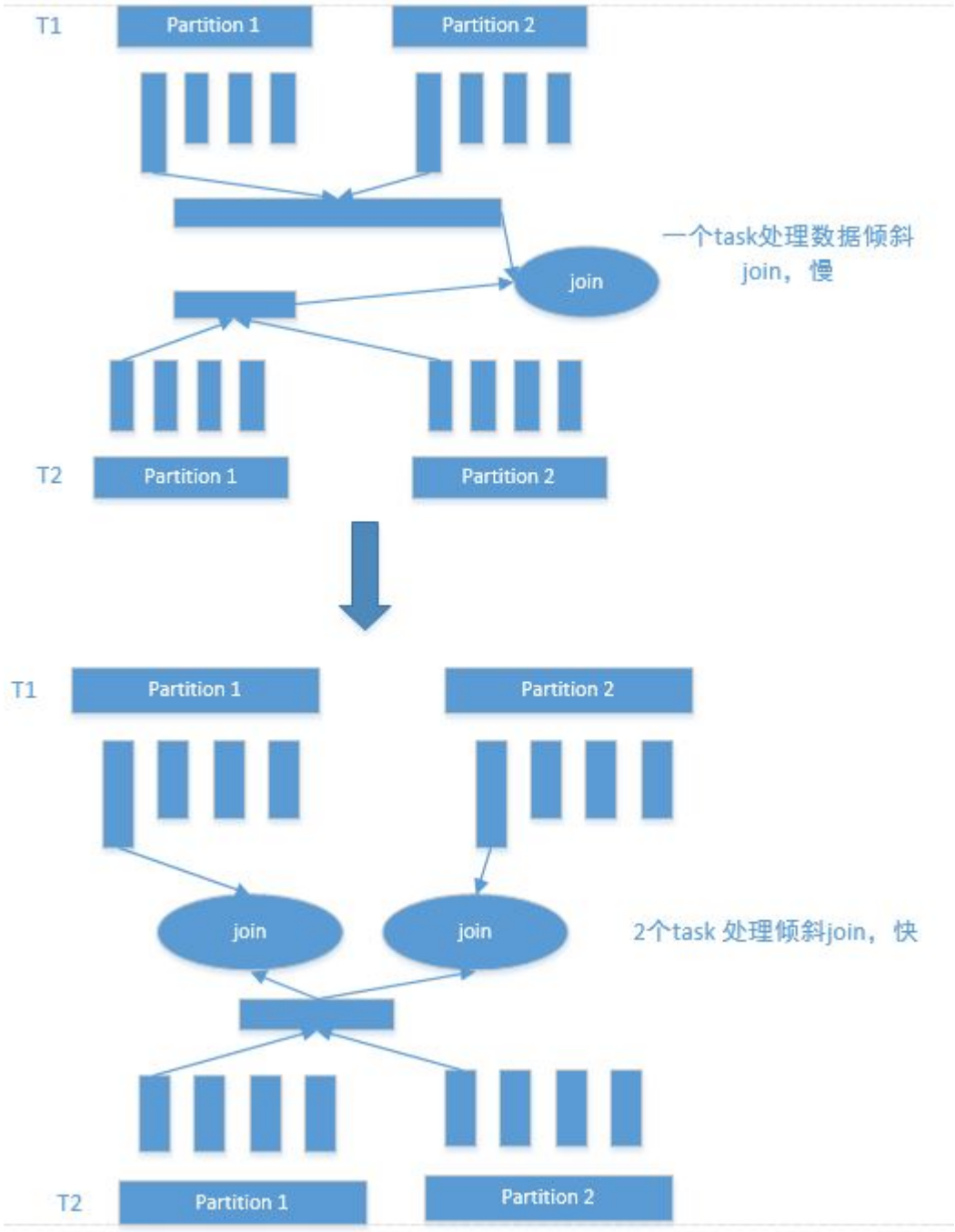
- 外部数据源是Hive的场景，通过Spark建的外表无法进行查询。
- 数据源只支持MySQL和Mppdb。

1.4.24.5.3 数据倾斜优化

在Spark SQL多表Join的场景下，会存在关联键严重倾斜的情况，导致Hash分桶后，部分桶中的数据远高于其他分桶。最终导致部分Task过重，运行得很慢；其他Task过轻，运行得很快。一方面，数据量大Task运行慢，使得计算性能低；另一方面，数据量少的Task在运行完成后，导致很多CPU空闲，造成CPU资源浪费。

针对数据倾斜的情况，可以通过配置“spark.sql.adaptive.skewjoin.threshold”配置项，打开数据倾斜优化特性，即可感知数据分桶的桶大小。此时，如果某个桶数据量过大，发生了数据倾斜，则把倾斜的那个桶拆小，把倾斜数据平均到多个task里边进行处理，每个task对join表相同桶的数据进行全量拉取，从而充分利用CPU资源，提升整体的性能。

图 1-124 倾斜数据 Join 转换



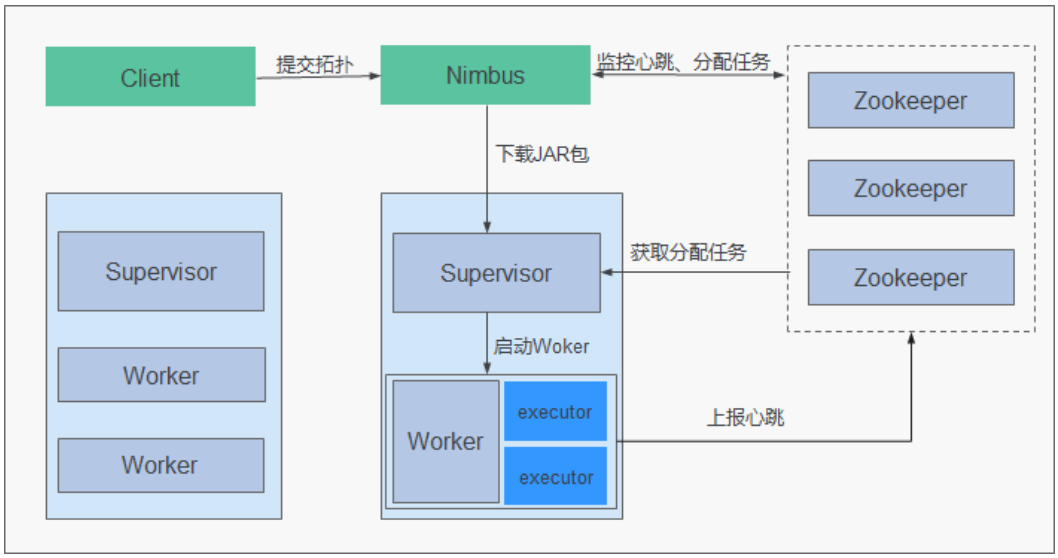
1.4.25 Storm

1.4.25.1 Storm 基本原理

Apache Storm是一个分布式、可靠、容错的实时流式数据处理的系统。在Storm中，先要设计一个用于实时计算的图状结构，称之为拓扑（topology）。这个拓扑将会被提交给集群，由集群中的主控节点（master node）分发代码，将任务分配给工作节点（worker node）执行。一个拓扑中包括spout和bolt两种角色，其中spout发送消息，

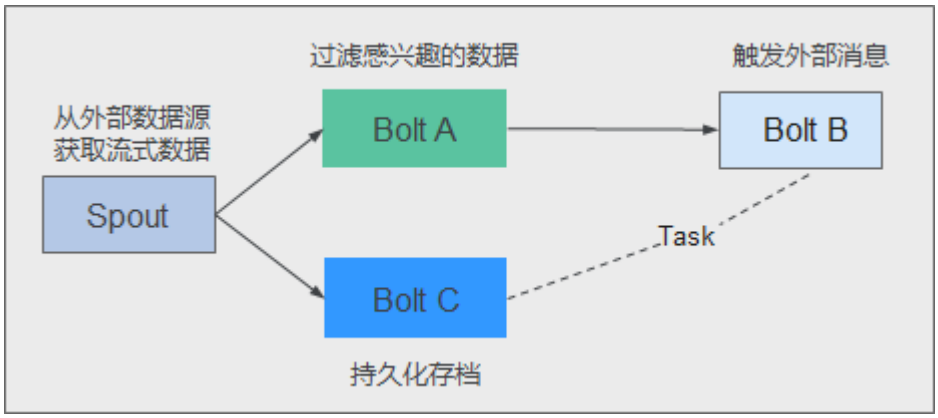
负责将数据流以tuple元组的形式发送出去；而bolt则负责转换这些数据流，在bolt中可以完成计算、过滤等操作，bolt自身也可以随机将数据发送给其他bolt。由spout发射出的tuple是不可变数组，对应着固定的键值对。

图 1-125 Storm 系统架构



业务处理逻辑被封装进Storm中的Topology中。一个Topology是由一组Spout组件（数据源）和Bolt组件（逻辑处理）通过Stream Groupings进行连接的有向无环图（DAG）。Topology里面的每一个Component（Spout/Bolt）节点都是并行运行的。在Topology里面， 可以指定每个节点的并行度，Storm则会在集群里面分配相应的Task来同时计算，以增强系统的处理能力。

图 1-126 Topology



Storm有众多适用场景：实时分析、持续计算、分布式ETL等。Storm有如下几个特点：

- 适用场景广泛
- 易扩展，可伸缩性高
- 保证无数据丢失
- 容错性好

- 易于构建和操控
- 多语言

Storm作为计算平台，在业务层为用户提供了更为易用的业务实现方式：CQL（Continuous Query Language—持续查询语言）。CQL具有以下几个特点：

- 使用简单：CQL语法和标准SQL语法类似，只要具备SQL基础，通过简单地学习，即可快速地进行业务开发。
- 功能丰富：CQL除了包含标准SQL的各类基本表达式等功能之外，还特别针对流处理场景增加了窗口、过滤、并发度设置等功能。
- 易于扩展：CQL提供了拓展接口，以支持日益复杂的业务场景，用户可以自定义输入、输出、序列化、反序列化等功能来满足特定的业务场景
- 易于调试：CQL提供了详细的异常码说明，降低了用户对各种错误的处理难度。

关于Storm的架构和详细原理介绍，请参见：**https://storm.apache.org/**。

Storm 原理

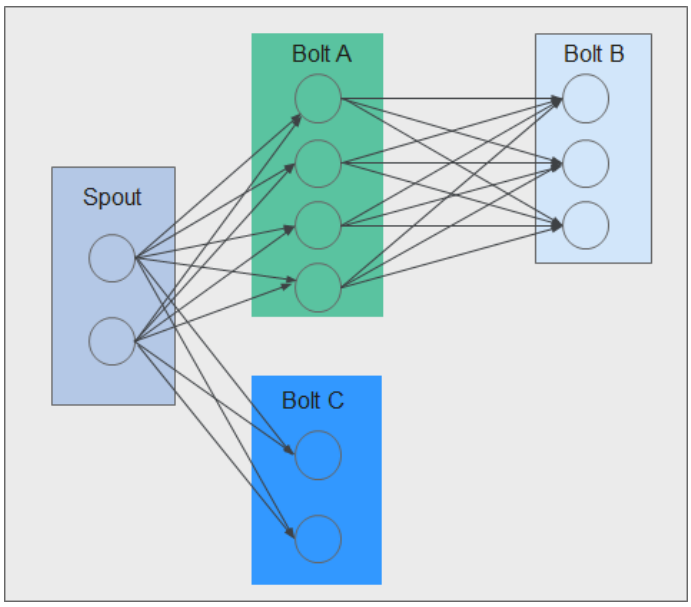
- **基本概念**

表 1-23 概念介绍

概念	说明
Tuple	Storm核心数据结构，是消息传递的基本单元，不可变Key-Value对，这些Tuple会以一种分布式的方式进行创建和处理。
Stream	Storm的关键抽象，是一个无边界的连续Tuple序列。
Topology	在Storm平台上运行的一个实时应用程序，由各个组件（Component）组成的一个DAG（Directed Acyclic Graph）。一个Topology可以并发地运行在多台机器上，每台机器上可以运行该DAG中的一部分。Topology与Hadoop中的MapReduce Job类似，不同的是，它是一个长驻程序，一旦开始就不会停止，除非人工中止。
Spout	Topology中产生源数据的组件，是Tuple的来源，通常可以从外部数据源（如消息队列、数据库、文件系统、TCP连接等）读取数据，然后转换为Topology内部的数据结构Tuple，由下一级组件处理。
Bolt	Topology中接受数据并执行具体处理逻辑（如过滤，统计、转换、合并、结果持久化等）的组件。
Worker	是Topology运行态的物理进程。每个Worker是一个JVM进程，每个Topology可以由多个Worker并行执行，每个Worker运行Topology中的一个逻辑子集。
Task	Worker中每一个Spout/Bolt的线程称为一个Task。
Stream groupings	Storm中的Tuple分发策略，即后一级Bolt以什么分发方式来接收数据。当前支持的策略有：Shuffle Grouping， Fields Grouping， All Grouping， Global Grouping， Non Grouping， Directed Grouping。

图1-127描述了一个由Spout、Bolt组成的DAG，即Topology。图中每个矩型框代表Spout或者Bolt，矩型框内的节点表示各个并发的Task，Task之间的“边”代表数据流——Stream。

图 1-127 Topology 示意图



- **可靠性**

  Storm提供三种级别的数据可靠性：

  - 至多一次：处理的数据可能会丢失，但不会被重复处理。此情况下，系统吞吐量最大。
  - 至少一次：保证数据传输可靠，但可能会被重复处理。此情况下，对在超时时间内没有获得成功处理响应的数据，会在Spout处进行重发，供后续Bolt再次处理，会对性能稍有影响。
  - 精确一次：数据成功传递，不丢失，不冗余处理。此情况下，性能最差。

  可靠性不同级别的选择，需要根据业务对可靠性的要求来选择、设计。例如对于一些对数据丢失不敏感的业务，可以在业务中不考虑数据丢失处理从而提高系统性能；而对于一些严格要求数据可靠性的业务，则需要使用精确一次的可靠性方案，以确保数据被处理且仅被处理一次。

- **容错**

  Storm是一个容错系统，提供较高可用性。表1-24从Storm的不同部件失效的情况角度解释其容错能力：

表 1-24 容错能力

失效场景	说明
Nimbus失效	Nimbus是无状态且快速失效的。当主Nimbus失效时，备Nimbus会接管，并对外提供服务。

失效场景	说明
Supervisor失效	Supervisor是工作节点的后台守护进程，是一种快速失效机制，且是无状态的，并不影响正在该节点上运行的Worker，但是会无法接收新的Worker分配。当Supervisor失效时，OMS会侦测到，并及时重启该进程。
Worker失效	该Worker所在节点上的Supervisor会在此节点上重新启动该Worker。如果多次重启失败，则Nimbus会将该任务重新分配到其它节点。
节点失效	该节点上的所有分配的任务会超时，而Nimbus会将这些Worker重新分配到其他节点。

Storm 开源特性

- 分布式实时计算框架

  开源Storm集群中的每台机器上都可以运行多个工作进程，每个工作进程又可创建多个线程，每个线程可以执行多个任务，任务是并发进行数据处理。
- 高容错

  如果在消息处理过程中有节点、进程等出现异常，提供重新部署该处理单元的能力。
- 可靠的消息保证

  支持At-Least Once、At-Most Once、Exactly Once的数据处理模式。
- 安全机制

  提供基于Kerberos的认证以及可插拔的授权机制，提供支持SSL的Storm UI以及Log Viewer界面，同时支持与大数据平台其他组件（如ZooKeeper，HDFS等）进行安全集成。
- 灵活的拓扑定义及部署

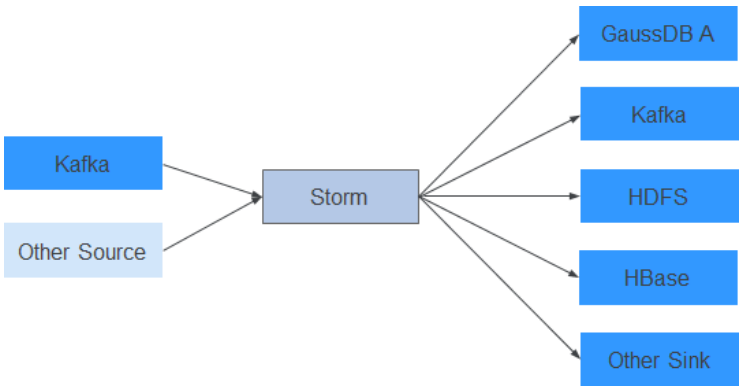
  使用Flux框架定义及部署业务拓扑，在业务DAG发生变化时，只需对YAML DSL（domain-specific language）定义进行修改，无需重新编译及打包业务代码。
- 与外部组件集成

  支持与多种外部组件集成，包括：Kafka、HDFS、HBase、Redis或JDBC/RDBMS等服务，便于实现涉及多种数据源的业务。

1.4.25.2 Storm 与其他组件的关系

Storm，提供实时的分布式计算框架，它可以从数据源（如Kafka、TCP连接等）中获得实时消息数据，在实时平台上完成高吞吐、低延迟的实时计算，并将结果输出到消息队列或者进行持久化。Storm与其他组件的关系如**图1-128**所示：

图 1-128 组件关系图



Storm 和 Streaming 的关系

Storm和Streaming都使用的开源Apache Storm内核，不同的是，Storm使用的内核版本是1.2.1，Streaming使用的是0.10.0。Streaming组件一般用来在升级场景继承过度业务，比如之前版本已经部署Streaming并且有业务在运行的情况下，升级后仍然可以使用Streaming。如果是新搭建的集群，则建议使用Storm。

Storm 1.2.1新增特性说明：

- **分布式缓存**：提供命令行工具共享和更新拓扑的所需要的外部资源（配置），无需重新打包和部署拓扑。
- **Native Streaming Window API**：提供基于窗口的API。
- **资源调度器**：新增基于资源的调度器插件，可以在拓扑定义时指定可使用的最大资源，并且通过配置的方式指定用户的资源配额，从而管理该用户名下的拓扑资源。
- **State Management**：提供带检查点机制的Bolt接口，当事件失败时，Storm会自动管理bolt的状态并且执行恢复。
- **消息采样和调试**：在Storm UI界面可以开关拓扑或者组件级别的调试，将流消息按采样比率输出到指定日志中。
- **Worker动态分析**：在Storm UI界面可以收集Wokrer进程的Jstack、Heap日志，并且可以重启Worker进程。
- **拓扑日志级别动态调整**：提供命令行和Storm UI两种方式对运行中的拓扑日志进行动态修改。
- **性能提升**：与之前的版本相比，Storm的性能得到了显著提升。虽然，拓扑的性能和用例场景及外部服务的依赖有很大的关系，但是对于大多数场景来说，性能可以提升3倍。

1.4.25.3 Storm 开源增强特性

- CQL

  CQL（Continuous Query Language），持续查询语言，是一种用于实时数据流上的查询语言，它是一种SQL-like的语言，相对于SQL，CQL中增加了（时序）窗口的概念，将待处理的数据保存在内存中，进行快速的内存计算，CQL的输出结果为数据流在某一时刻的计算结果。使用CQL，可以快速进行业务开发，并方便地将业务提交到Storm平台开启实时数据的接收、处理及结果输出；并可以在合适的时候中止业务。

- 高可用性

  Nimbus HA机制，避免了开源Storm集群中Nimbus出现单点故障而导致集群无法提供Topology的新增及管理操作的问题，增强了集群可用性。

1.4.26 Tez

Tez是Apache最新的支持DAG（有向无环图）作业的开源计算框架，它可以将多个有依赖的作业转换为一个作业从而大幅提升DAG作业的性能。如果 Hive这样的项目使用Tez而不是MapReduce作为其数据处理的骨干，那么将会显著提升它们的响应时间，Tez构建在YARN之上，能够不需要做任何改动地运行MapReduse任务。

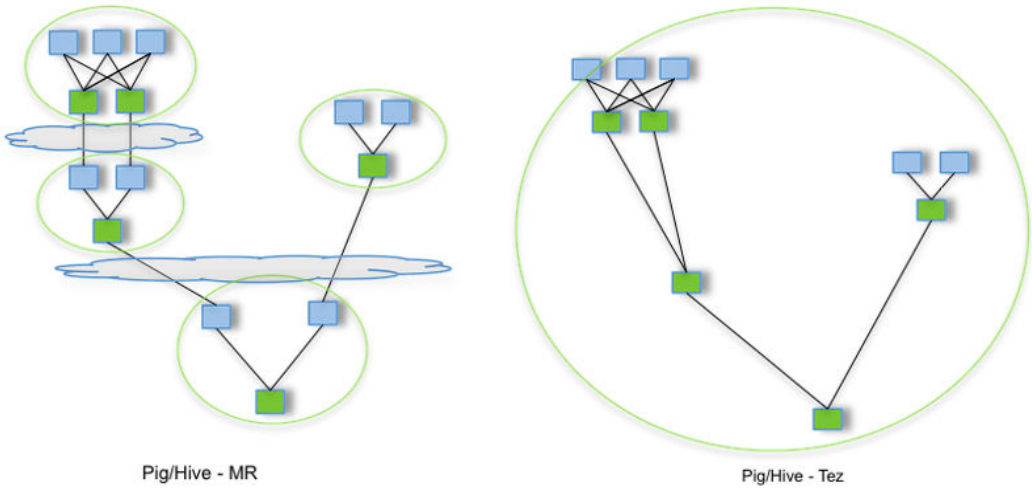
MRS将Tez作为Hive的默认执行引擎，执行效率远远超过原先的MapReduce的计算引擎。

有关Tez的详细说明，请参见：**https://tez.apache.org/**。

Tez 和 MapReduce 间的关系

Tez采用了DAG来组织MapReduce任务（DAG中一个节点就是一个RDD，边表示对RDD的操作）。它的核心思想是把将Map任务和Reduce任务进一步拆分，Map任务拆分为Input-Processor-Sort-Merge-Output，Reduce任务拆分为Input-Shuffer-Sort-Merge-Process-output，Tez将若干小任务灵活重组，形成一个大的DAG作业。

图 1-129 Hive 基于 MapReduce 提交任务和基于 Tez 提价任务流程图



Hive on MR任务中包含多个MapReduce任务，每个任务都会将中间结果存储到HDFS上——前一个步骤中的reducer为下一个步骤中的mapper提供数据。Hive on Tez任务仅在一个任务中就能完成同样的处理过程，任务之间不需要访问HDFS。

Tez 和 Yarn 间的关系

Tez是运行在Yarn之上的计算框架，运行时环境由Yarn的ResourceManager和ApplicationMaster组成。其中ResourceManager是一个全新的资源管理系统，而ApplicationMaster则负责MapReduce作业的数据切分、任务划分、资源申请和任务调度与容错等工作。此外，TezUI依赖Yarn提供的TimelineServer实现Tez任务运行过程呈现。

1.4.27 YARN

1.4.27.1 YARN 基本原理

为了实现一个Hadoop集群的集群共享、可伸缩性和可靠性，并消除早期MapReduce框架中的JobTracker性能瓶颈，开源社区引入了统一的资源管理框架**YARN**。

YARN是将JobTracker的两个主要功能（资源管理和作业调度/监控）分离，主要方法是创建一个全局的ResourceManager（RM）和若干个针对应用程序的ApplicationMaster（AM）。

说明

应用程序是指传统的MapReduce作业或作业的DAG（有向无环图）。

YARN 结构

YARN分层结构的本质是ResourceManager。这个实体控制整个集群并管理应用程序向基础计算资源的分配。ResourceManager将各个资源部分（计算、内存、带宽等）精心安排给基础NodeManager（YARN的每节点代理）。ResourceManager还与Application Master一起分配资源，与NodeManager一起启动和监视它们的基础应用程序。在此上下文中，Application Master承担了以前的TaskTracker的一些角色，ResourceManager 承担了JobTracker的角色。

Application Master管理一个在YARN内运行的应用程序的每个实例。Application Master负责协调来自ResourceManager的资源，并通过NodeManager监视容器的执行和资源使用（CPU、内存等的资源分配）。

NodeManager管理一个YARN集群中的每个节点。NodeManager提供针对集群中每个节点的服务，从监督对一个容器的终生管理到监视资源和跟踪节点健康。MRv1通过插槽管理Map和Reduce任务的执行，而NodeManager管理抽象容器，这些容器代表着可供一个特定应用程序使用的针对每个节点的资源。

图 1-130 YARN 结构

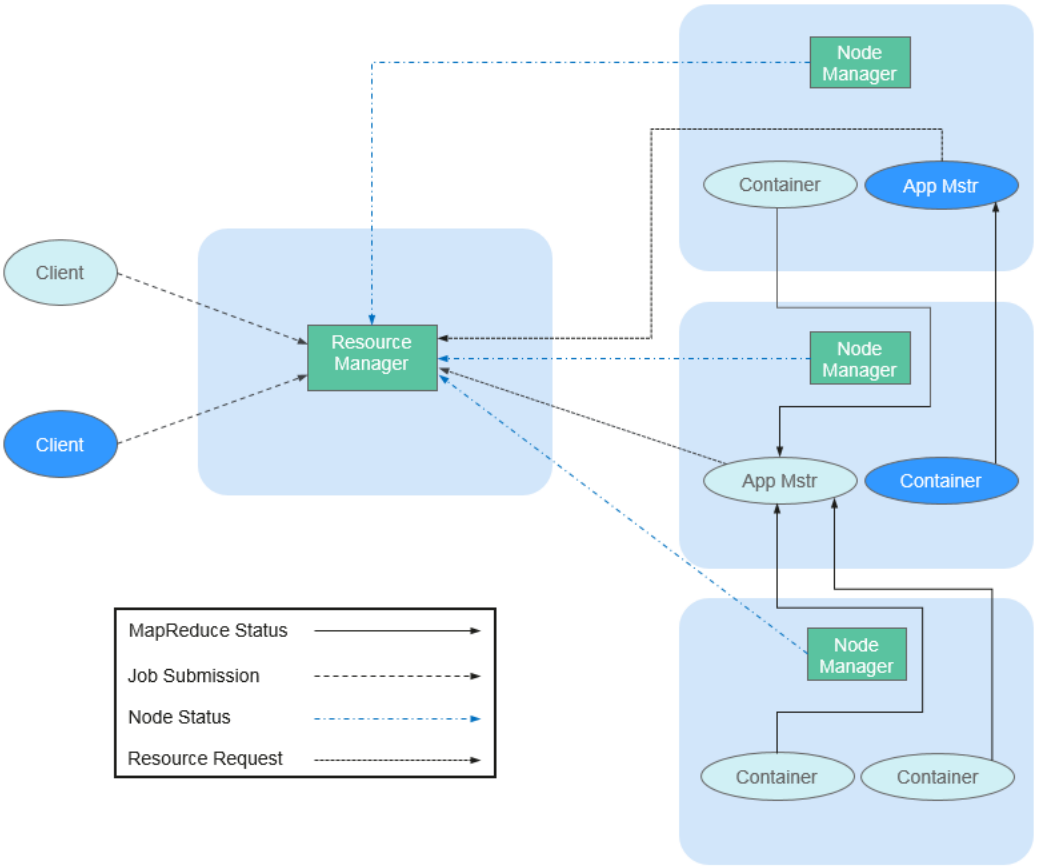


图1-130中各部分的功能如表1-25所示。

表 1-25 结构图说明

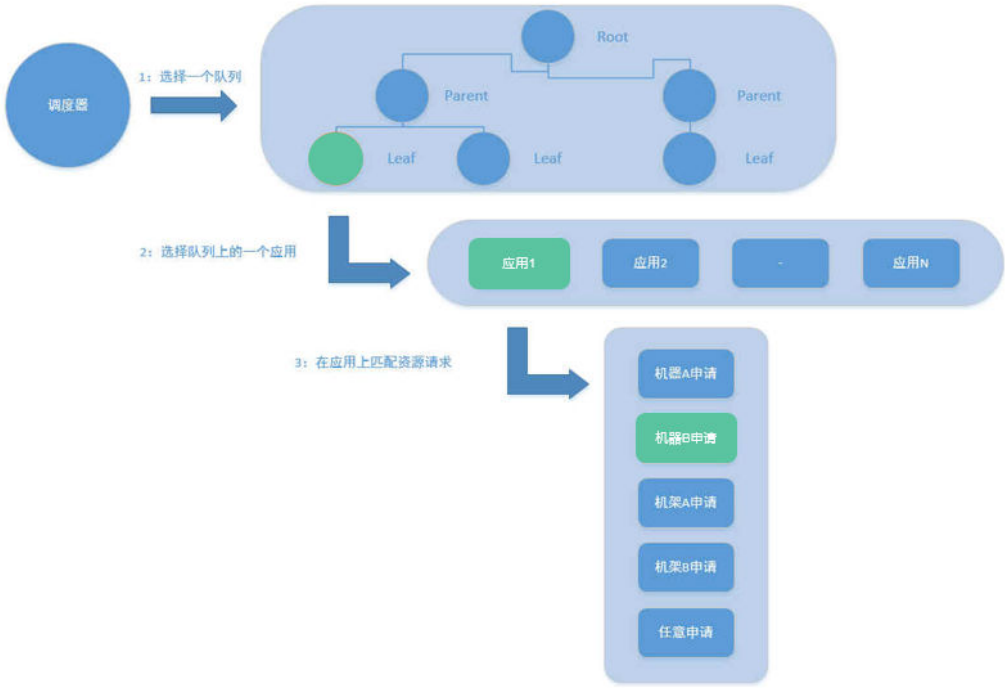
名称	描述
Client	YARN Application客户端，用户可以通过客户端向ResourceManager提交任务，查询Application运行状态等。
ResourceManager(RM)	负责集群中所有资源的统一管理和分配。接收来自各个节点（NodeManager）的资源汇报信息，并根据收集的资源按照一定的策略分配给各个应用程序。
NodeManager(NM)	NodeManager（NM）是YARN中每个节点上的代理，管理Hadoop集群中单个计算节点，包括与ResourceManger保持通信，监督Container的生命周期管理，监控每个Container的资源使用（内存、CPU等）情况，追踪节点健康状况，管理日志和不同应用程序用到的附属服务（auxiliary service）。
ApplicationMaster(AM)	即图中的App Mstr，负责一个Application生命周期内的所有工作。包括：与RM调度器协商以获取资源；将得到的资源进一步分配给内部任务（资源的二次分配）；与NM通信以启动/停止任务；监控所有任务运行状态，并在任务运行失败时重新为任务申请资源以重启任务。

名称	描述
Container	Container是YARN中的资源抽象，封装了某个节点上的多维度资源，如内存、CPU、磁盘、网络等（目前仅封装内存和CPU），当AM向RM申请资源时，RM为AM返回的资源便是用Container表示。YARN会为每个任务分配一个Container，且该任务只能使用该Container中描述的资源。

在YARN中，资源调度器是以层级队列方式组织资源的，这种组织方式有利于资源在不同队列间分配和共享，进而提高集群资源利用率。如下图所示，Superior Scheduler和Capacity Scheduler的核心资源分配模型相同。

调度器会维护队列的信息。用户可以向一个或者多个队列提交应用。每次NM心跳的时候，调度器会根据一定规则选择一个队列，再选择队列上的一个应用，并尝试在这个应用上分配资源。若因参数限制导致分配失败，将选择下一个应用。选择一个应用后，调度器会处理此应用的资源申请。其优先级从高到低依次为：本地资源的申请、同机架的申请，任意机器的申请。

图 1-131 资源分配模型



YARN 原理

新的Hadoop MapReduce框架被命名为MRv2或YARN。YARN主要包括ResourceManager、ApplicationMaster与NodeManager三个部分。

- ResourceManager：RM是一个全局的资源管理器，负责整个系统的资源管理和分配。主要由两个组件构成：调度器（Scheduler）和应用程序管理器（Applications Manager）。
  - 调度器根据容量、队列等限制条件（如每个队列分配一定的资源，最多执行一定数量的作业等），将系统中的资源分配给各个正在运行的应用程序。调

度器仅根据各个应用程序的资源需求进行资源分配，而资源分配单位用一个抽象概念Container表示。Container是一个动态资源分配单位，将内存、CPU、磁盘、网络等资源封装在一起，从而限定每个任务使用的资源量。此外，该调度器是一个可插拔的组件，用户可根据自己的需要设计新的调度器，YARN提供了多种直接可用的调度器，比如Fair Scheduler和Capacity Scheduler等。

- 应用程序管理器负责管理整个系统中所有应用程序，包括应用程序提交、与调度器协商资源以启动ApplicationMaster、监控ApplicationMaster运行状态并在失败时重新启动等。

- NodeManager：NM是每个节点上的资源和任务管理器，一方面，会定时向RM汇报本节点上的资源使用情况和各个Container的运行状态；另一方面，接收并处理来自AM的Container启动/停止等请求。
- ApplicationMaster：AM负责一个Application生命周期内的所有工作。包括：
  - 与RM调度器协商以获取资源。
  - 将得到的资源进一步分配给内部的任务(资源的二次分配)。
  - 与NM通信以启动/停止任务。
  - 监控所有任务运行状态，并在任务运行失败时重新为任务申请资源以重启任务。

开源容量调度器 Capacity Scheduler 原理

Capacity Scheduler是一种多用户调度器，它以队列为单位划分资源，为每个队列设定了资源最低保证和使用上限。同时，也为每个用户设定了资源使用上限以防止资源滥用。而当一个队列的资源有剩余时，可暂时将剩余资源共享给其他队列。

Capacity Scheduler支持多个队列，为每个队列配置一定的资源量，并采用FIFO调度策略。为防止同一用户的应用独占队列资源，Capacity Scheduler会对同一用户提交的作业所占资源量进行限定。调度时，首先计算每个队列使用的资源，选择使用资源最少的队列；然后按照作业优先级和提交时间顺序选择，同时考虑用户资源量的限制和内存限制。Capacity Scheduler主要有如下特性：

- 容量保证。MRS集群管理员可为每个队列设置资源最低保证和资源使用上限，而所有提交到队列的应用程序共享这些资源。
- 灵活性。如果一个队列中的资源有剩余，可以暂时共享给那些需要资源的队列，而一旦该队列有新的应用程序提交，则占用资源的队列将资源释放给该队列。这种资源灵活分配的方式可明显提高资源利用率。
- 多重租赁。支持多用户共享集群和多应用程序同时运行。为防止单个应用程序、用户或者队列独占集群中的资源，MRS集群管理员可为之增加多重约束（比如单个应用程序同时运行的任务数等）。
- 安全保证。每个队列有严格的ACL列表规定它的访问用户，每个用户可指定哪些用户允许查看自己应用程序的运行状态或者控制应用程序。此外，MRS集群管理员可指定队列管理员和集群系统管理员。
- 动态更新配置文件。MRS集群管理员可根据需要动态修改配置参数以实现在线集群管理

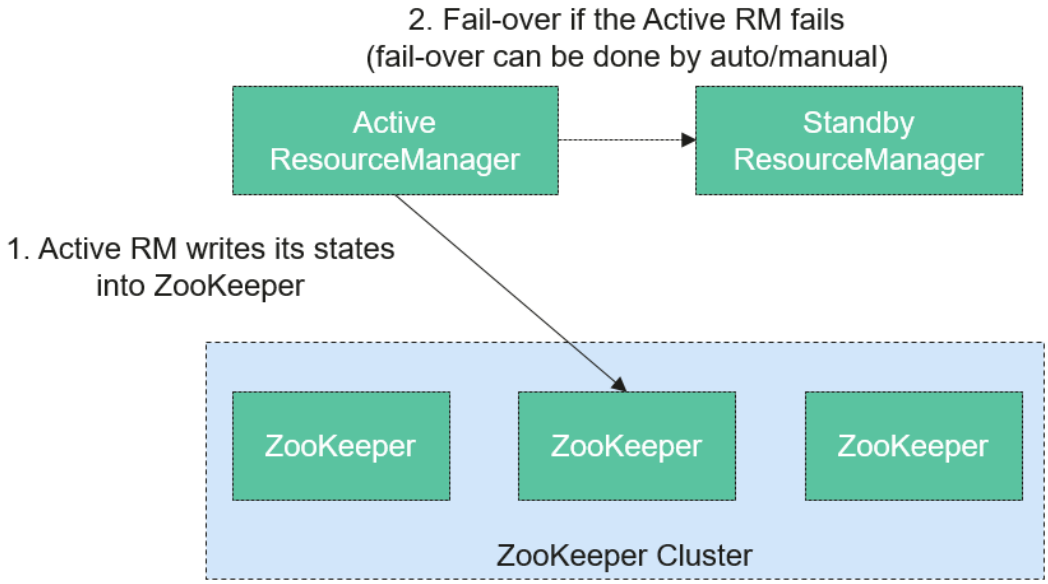
Capacity Scheduler中每个队列可以限制资源使用量。队列间的资源分配以使用量作为排列依据，使得容量小的队列有竞争优势。集群整体吞吐较大，延迟调度机制使得应用可以有机会放弃跨机器或者跨机架的调度，争取本地调度。

1.4.27.2 YARN HA 方案介绍

YARN HA 原理与实现方案

YARN中的ResourceManager负责整个集群的资源管理和任务调度，在Hadoop2.4版本之前，ResourceManager在YARN集群中存在单点故障的问题。YARN高可用性方案通过引入冗余的ResourceManager节点的方式，解决了这个基础服务的可靠性和容错性问题。

图 1-132 ResourceManager 高可用性实现架构



ResourceManager的高可用性方案是通过设置一组Active/Standby的ResourceManager节点来实现的（如图1-132）。与HDFS的高可用性方案类似，任何时间点上都只能有一个ResourceManager处于Active状态。当Active状态的ResourceManager发生故障时，可通过自动或手动的方式触发故障转移，进行Active/Standby状态切换。

在未开启自动故障转移时，YARN集群启动后，MRS集群管理员需要在命令行中使用***yarn rmadmin***命令手动将其中一个ResourceManager切换为Active状态。当需要执行计划性维护或故障发生时，则需要先手动将Active状态的ResourceManager切换为Standby状态，再将另一个ResourceManager切换为Active状态。

开启自动故障转移后，ResourceManager会通过内置的基于ZooKeeper实现的ActiveStandbyElector来决定哪一个ResouceManager应该成为Active节点。当Active状态的ResourceManager发生故障时，另一个ResourceManager将自动被选举为Active状态以接替故障节点。

当集群的ResourceManager以HA方式部署时，客户端使用的“yarn-site.xml”需要配置所有ResourceManager地址。客户端（包括ApplicationMaster和NodeManager）会以轮询的方式寻找Active状态的ResourceManager，也就是说客户端需要自己提供容错机制。如果当前Active状态的ResourceManager无法连接，那么会继续使用轮询的方式找到新的ResourceManager。

备RM升主后，能够恢复故障发生时上层应用运行的状态（详见**ResourceManger Restart**）。当启用ResourceManager Restart时，重启后的ResourceManager就可以通过加载之前Active的ResourceManager的状态信息，并通过接收所有NodeManager

上container的状态信息重构运行状态继续执行。这样应用程序通过定期执行检查点操作保存当前状态信息，就可以避免工作内容的丢失。状态信息需要让Active/Standby的ResourceManager都能访问。当前系统提供了三种共享状态信息的方法：通过文件系统共享（FileSystemRMStateStore）、通过LevelDB数据库共享（LeveldbRMStateStore）或通过ZooKeeper共享（ZKRMStateStore）。这三种方式中只有ZooKeeper共享支持Fencing机制。Hadoop默认使用ZooKeeper共享。

1.4.27.3 Yarn 与其他组件的关系

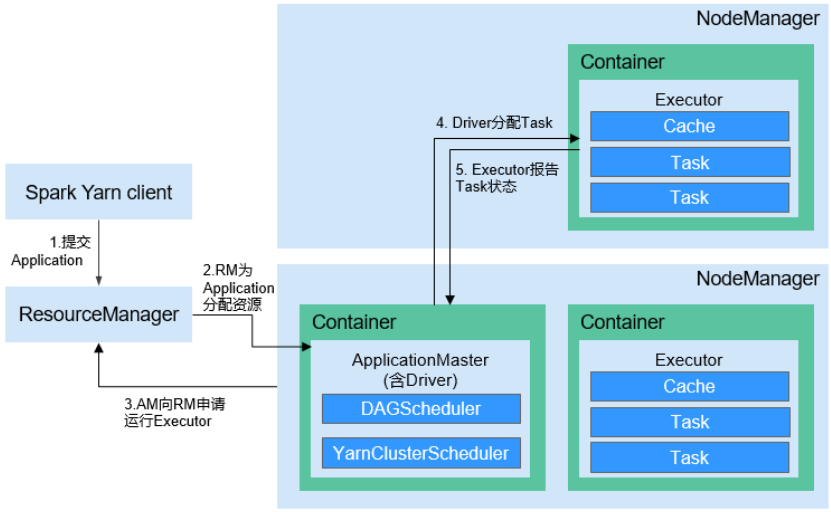
Yarn 和 Spark 组件的关系

Spark的计算调度方式，可以通过Yarn的模式实现。Spark共享Yarn集群提供丰富的计算资源，将任务分布式的运行起来。Spark on Yarn分两种模式：Yarn Cluster和Yarn Client。

- Yarn Cluster模式

  运行框架如图1-133所示。

  图 1-133 Spark on yarn-cluster 运行框架

  Spark on yarn-cluster实现流程：

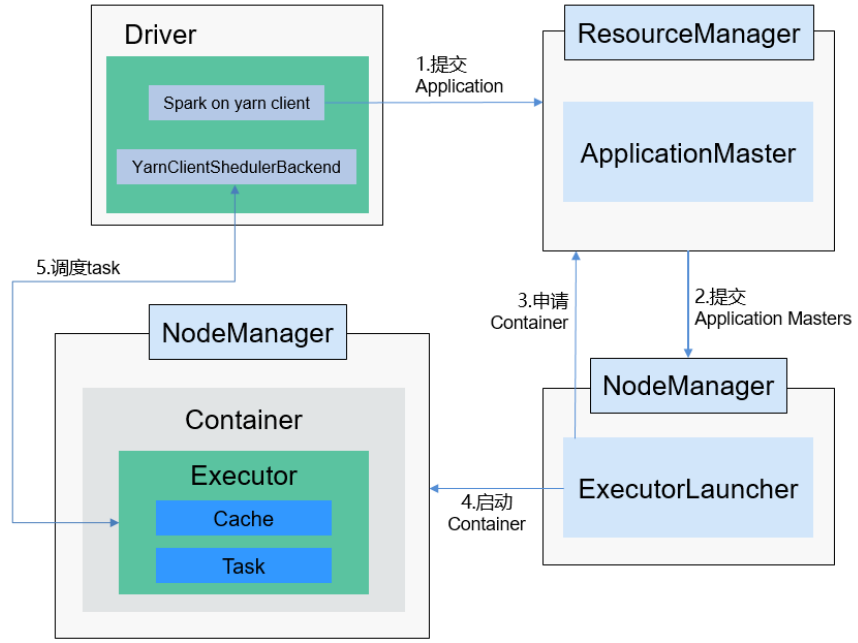
  a. 首先由客户端生成Application信息，提交给ResourceManager。
  b. ResourceManager为Spark Application分配第一个Container(ApplicationMaster)，并在该Container上启动Driver。
  c. ApplicationMaster向ResourceManager申请资源以运行Container。

     ResourceManager分配Container给ApplicationMaster，ApplicationMaster和相关的NodeManager通讯，在获得的Container上启动Executor，Executor启动后，开始向Driver注册并申请Task。
  d. Driver分配Task给Executor执行。
  e. Executor执行Task并向Driver汇报运行状况。

- Yarn Client模式

  运行框架如图1-134所示。

图 1-134 Spark on yarn-client 运行框架



Spark on yarn-client实现流程：

说明

在yarn-client模式下，Driver部署在Client端，在Client端启动。yarn-client模式下，不兼容老版本的客户端。推荐使用yarn-cluster模式。

a. 客户端向ResourceManager发送Spark应用提交请求，ResourceManager为其返回应答，该应答中包含多种信息(如ApplicationId、可用资源使用上限和下限等)。Client端将启动ApplicationMaster所需的所有信息打包，提交给ResourceManager上。

b. ResourceManager收到请求后，会为ApplicationMaster寻找合适的节点，并在该节点上启动它。ApplicationMaster是Yarn中的角色，在Spark中进程名字是ExecutorLauncher。

c. 根据每个任务的资源需求，ApplicationMaster可向ResourceManager申请一系列用于运行任务的Container。

d. 当ApplicationMaster（从ResourceManager端）收到新分配的Container列表后，会向对应的NodeManager发送信息以启动Container。

   ResourceManager分配Container给ApplicationMaster，ApplicationMaster和相关的NodeManager通讯，在获得的Container上启动Executor，Executor启动后，开始向Driver注册并申请Task。

   说明

   正在运行的container不会被挂起释放资源。

e. Driver分配Task给Executor执行。Executor执行Task并向Driver汇报运行状况。

Yarn 和 MapReduce 的关系

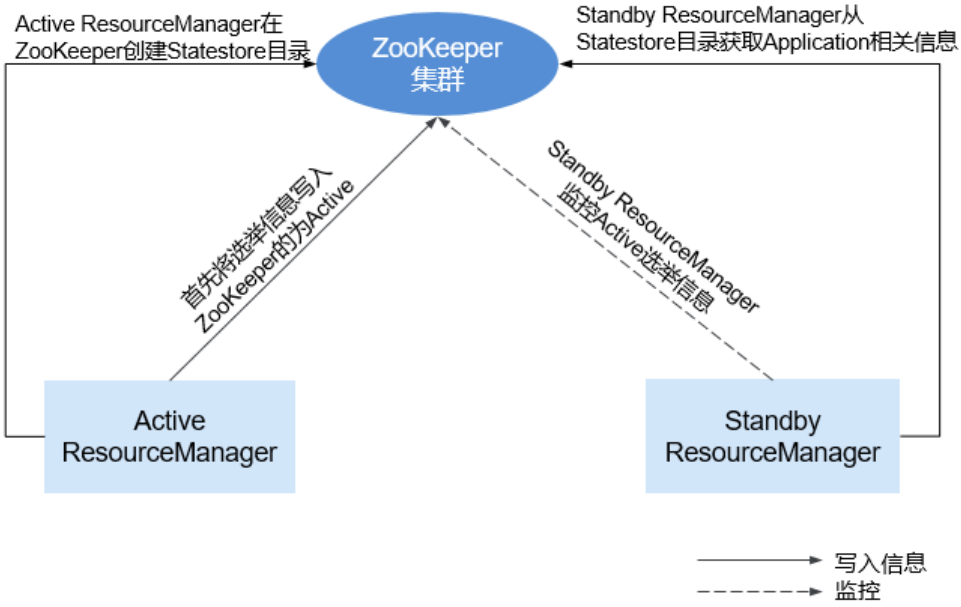
MapReduce是运行在Yarn之上的一个批处理的计算框架。MRv1是Hadoop 1.0中的MapReduce实现，它由编程模型（新旧编程接口）、运行时环境（由JobTracker和

TaskTracker组成）和数据处理引擎（MapTask和ReduceTask）三部分组成。该框架在扩展性、容错性（JobTracker单点）和多框架支持（仅支持MapReduce一种计算框架）等方面存在不足。MRv2是Hadoop 2.0中的MapReduce实现，它在源码级重用了MRv1的编程模型和数据处理引擎实现，但运行时环境由Yarn的ResourceManager和ApplicationMaster组成。其中ResourceManager是一个全新的资源管理系统，而ApplicationMaster则负责MapReduce作业的数据切分、任务划分、资源申请和任务调度与容错等工作。

Yarn 和 ZooKeeper 的关系

ZooKeeper与Yarn的关系如图1-135所示。

图 1-135 ZooKeeper 与 Yarn 的关系



1. 在系统启动时，ResourceManager会尝试把选举信息写入ZooKeeper，第一个成功写入ZooKeeper的ResourceManager被选举为Active ResourceManager，另一个为Standby ResourceManager。Standby ResourceManager定时去ZooKeeper监控Active ResourceManager选举信息。
2. Active ResourceManager还会在ZooKeeper中创建Statestore目录，存储Application相关信息。当Active ResourceManager产生故障时，Standby ResourceManager会从Statestore目录获取Application相关信息，恢复数据。

Yarn 和 Tez 的关系

Hive on Tez作业信息需要Yarn提供TimeLine Server能力，以支持Hive任务展示应用程序的当前和历史状态，便于存储和检索。

1.4.27.4 YARN 开源增强特性

任务优先级调度

在原生的YARN资源调度机制中，如果先提交的MapReduce Job长时间地占据整个Hadoop集群的资源，会使得后提交的Job一直处于等待状态，直到Running中的Job执行完并释放资源。

MRS集群提供了任务优先级调度机制。此机制允许用户定义不同优先级的Job，后启动的高优先级Job能够获取运行中的低优先级Job释放的资源；低优先级Job未启动的计算容器被挂起，直到高优先级Job完成并释放资源后，才被继续启动。

该特性使得业务能够更加灵活地控制自己的计算任务，从而达到更佳的集群资源利用率。

说明

容器可重用与任务优先级调度有冲突，若启用容器重用，资源会被持续占用，优先级调度将不起作用。

YARN 的权限控制

Hadoop YARN的权限机制是通过访问控制列表（ACL）实现的。按照不同用户授予不同权限控制，主要介绍下面两个部分：

- 集群运维管理员控制列表（Admin Acl）

  该功能主要用于指定YARN集群的运维管理员，其中，管理员列表由参数“yarn.admin.acl”指定。集群运维管理员可以访问ResourceManager WebUI，还能操作NodeManager节点、队列、NodeLabel等，**但不能提交任务**。

- 队列访问控制列表（Queue Acl）

  为了方便管理集群中的用户，YARN将用户/用户组分成若干队列，并指定每个用户/用户组所属的队列。每个队列包含两种权限：提交应用程序权限和管理应用程序权限（比如终止任意应用程序）。

开源功能：

虽然目前YARN服务的用户层面上支持如下三种角色：

- 集群运维管理员
- 队列管理员
- 普通用户

但是当前开源YARN提供的WebUI/RestAPI/JavaAPI等接口上不会根据用户角色进行权限控制，任何用户都有权限访问应用和集群的信息，无法满足多租户场景下的隔离要求。

增强：

安全模式下，对开源YARN提供的WebUI/RestAPI/JavaAPI等接口上进行了权限管理上的增强，支持根据不同的用户角色，进行相应的权限控制。

各个角色对应的权限如下：

- 集群运维管理员：拥有在YARN集群上执行管理操作（如访问ResourceManager WebUI、刷新队列、设置NodeLabel、主备倒换等）的权限。

- 队列管理员：拥有在YARN集群上所管理队列的修改和查看权限。
- 普通用户：拥有在YARN集群上对自己提交应用的修改和查看权限。

自研超级调度器 Superior Scheduler 原理

Superior Scheduler是一个专门为Hadoop YARN分布式资源管理系统设计的调度引擎，是针对企业客户融合资源池，多租户的业务诉求而设计的高性能企业级调度器。

Superior Scheduler可实现开源调度器、Fair Scheduler以及Capacity Scheduler的所有功能。另外，相较于开源调度器，Superior Scheduler在企业级多租户调度策略、租户内多用户资源隔离和共享、调度性能、系统资源利用率和支持大集群扩展性方面都做了针对性的增强。设计的目标是让Superior Scheduler直接替代开源调度器。

类似于开源Fair Scheduler和Capacity Scheduler，Superior Scheduler通过YARN调度器插件接口与YARN Resource Manager组件进行交互，以提供资源调度功能。图1-136为其整体系统图。

图 1-136 Superior Scheduler 内部架构

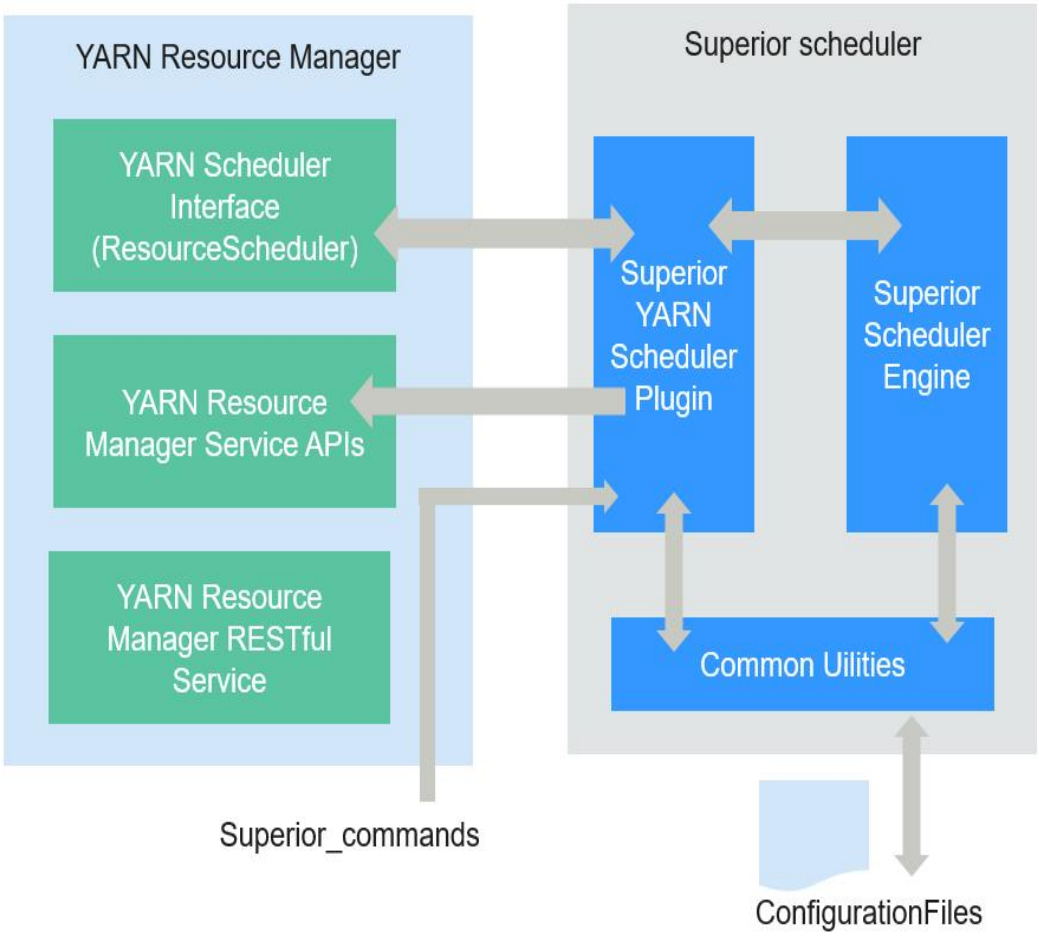


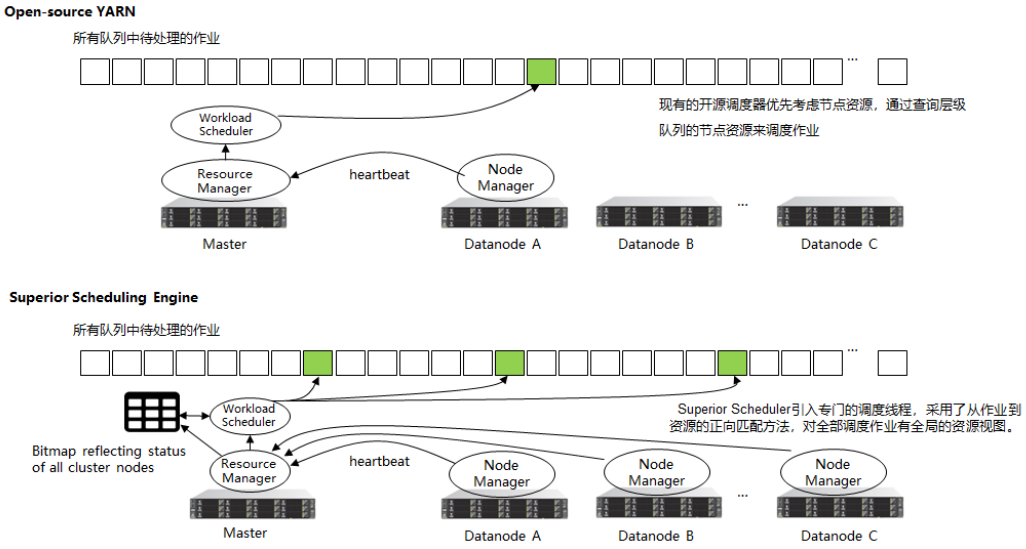
图1-136中，Superior Scheduler的主要模块如下：

- Superior Scheduler Engine：具有丰富调度策略的高性能调度器引擎。
- Superior YARN Scheduler Plugin：YARN Resource Manager和Superior Scheduler Engine之间的桥梁，负责同YARN Resource Manager交互。

在调度原理上，开源的调度器都是基于计算节点心跳驱动的资源反向匹配作业的调度机制。具体来讲，每个计算节点定期发送心跳到YARN的Resource Manager通知该节点状态并同时启动调度器为这个节点分配作业。这种调度机制把调度的周期同心跳结合在一起，当集群规模增大时，会遇到系统扩展性以及调度性能瓶颈。另外，因为采用了资源反向匹配作业的调度机制，开源调度器在调度精度上也有局限性，例如数据亲和性偏于随机，另外系统也无法支持基于负载的调度策略等。主要原因是调度器在选择作业时，缺乏全局的资源视图，很难做到最优选择。

Superior Scheduler内部采用了不同的调度机制。Superior Scheduler的调度器引入了专门的调度线程，把调度同心跳剥离开，避免了系统心跳风暴问题。另外，Superior Scheduler调度流程采用了从作业到资源的正向匹配方法，这样每个调度的作业都有全局的资源视图，可以很大的提到调度的精度。相比开源调度器，Superior Scheduler在系统吞吐量、利用率、数据亲和性等方面都有很大提升。

图 1-137 Superior Scheduler 性能对比



Superior Scheduler除了提高系统吞吐量和利用率，还提供了以下主要调度功能：

- 多资源池

  多资源池有助于在逻辑上划分集群资源并在多个租户/队列之间共享它们。资源池的划分可以基于异构的资源或完全按照应用资源隔离的诉求来划分。对于一个资源池，不同队列可配置进一步的策略。

- 每个资源池多租户调度（reserve、min、share、max）

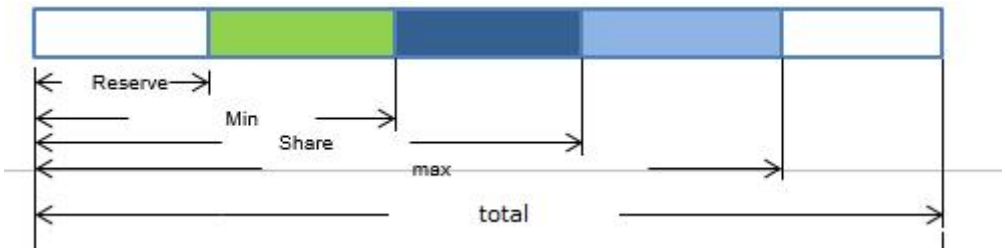
  Superior Scheduler提供了灵活的层级多租户调度策略。并允许针对不同的资源池可以访问的租户/队列，配置不同策略，如下所示。

表 1-26 策略描述

策略名称	描述
reserve	预留租户资源。即使租户没有作业，其他租户也不能使用该预留的资源。其值可以是百分比或绝对值。如果两者都配置，调度系统动态计算转换为资源绝对值，并取两者的最大值。缺省的reserve值为0。相对于定义一个专用资源池并指定具体机器的方式，reserve的策略可以认为提供了一种灵活的浮动预留功能，由于并不限定具体的机器，可以提高计算的数据亲和性，也不会受具体机器故障的影响。
min	具有抢占支持的最低保证资源。其他租户可以使用这部分资源，但是本租户享有优先使用权。其值可以是百分比或绝对值。如果两者都配置，调度系统动态计算转换为资源绝对值，并取两者的最大值。缺省值是0。
share	不支持抢占的共享资源。本租户要使用这部分资源时，需要等待其他租户完成作业并释放资源。其值是百分比或绝对值。
max	允许的最大资源数量。租户无法获得比允许的最大资源多的资源。其值是百分比或绝对值。如果两者都配置，调度系统动态计算转换为资源绝对值，并取两者最大值。缺省值不受限制。

租户资源分配策略示意图，如图1-138所示。

图 1-138 策略示意图



说明

其中“total”表示总资源，不是调度策略。

同开源的调度器相比，Superior Scheduler同时提供了租户级百分比和绝对值的混配策略，可以很好的适应各种灵活的企业级租户资源调度诉求。例如，用户可以在一级租户提供最大绝对值的资源保障，这样租户的资源不会因为集群的规模改变而受影响。但在下层的子租户之间，可以提供百分比的分配策略，这样可以尽可能提升一级租户内的资源利用率。

- 异构和多维资源调度

  Superior Scheduler支持CPU和内存资源的调度外，还支持扩展支持以下功能：

  - 节点标签可用于识别像GPU_ENABLED，SSD_ENBALED等节点的多维属性，可以根据这些标签进行调度。
  - 资源池可用于对同一类别的资源进行分组并分配给特定的租户/队列。
- 租户内多用户公平调度

在叶子租户里，多个用户可以使用相同的队列来提交作业。相比开源调度器，Superior Scheduler可以支持在同一租户内灵活配置不同用户的资源共享策略。例如可以为VIP用户配置更多的资源访问权重。

- 数据位置感知调度

  Superior Scheduler采用“从作业到节点的调度策略”，即尝试在可用节点之间调度给定的作业，使得所选节点适合于给定作业。通过这样做，调度器将具有集群和数据的整体视图。如果有机会使任务更接近数据，则保证了本地化。而开源调度器采用“从节点到作业的调度策略”，在给定节点中尝试匹配适当的作业。

- Container调度时动态资源预留

  在异构和多样化的计算环境中，一些container需要更多的资源或多种资源，例如Spark作业可能需要更大的内存。当这些container与其他需要较小资源的container竞争时，可能没有机会在合理的时间内获得所需的资源而处于饥饿状态。由于开源的调度器是基于资源反向匹配作业的调度方式，会为这些作业盲目的进行资源预留以防进入饥饿状态。这就导致了系统资源的整体浪费。Superior Scheduler与开源特性的不同之处在于：

  - 基于需求的匹配：由于Superior Scheduler采用“从作业到节点的调度”，能够选择合适的节点来预留资源提升这些特殊container的启动时间，并避免浪费。
  - 租户重新平衡：启用预留逻辑时，开源调度器并不遵循配置的共享策略。Superior Scheduler采取不同的方法。在每个调度周期中，Superior Scheduler将遍历租户，并尝试基于多租户策略重新达到平衡，且尝试满足所有策略（reserve，min，share等），以便可以释放预留的资源，将可用资源流向不同租户下的其他本应得到资源的container。

- 动态队列状态控制（Open/Closed/Active/Inactive）

  支持多个队列状态，有助于管理员操作和维护多个租户。

  - Open状态（Open/Closed）：如果是Open（默认）状态，将接受提交到此队列的应用程序，如果是Closed状态，则不接受任何应用程序。
  - Active状态（Active/Inactive）：如果处于Active（默认）状态，租户内的应用程序是可以被调度和分配资源。如果处于Inactive状态则不会进行调度。

- 应用等待原因

  如果应用程序尚未启动，则提供作业等待原因信息。

Superior Scheduler和YARN开源调度器作了对比分析，如表1-27所示：

表 1-27 对比分析

领域	YARN开源调度器	Superior Scheduler
多租户调度	在同构集群上，只能选择容量调度器（Capacity Scheduler）或公平调度器（Fair Scheduler）两者之一，且集群当前不支持公平调度器（Fair Scheduler）。容量调度器只支持百分比方式配置，而公平调度器只支持绝对值方式。	• 支持**异构集群**和**多资源池**。 • 支持**预留**，以保证直接访问资源。

领域	YARN开源调度器	Superior Scheduler
数据位置感知调度	从节点到作业的调度策略导致降低数据本地话命中率，潜在影响应用的执行性能。	**从作业到节点的调度策略。**可具有更精确的数据位置感知，数据本地化调度的作业命中率比较高。
基于机器负载的均衡调度	不支持	**Superior Scheduler在调度时考虑机器的负载和资源分配情况，做到均衡调度。**
租户内多用户公平调度	不支持	租户内用户的公平调度，支持关键字default、others。
作业等待原因	不支持	作业等待原因信息可显示为什么作业需等待。

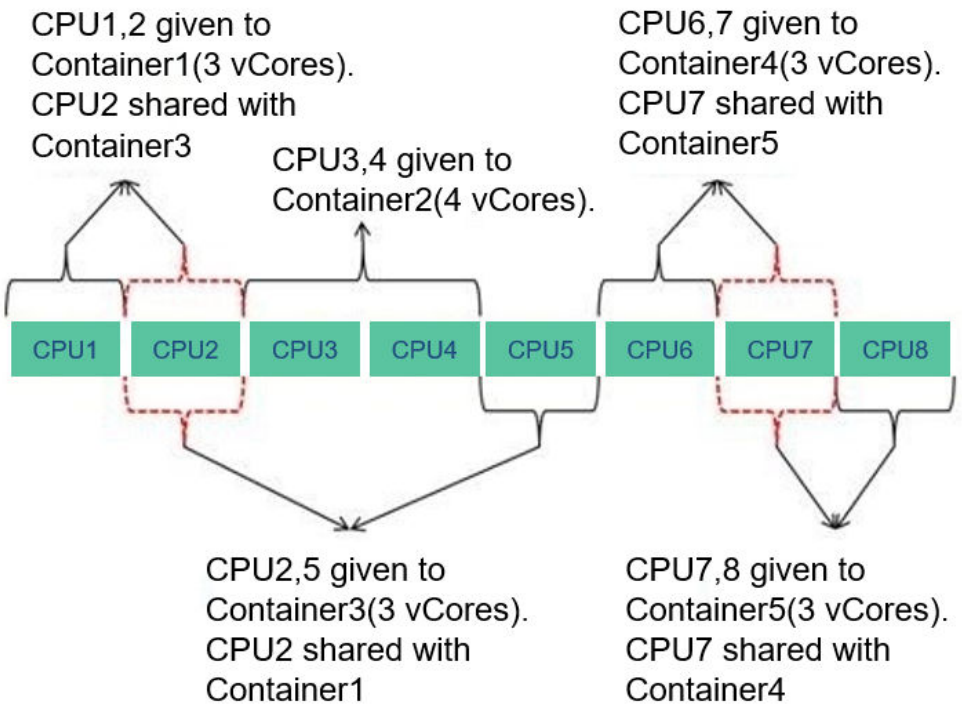
综上所述，Superior Scheduler是一个高性能调度器，拥有丰富的调度策略，在功能、性能、资源利用率和扩展性方面都优于Capacity Scheduler。

支持 CPU 硬隔离

YARN无法严格控制每个container使用的CPU资源。在使用CPU子系统时，container可能会超额占用资源。此时使用CPUset控制资源分配。

为了解决这个问题，CPU将会被严格按照虚拟核和物理核的比例分配至各个container。如果container需要一整个物理核，则分配给它一整个物理核。若container只需要部分物理核，则可能发生几个container共享同一个物理核的情况。下图为CPU配额示例，假定虚拟核和物理核的比例为2:1。

图 1-139 CPU 配额

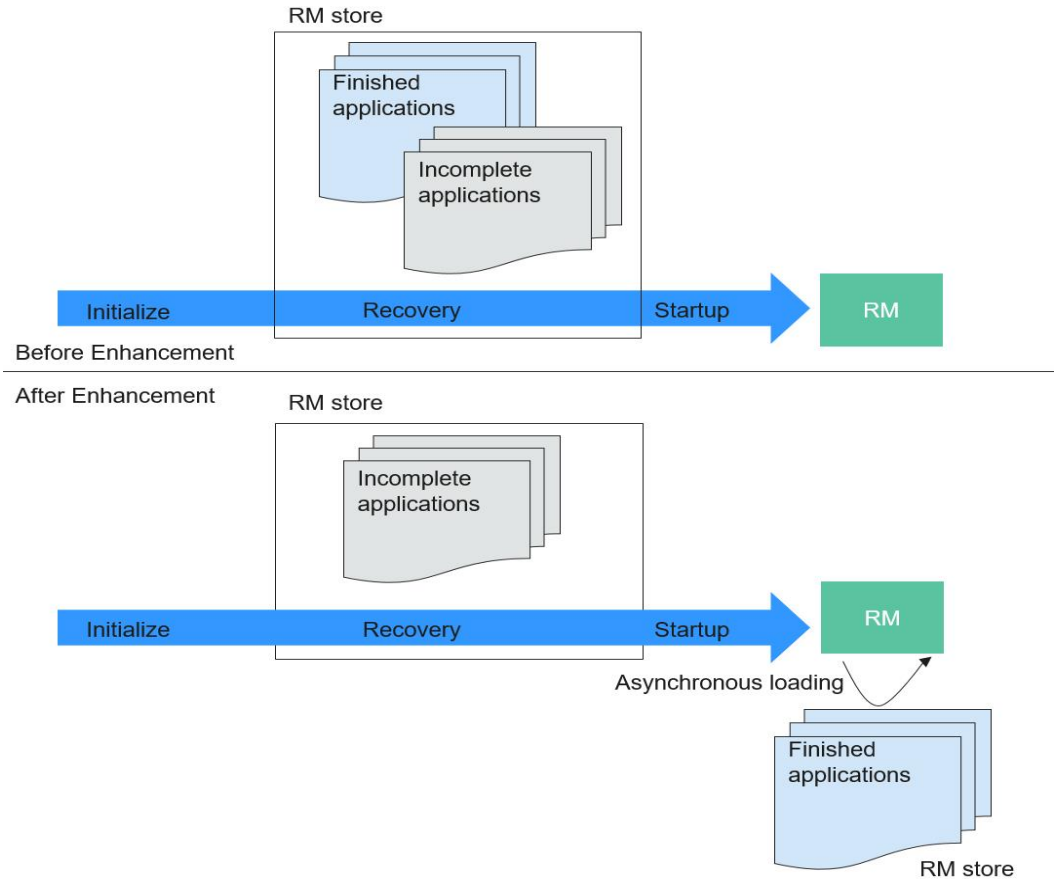


YARN 开源增强特性：重启性能优化

一般情况下，RM恢复会获取正在运行和已完成的应用。而大量的已完成的应用可能导致RM启动过慢、HA切换/重启耗时过长等问题。

为了加速RM的启动，现在优先获取未完成的应用列表，再启动RM。此时，已完成的应用会在一个后台异步线程中继续恢复。下图展示了RM的启动恢复流程。

图 1-140 RM 启动恢复流程



1.4.28 ZooKeeper

1.4.28.1 ZooKeeper 基本原理

ZooKeeper 简介

ZooKeeper是一个分布式、高可用性的协调服务。在大数据产品中主要提供两个功能：

- 帮助系统避免单点故障，建立可靠的应用程序。
- 提供分布式协作服务和维护配置信息。

ZooKeeper 结构

ZooKeeper集群中的节点分为三种角色：Leader、Follower和Observer，其结构和相互关系如图1-141所示。通常来说，需要在集群中配置奇数个（2N+1）ZooKeeper服务，至少（N+1）个投票才能成功的执行写操作。

图 1-141 ZooKeeper 结构

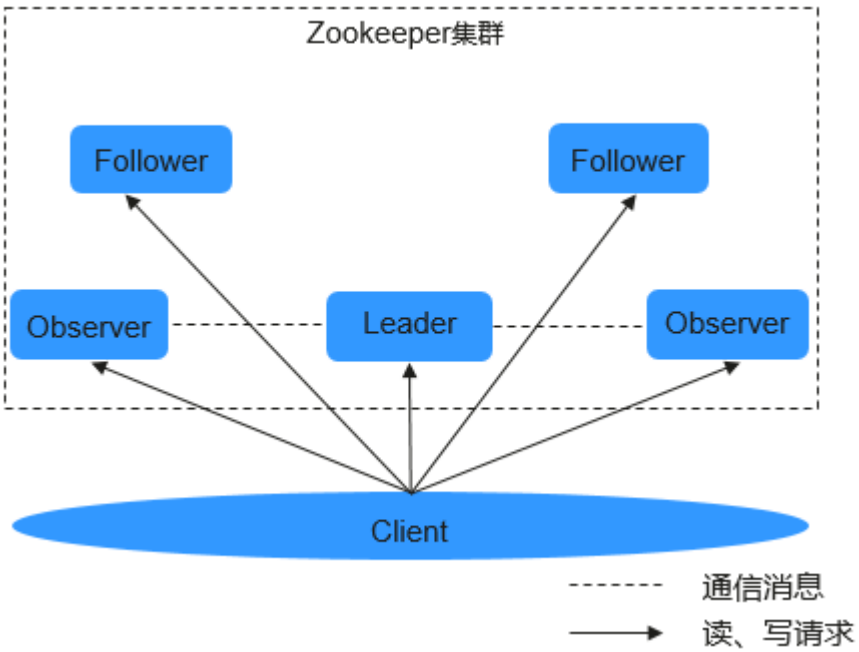


图1-141中各部分的功能说明如表1-28所示。

表 1-28 结构图说明

名称	描述
Leader	在ZooKeeper集群中只有一个节点作为集群的Leader，由各Follower通过ZooKeeper Atomic Broadcast(ZAB)协议选举产生，主要负责接收和协调所有写请求，并把写入的信息同步到Follower和Observer。
Follower	Follower的功能有两个： • 每个Follower都作为Leader的储备，当Leader故障时重新选举Leader，避免单点故障。 • 处理读请求，并配合Leader一起进行写请求处理。
Observer	Observer不参与选举和写请求的投票，只负责处理读请求、并向Leader转发写请求，避免系统处理能力浪费。
Client	ZooKeeper集群的客户端，对ZooKeeper集群进行读写操作。例如HBase可以作为ZooKeeper集群的客户端，利用ZooKeeper集群的仲裁功能，控制其HMaster的“Active”和“Standby”状态。

如果集群启用了安全服务，在连接ZooKeeper时需要进行身份认证，认证方式有以下两种：

- keytab方式：需要从MRS集群管理员处获取一个“人机”用户，用于登录MRS平台并通过认证，并且获取到该用户的keytab文件。
- 票据方式：从MRS集群管理员处获取一个“人机”用户，用于后续的安全登录，开启Kerberos服务的renewable和forwardable开关并且设置票据刷新周期，开启成功后重启kerberos及相关组件。

  说明

  - 默认情况下，用户的密码有效期是90天，所以获取的keytab文件的有效期是90天。
  - Kerberos服务的renewable、forwardable开关和票据刷新周期的设置在Kerberos服务的配置页面的“系统”标签下，票据刷新周期的修改可以根据实际情况修改“kdc_renew_lifetime”和“kdc_max_renewable_life”的值。

ZooKeeper 原理

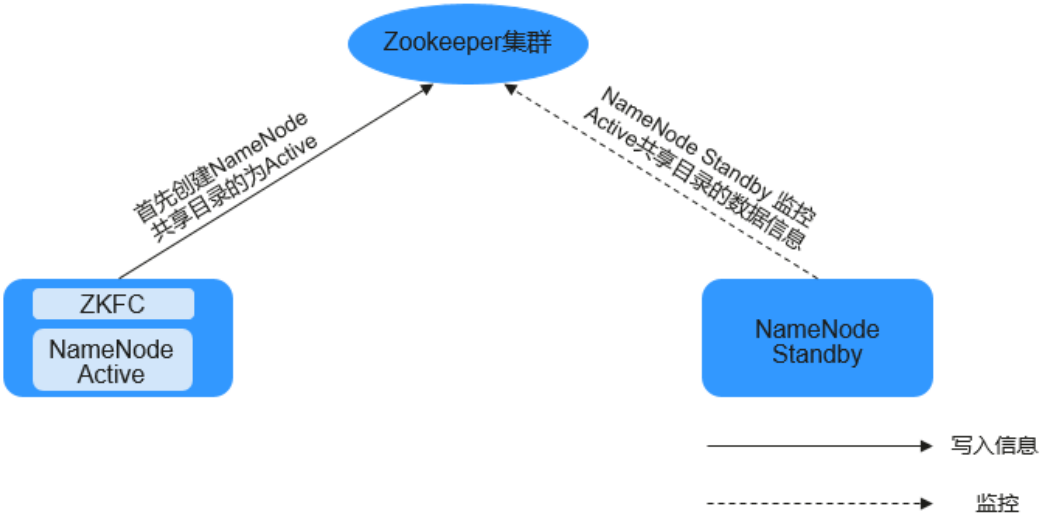
- **写请求**
  a. Follower或Observer接收到写请求后，转发给Leader。
  b. Leader协调各Follower，通过投票机制决定是否接受该写请求。
  c. 如果超过半数以上的Leader、Follower节点返回写入成功，那么Leader提交该请求并返回成功，否则返回失败。
  d. Follower或Observer返回写请求处理结果。
- **只读请求**
  客户端直接向Leader、Follower或Observer读取数据。

1.4.28.2 ZooKeeper 与其他组件的关系

ZooKeeper 和 HDFS 的关系

ZooKeeper与HDFS的关系如图1-142所示。

图 1-142 ZooKeeper 和 HDFS 的关系



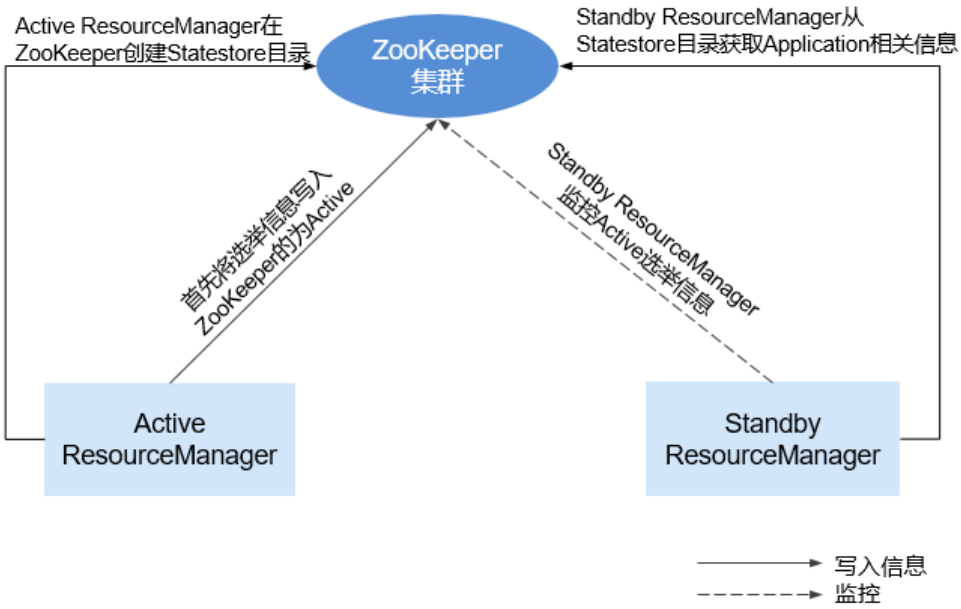
ZKFC（ZKFailoverController）作为一个ZooKeeper集群的客户端，用来监控NameNode的状态信息。ZKFC进程仅在部署了NameNode的节点中存在。HDFS NameNode的Active和Standby节点均部署有zkfc进程。

1. HDFS NameNode的ZKFC连接到ZooKeeper，把主机名等信息保存到ZooKeeper中，即“/hadoop-ha”下的znode目录里。先创建znode目录的NameNode节点为主节点，另一个为备节点。HDFS NameNode Standby通过ZooKeeper定时读取NameNode信息。
2. 当主节点进程异常结束时，HDFS NameNode Standby通过ZooKeeper感知“/hadoop-ha”目录下发生了变化，NameNode会进行主备切换。

ZooKeeper 和 YARN 的关系

ZooKeeper与YARN的关系如图1-143所示。

图 1-143 ZooKeeper 与 YARN 的关系

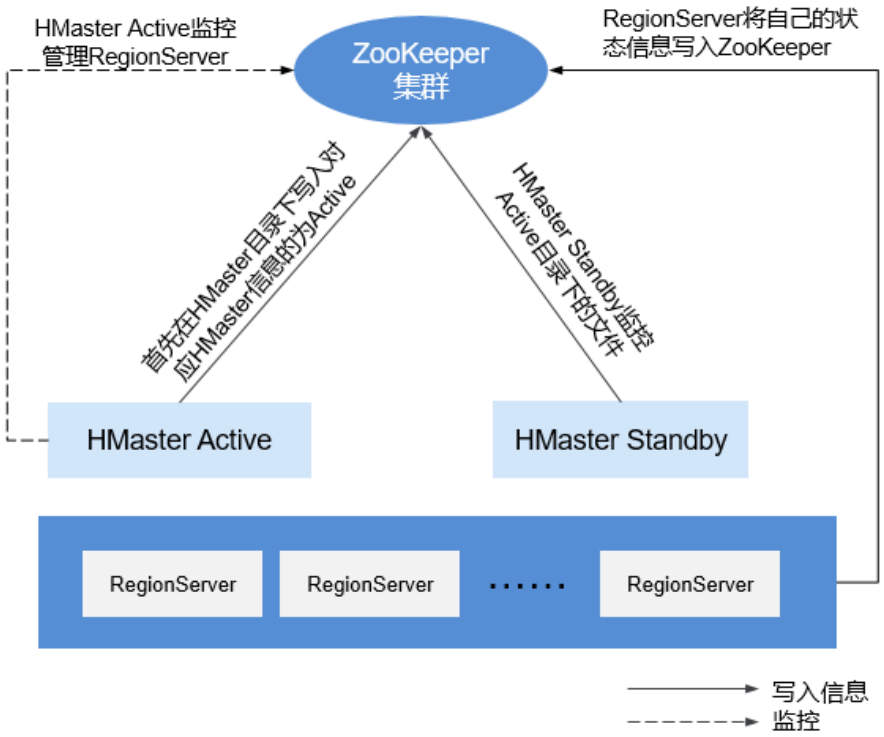


1. 在系统启动时，ResourceManager会尝试把选举信息写入ZooKeeper，第一个成功写入ZooKeeper的ResourceManager被选举为Active ResourceManager，另一个为Standby ResourceManager。Standby ResourceManager定时去ZooKeeper监控Active ResourceManager选举信息。
2. Active ResourceManager还会在ZooKeeper中创建Statestore目录，存储Application相关信息。当Active ResourceManager产生故障时，Standby ResourceManager会从Statestore目录获取Application相关信息，恢复数据。

ZooKeeper 和 HBase 的关系

ZooKeeper与HBase的关系如图1-144所示。

图 1-144 ZooKeeper 和 HBase 的关系

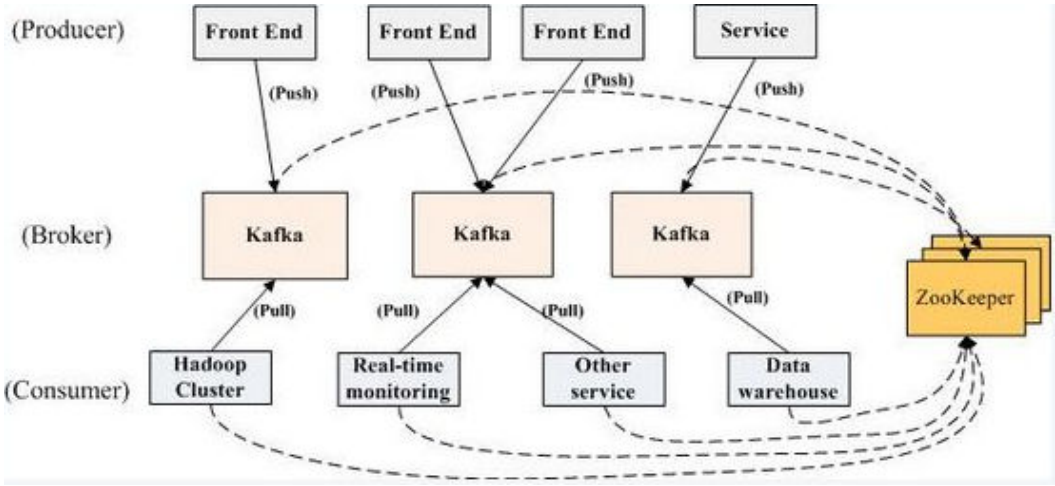


1. HRegionServer以Ephemeral node的方式注册到ZooKeeper中。其中ZooKeeper存储HBase的如下信息：HBase元数据、HMaster地址。
2. HMaster通过ZooKeeper随时感知各个HRegionServer的健康状况，以便进行控制管理。
3. HBase也可以部署多个HMaster，类似HDFS NameNode，当HMaster主节点出现故障时，HMaster备用节点会通过ZooKeeper获取主HMaster存储的整个HBase集群状态信息。即通过ZooKeeper实现避免HBase单点故障问题的问题。

ZooKeeper 和 Kafka 的配合关系

ZooKeeper与Kafka的关系如图 ZooKeeper和Kafka的关系所示。

图 1-145 ZooKeeper 和 Kafka 的关系



1. Broker端使用ZooKeeper用来注册broker信息，并进行partition leader选举。
2. Consumer端使用ZooKeeper用来注册consumer信息，其中包括consumer·消费的partition列表等，同时也用来发现broker列表，并和partition leader建立socket连接，并获取消息。

1.4.28.3 ZooKeeper 开源增强特性

日志增强

安全模式下，Ephemeral node（临时节点）在session过期之后就会被系统删除，在审计日志中添加Ephemeral node被删除的审计日志，以便了解当时Ephemeral node的状态信息。

所有ZooKeeper客户端的操作都要在审计日志中添加Username。

从ZooKeeper客户端创建znode，其kerberos principal是“zkcli/hadoop.*<系统域名>*@*<系统域名>*”。

例如打开日志<ZOO_LOG_DIR>/zookeeper_audit.log，内容如下：

```
2016-12-28 14:17:10,505 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test1?result=success
2016-12-28 14:17:10,530 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test2?result=success
2016-12-28 14:17:10,550 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test3?result=success
2016-12-28 14:17:10,570 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test4?result=success
2016-12-28 14:17:10,592 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test5?result=success
2016-12-28 14:17:10,613 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test6?result=success
2016-12-28 14:17:10,633 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test7?result=success
```

输出显示了在审计日志中添加了ZooKeeper客户端用户“zkcli/hadoop.hadoop.com@HADOOP.COM”的日志。

ZooKeeper中的用户详情：

在ZooKeeper中，不同的认证方案使用不同的凭证作为用户。 基于认证供应商的要求，任何参数都可以被认为是用户。

示例：

- SAMLAuthenticationProvider使用客户端主体作为用户。
- X509AuthenticationProvider使用户客户端证书作为用户。
- IAuthenticationProvider使用客户端IP作为用户。
- 自定义认证提供程序实现org.apache.zookeeper.server.auth.ExtAuthenticationProvider.getUserName（String）方法以获取用户名。 如果没有实现，从认证提供程序实例获取用户名将被跳过。

ZooKeeper 开源增强特性：ZooKeeper SSL 通信（Netty 连接）

ZooKeeper设计最初含有Nio包，且不能较好的支持3.5版本后的SSL。为了解决这个问题，Netty被加入到ZooKeeper中。所以如果用户需要使用SSL，启用Netty并设置Server端和Client端的以下参数。

开源的服务端只支持简单的文本密码，这可能导致相关安全问题。为此在服务端将不再使用此类文本密码。

- Client端
  a. 将“zkCli.sh/zkEnv.sh”文件中的参数“-Dzookeeper.client.secure”设置为“true”以在Client端使用安全通信。之后客户端可以连接服务端的secureClientPort。
  b. 通过设置“zkCli.sh/zkEnv.sh”文件中的以下参数配置客户端环境。

参数	描述
-Dzookeeper.clientCnxnSocket	用于客户端的Netty通信。 默认值： "org.apache.zookeeper.ClientCnxnSocketNetty"
-Dzookeeper.ssl.keyStore.location	keystore文件路径。
-Dzookeeper.ssl.keyStore.password	加密密码。
-Dzookeeper.ssl.trustStore.location	truststore文件路径。
-Dzookeeper.ssl.trustStore.password	加密密码。
-Dzookeeper.config.crypt.class	用于加密密码的解密。
-Dzookeeper.ssl.password.encrypted	默认值：false 当keystore和truststore的密码为加密密码时设置为true。

参数	描述
-Dzookeeper.ssl.enabled.protocols	通过配置此参数定义SSL协议以适用于SSL上下文。
-Dzookeeper.ssl.exclude.cipher.ext	通过配置此参数定义SSL上下文中应排除的密码列表，之间以逗号间隔。

说明

以上参数须在“zkCli.sh/zk.Env.sh”文件内设置。

- Server端
  a. 在文件“zoo.cfg”中将监听SSL端口参数“secureClientPort”设置为“3381”。
  b. 在server端将文件“zoo.cfg”中的参数“zookeeper.serverCnxnFactory”设置为“org.apache.zookeeper.server.NettyServerCnxnFactory”。
  c. 设置文件zoo.cfg(路径：“zookeeper/conf/zoo.cfg”)中的以下参数来配置服务端环境。

参数	描述
ssl.keyStore.location	keystore.jks文件路径。
ssl.keyStore.password	加密密码。
ssl.trustStore.location	truststore文件路径。
ssl.trustStore.password	加密密码。
config.crypt.class	用于加密密码的解密。
ssl.keyStore.password.encrypted	默认值：false 设置为true时可使用加密密码。
ssl.trustStore.password.encrypted	默认值：false 设置为true时可使用加密密码。
ssl.enabled.protocols	通过配置此参数定义SSL协议以适用于SSL上下文。
ssl.exclude.cipher.ext	通过配置此参数定义SSL上下文中应排除的密码列表，之间以逗号间隔。

  d. 启动ZKserver，然后将安全客户端连接到安全端口。
- 凭证
  ZooKeeper上Client和Server之间的凭证由X509AuthenticationProvider执行。根据以下参数指定服务端证书及信任客户端证书，并通过这些证书初始化X509AuthenticationProvider。

- zookeeper.ssl.keyStore.location
- zookeeper.ssl.keyStore.password
- zookeeper.ssl.trustStore.location
- zookeeper.ssl.trustStore.password

说明

若用户不想使用ZooKeeper的默认机制，可根据所需配置不同的ZooKeeper信任机制。

1.5 产品功能

1.5.1 多租户

特性简介

现代企业的数据集群在向集中化和云化方向发展，企业级大数据集群需要满足：

- 不同用户在集群上运行不同类型的应用和作业（分析、查询、流处理等），同时存放不同类型和格式的数据。
- 某些类型的用户（例如银行、政府单位等）对数据安全非常关注，很难容忍将自己的数据与其他用户放在一起。

这给大数据集群带来了以下挑战：

- 合理地分配和调度资源，以支持多种应用和作业在集群上平稳运行。
- 对不同的用户进行严格的访问控制，以保证数据和业务的安全。

多租户将大数据集群的资源隔离成一个个资源集合，彼此互不干扰，用户通过“租用”需要的资源集合，来运行应用和作业，并存放数据。在大数据集群上可以存在多个资源集合来支持多个用户的不同需求。

因此，MRS大数据集群提供了完整的企业级大数据多租户解决方案。多租户是MRS大数据集群中的多个资源集合（每个资源集合是一个租户），具有分配和调度资源（资源包括计算资源和存储资源）的能力。

特性优势

- 合理配置和隔离资源

  租户之间的资源是隔离的，一个租户对资源的使用不影响其它租户，保证了每个租户根据业务需求去配置相关的资源，可提高资源利用效率。
- 测量和统计资源消费

  系统资源以租户为单位进行计划和分配，租户是系统资源的申请者和消费者，其资源消费能够被测量和统计。
- 保证数据安全和访问安全

  多租户场景下，分开存放不同租户的数据，以保证数据安全；控制用户对租户资源的访问权限，以保证访问安全。

调度器增强

多租户根据调度器类型分为开源的Capacity调度器和自主研发的增强型Superior调度器。

为满足企业需求，克服Yarn社区在调度上遇到的挑战与困难，自主研发的Superior调度器，不仅集合了当前Capacity调度器与Fair调度器的优点，还做了以下增强：

- 增强资源共享策略

  Superior调度器支持队列层级，在同集群集成开源调度器的特性，并基于可配置策略进一步共享资源。针对实例，MRS集群管理员可通过Superior调度器为队列同时配置绝对值或百分比的资源策略计划。Superior调度器的资源共享策略将YARN的标签调度增强为资源池特性，YARN集群中的节点可根据容量或业务类型不同，进行分组以使队列更有效地利用资源。

- 基于租户的资源预留策略

  部分租户可能在某些时间中运行关键任务，租户所需的资源应保证可用。Superior调度器构建了支持资源预留策略的机制，在这些租户队列运行的任务可立即获取到预留资源，以保证计划的关键任务可正常执行。

- 租户和资源池的用户公平共享

  Superior调度器提供了队列内用户间共享资源的配置能力。每个租户中可能存在不同权重的用户，高权重用户可能需要更多共享资源。

- 大集群环境下的调度性能优势

  Superior调度器接收到各个NodeManager上报的心跳信息，并将资源信息保存在内存中，使得调度器能够全局掌控集群的资源使用情况。Superior调度器采用了push调度模型，令调度更加精确、高效，大大提高了大集群下的资源使用率。另外，Superior调度器在NodeManager心跳间隔较大的情况下，调度性能依然优异，不牺牲调度性能，也能避免大集群环境下的“心跳风暴”。

- 优先策略

  当某个服务在获取所有可用资源后还无法满足最小资源的要求，则会发生优先抢占。抢占功能默认关闭。

1.5.2 安全增强

MRS作为一个海量数据管理和分析的平台，具备高安全性。MRS主要从以下几个方面保障用户的数据和业务运行安全。

- 网络隔离

  整个系统部署在公有云上的虚拟私有云中，提供隔离的网络环境，保证集群的业务、管理的安全性。结合虚拟私有云的子网划分、路由控制、安全组等功能，为用户提供高安全、高可靠的网络隔离环境。

- 资源隔离

  MRS服务支持资源专属区内部署，专属区内物理资源隔离，用户可以在专属区内灵活地组合计算存储资源，包括专属计算资源+共享存储资源、共享计算资源+专属存储资源、专属计算资源+专属存储资源。

- 主机安全

  MRS支持与公有云安全服务集成，支持漏洞扫描、安全防护、应用防火墙、堡垒机、网页防篡改等。针对操作系统和端口部分，提供如下安全措施：

  - 操作系统内核安全加固
  - 更新操作系统最新补丁
  - 操作系统权限控制
  - 操作系统端口管理
  - 操作系统协议与端口防攻击

- 应用安全

  通过如下措施保证大数据业务正常运行：

  - 身份鉴别和认证
  - Web应用安全
  - 访问控制
  - 审计安全
  - 密码安全

- 数据安全

  针对海量用户数据，提供如下措施保障客户数据的机密性、完整性和可用性。

  - 容灾：MRS支持将数据备份到OBS（对象存储服务）中，支持跨区域的高可靠性。
  - 备份：MRS支持针对DBService、NameNode、LDAP的元数据备份和对HDFS、HBase的业务数据备份。

- 数据完整性

  通过数据校验，保证数据在存储、传输过程中的数据完整性。

  - 用户数据保存在HDFS上，HDFS默认采用CRC32C校验数据的正确性。
  - HDFS的DataNode节点负责存储校验数据，如果发现客户端传递过来的数据有异常（不完整）就上报异常给客户端，让客户端重新写入数据。
  - 客户端从DataNode读数据的时候会同步检查数据是否完整，如果发现数据不完整，尝试从其它的DataNode节点上读取数据。

- 数据保密性

  MRS分布式文件系统在Apache Hadoop版本基础上，提供对文件内容的加密存储功能，避免敏感数据明文存储，提升数据安全性。业务应用只需对指定的敏感数据进行加密，加解密过程业务完全不感知。在文件系统数据加密基础上，Hive实现表级加密，HBase实现列族级加密，在创建表时指定采用的加密算法，即可实现对敏感数据的加密存储。

  从数据的存储加密、访问控制来保障用户数据的保密性。

  - HBase支持将业务数据存储到HDFS前进行压缩处理，且用户可以配置AES和SMS4算法加密存储。
  - 各组件支持本地数据目录访问权限设置，无权限用户禁止访问数据。
  - 所有集群内部用户信息提供密文存储。

- 安全认证
  - 基于用户和角色的认证统一体系，遵从帐户/角色RBAC（Role-Based Access Control）模型，实现通过角色进行权限管理，对用户进行批量授权管理。
  - 支持安全协议Kerberos，MRS使用LDAP作为帐户管理系统，并通过Kerberos对帐户信息进行安全认证。
  - 提供单点登录，统一了MRS系统用户和组件用户的管理及认证。
  - 对登录Manager的用户进行审计。

1.5.3 组件 WebUI 便捷访问

大数据组件都有自己的WebUI页面管理自身系统，但是由于网络隔离的原因，用户并不能很简便地访问到该页面。比如访问HDFS的WebUI页面，传统的操作方法是需要用户创建ECS，使用ECS远程登录组件的UI，这使得组件的页面UI访问很是繁琐，对于很多初次接触大数据的用户很不友好。

MRS提供了基于弹性公网IP来便捷访问组件WebUI的安全通道，并且比用户自己绑定弹性公网IP更便捷，只需界面鼠标操作，即可简化原先用户需要自己登录虚拟私有云添加安全组规则，获取公网IP等步骤，减少了用户操作步骤。分析集群Hadoop、Spark、HBase、Hue及流式集群Storm，都可以在Manager上找到组件页面入口，快速访问。

1.5.4 可靠性增强

MRS在基于Apache Hadoop开源软件的基础上，在主要业务部件的可靠性、性能调优等方面进行了优化和提升。

系统可靠性

- 管理节点均实现HA

  Hadoop开源版本的数据、计算节点已经是按照分布式系统进行设计的，单节点故障不影响系统整体运行；而以集中模式运作的管理节点可能出现的单点故障，就成为整个系统可靠性的短板。

  MRS对所有业务组件的管理节点都提供了类似的双机的机制，包括Manager、HDFS NameNode、HiveServer、HBase HMaster、YARN ResourceManager、KerberosServer、LdapServer等，全部采用主备或负荷分担配置，有效避免了单点故障场景对系统可靠性的影响。

- 异常场景下的可靠性保证

  通过可靠性分析方法，梳理软件、硬件异常场景下的处理措施，提升系统的可靠性。

  - 保障意外掉电时的数据可靠性，不论是单节点意外掉电，还是整个集群意外断电，恢复供电后系统能够正常恢复业务，除非硬盘介质损坏，否则关键数据不会丢失。
  - 硬盘亚健康检测和故障处理，对业务不造成实际影响。
  - 自动处理文件系统的故障，自动恢复受影响的业务。
  - 自动处理进程和节点的故障，自动恢复受影响的业务。
  - 自动处理网络故障，自动恢复受影响的业务。

- 数据备份与恢复

  为应对数据丢失或损坏对用户业务造成不利影响，在异常情况下快速恢复系统，MRS根据用户业务的需要提供全量备份、增量备份和恢复功能。

  - 自动备份

    MRS对集群管理系统Manager上的数据提供自动备份功能，根据制定的备份策略可自动备份集群上的数据，包括LdapServer、DBService的数据。

  - 手动备份

    在系统进行扩容、打补丁等重大操作前，需要通过手动备份集群管理系统的数据，以便在系统故障时，恢复集群管理系统功能。

    为进一步提供系统的可靠性，在将Manager、HBase上的数据备份到第三方服务器时，也需要通过手动备份。

节点可靠性

- 操作系统健康状态监控

  周期采集操作系统硬件资源使用率数据，包括CPU、内存、硬盘、网络等资源的使用率状态。

- 进程健康状态监控

  MRS提供业务实例的状态以及业务实例进程的健康指标的检查，能够让用户第一时间感知进程健康状态。

- 硬盘故障的自动处理

  MRS对开源版本进行了增强，可以监控各节点上的硬盘以及文件系统状态。如果出现异常，立即将相关分区移出存储池；如果硬盘恢复正常（通常是因为用户更换了新硬盘），也会将新硬盘重新加入业务运作。这样极大简化了维护人员的工作，更换故障硬盘可以在线完成；同时用户可以设置热备盘，从而极大缩减了故障硬盘的修复时间，有利于提高系统的可靠性。

- 节点磁盘LVM配置

  MRS支持将多个磁盘配置成LVM（Logic Volume Management），多个磁盘规划成一个逻辑卷组。配置成LVM可以避免各磁盘间使用不均的问题，保持各个磁盘间均匀使用在HDFS和Kafka等能够利用多磁盘能力的组件上尤其重要。并且LVM可以支持磁盘扩容时不需要重新挂载，避免了业务中断。

数据可靠性

MRS可以利用弹性云服务器ECS提供的反亲和节点组以及放置组的能力，结合Hadoop的机架感知能力，将数据冗余到多个物理宿主机上，避免物理硬件的失效造成数据的失效。

1.5.5 作业管理

作业管理为用户提供向集群提交作业的入口，支持包括MapReduce、Spark、HiveQL和SparkSQL等类型的作业。结合数据湖工厂（DLF），提供一站式的大数据协同开发环境、全托管的大数据调度能力，帮助用户快速构建大数据处理中心。

通过数据湖工厂（DLF），用户可以先在线开发调试MRS HiveQL/SparkSQL脚本、拖拽式地开发MRS作业，完成MRS与其他20多种异构数据源之间的数据迁移和数据集成；通过强大的作业调度与灵活的监控告警，轻松管理数据作业运维。

1.5.6 自定义引导操作

特性简介

MRS提供标准的云上弹性大数据集群，目前可安装部署包括hadoop、spark等9种大数据组件。当前标准的云上大数据集群不能满足所有用户需求，例如如下几种场景：

- 通用的操作系统配置不能满足实际数据处理需求，例如需调大系统最大连接数。
- 需要安装自身业务所需的软件工具或运行环境，例如须安装gradle、业务需要依赖R语言包。
- 根据自身业务对大数据组件包做修改，例如对hadoop或spark安装包做修改。
- 需要安装其他MRS还未支持的大数据组件。

对于上述定制化的场景，可以选择登录到每个节点上手动操作，之后每扩容一个新节点，再执行一次同样的操作，操作相对繁琐，也容易出错。同时手动执行记录不便追溯，不能实现“按需创建、创建成功后即处理数据”的目标。

因此，MRS提供了自定义引导操作，在启动集群组件前（或后）可以在指定的节点上执行脚本。用户可以通过引导操作来完成安装MRS还没支持的第三方软件，修改集群运行环境等自定义操作。如果集群扩容，选择执行引导操作，则引导操作也会以相同

方式在新增节点上执行。MRS会使用root用户执行您指定的脚本，脚本内部您可以通过**su - *xxx***命令切换用户。

客户价值

MRS提供了自定义引导操作，用户可以以此为入口，灵活、便捷地配置自己的专属集群，自定义安装软件。

1.5.7 企业项目管理

企业项目是一种云资源管理方式。企业管理提供面向企业客户的云上资源管理、人员管理、权限管理、财务管理等综合管理服务。区别于管理控制台独立操控、配置云产品的方式，企业管理控制台以面向企业资源管理为出发点，帮助企业以公司、部门、项目等分级管理方式实现企业云上的人员、资源、权限、财务的管理。

MRS支持已开通企业项目服务的用户在创建集群时为集群配置对应的项目，然后使用企业项目管理对MRS上的资源进行分组管理：

- 支持用户为多个资源进行分组管理。
- 支持用户查看企业项目下的资源信息、消费明细。
- 支持用户对企业项目级别的访问权限控制。
- 支持用户分企业项目查看具体的财务信息，包括订单、消费汇总、消费明细等。

1.5.8 元数据

当创建MRS集群时选择部署Hive和Ranger组件时，MRS提供多种元数据存储方式，您可以根据自身需要进行选择：

- 本地元数据：元数据存储于集群内的本地GaussDB中，当集群删除时元数据同时被删除，如需保存元数据，需提前前往数据库手动保存元数据。
- 数据连接：可选择关联与当前集群同一虚拟私有云和子网的RDS服务中的PostgresDB或MySQL数据库，元数据将存储于关联的数据库中，不会随当前集群的删除而删除，多个MRS集群可共享同一份元数据。

说明

Hive组件可选元数据存储方式功能在MRS 1.9.x及之后版本支持。

Ranger组件可选元数据存储方式功能目前仅在MRS 1.9.x版本支持关联RDS服务中的MySQL数据库。

1.5.9 集群管理

1.5.9.1 集群生命周期管理

MRS支持集群的生命周期管理包括创建集群和删除集群。

- 创建集群：支持用户定制集群的类型，组件范围，各类型的节点数、虚拟机规格、可用区、VPC网络、认证信息，MRS将为用户自动创建一个符合配置的集群，全程无需用户参与；同时支持用户在集群中运行自定义内容；支持快速创建多应用场景集群，比如创建Hadoop分析集群、HBase集群、Kafka集群。大数据平台同时支持部署异构集群，在集群中存在不同规格的虚机，允许在CPU类型，硬盘容量，硬盘类型，内存大小灵活组合。在集群中支持多种虚机规格混合使用。

- 删除集群：当集群不再需要时（包括集群中的数据和配置），用户可以选择删除集群，MRS会将集群相关的资源全部删除。

创建集群

通过在MRS服务管理面，客户可以创建MRS集群，通过选择集群所建的区域及使用的云资源规格，一键式创建适合企业业务的MRS集群。MRS服务会根据用户选择的集群类型、版本和节点规格，帮助客户自动完成企业级大数据平台的安装部署和参数调优。

MRS服务为客户提供完全可控的大数据集群，客户在创建时可设置虚拟机的登录方式（密码或者密钥对），所创建的MRS集群资源完全归客户所用。同时MRS支持在最小可在两节点4U8G的ECS上部署大数据集群，为客户测试开发提供更多的灵活选择。

MRS集群类型包括分析集群、流式集群和混合集群。

- 分析集群：用来做离线数据分析，提供的是Hadoop体系的组件。
- 流式集群：用来做流处理任务，提供的是流式处理组件。
- 混合集群：既可以用来做离线数据分析，又可以用来做流处理任务，提供的是Hadoop体系的组件和流式处理组件。
- 自定义：根据业务需求，可以灵活搭配所需组件（MRS 3.x及后续版本）。

MRS集群节点类型包括Master节点、Core节点和Task节点。

- Master节点：集群中的管理节点，分布式系统的Master进程和Manager以及数据库均部署在该节点；该类型节点不可扩容。该类型节点的处理能力决定了整个集群的管理上限，MRS服务支持将Master节点规格提高，以支持更大集群的管理。
- Core节点：支持存储和计算两种目标的节点，可扩容、缩容。因承载的数据存储，因此在缩容时，为保证数据不丢失，有较多限制，无法进行弹性伸缩。
- Task节点：仅用于计算的节点，可扩容、缩容。因只承载计算任务，因此可以进行弹性伸缩。

MRS创建集群方式支持自定义创建集群和快速创建集群两种。

- 自定义创建集群：自定义可以灵活地选择配置项，针对不同的应用场景，可以选择不同规格的弹性云服务器，全方位贴合您的业务诉求。
- 快速创建集群：用户可以根据应用场景，快速创建对应配置的集群，提高了配置效率，更加方便快捷。当前支持快速创建Hadoop分析集群、HBase集群、Kafka集群。
  - Hadoop分析集群：Hadoop分析集群完全使用开源Hadoop生态，采用YARN管理集群资源，提供Hive、Spark离线大规模分布式数据存储和计算，SparkStreaming、Flink流式数据计算，Presto交互式查询，Tez有向无环图的分布式计算框等Hadoop生态圈的组件，进行海量数据分析与查询。
  - HBase集群：HBase集群使用Hadoop和HBase组件提供一个稳定可靠，性能卓越、可伸缩、面向列的分布式云存储系统，适用于海量数据存储以及分布式计算的场景，用户可以利用HBase搭建起TB至PB级数据规模的存储系统，对数据轻松进行过滤分析，毫秒级得到响应，快速发现数据价值。
  - Kafka集群：Kafka集群使用Kafka和Storm组件提供一个开源高吞吐量，可扩展性的消息系统。广泛用于日志收集、监控数据聚合等场景，实现高效的流式数据采集，实时数据处理存储等。

删除集群

MRS服务支持用户在不需要大数据集群时执行删除集群操作，集群删除后，所有大数据使用的相关云资源都会同时被释放。删除集群前，建议完成数据搬迁或者备份，确认集群无任何业务运行或者集群异常且经运维分析无法继续提供服务时再执行集群删除操作。对于数据存放在云硬盘EVS或直通盘的大数据集群，集群删除后，数据也随之删除，强烈建议您慎重选择删除集群。

1.5.9.2 集群扩缩容

大数据集群的处理能力通常可以通过增加集群的节点数来横向扩展，当集群规模不符合业务要求时，用户可以通过该功能进行集群节点规模的调整，进行扩容或者缩容；在缩容节点时，MRS会智能地选择负载最少或者迁移数据量最小节点，并且在缩容过程中，缩容节点不再接收新的任务，正在执行的任务继续执行，同时将该节点数据拷贝至其他节点，该节点进入退服状态，当该节点任务长时间运行无法结束时，会迁移至其他节点运行，最大限度地减少对集群业务的影响。

扩容集群

目前支持扩容集群Core节点或Task节点，用户可通过增加节点数量处理业务峰值负载。MRS集群节点扩容中和扩容后对现有集群的业务没有影响。

缩容集群

用户可以根据业务需求量，通过简单的缩减Core节点或者Task节点，对集群进行缩容，以使MRS拥有更优的存储、计算能力，降低运维成本。用户执行MRS集群缩容后，MRS服务将根据节点已安装的服务类型自动选择可以缩容的节点。

Core节点在缩容的时候，会对原节点上的数据进行迁移。业务上如果对数据位置做了缓存，客户端自动刷新位置信息可能会影响时延。缩容节点可能会影响部分HBase on HDFS数据的第一次访问响应时长，可以重启HBase或者对相关的表Disable/Enable来避免。

Task节点本身不存储集群数据，属于计算节点，不存在节点数据迁移的问题。

1.5.9.3 自动弹性伸缩

特性简介

随着企业的数据越来越多，越来越多的企业选择使用Spark/Hive等技术来进行分析，由于数据量大，处理任务繁重，资源的消耗比较高，因此使用成本也是比较高。当前并不是每个企业在每时每刻在进行分析，而一般是在一天的一个时间段内进行分析汇总，因此MRS提供了弹性伸缩能力，可以自动在业务在繁忙时申请额外资源，业务不繁忙时释放闲置资源，让用户按需使用，尽可能的帮助客户降低使用成本，聚焦核心业务。

在大数据应用，尤其是周期性的数据分析处理场景中，需要根据业务数据的周期变化，动态调整集群计算资源以满足业务需要。MRS的弹性伸缩规则功能支持根据集群负载对集群进行弹性伸缩。此外，如果数据量为周期有规律的变化，并且希望在数据量变化前提前完成集群的扩缩容，可以使用MRS的资源计划特性。

MRS服务支持规则和时间计划两种弹性伸缩的策略：

- 弹性伸缩规则：根据集群实时负载对Task节点数量进行调整，数据量变化后触发扩缩容，有一定的延后性。

- 资源计划：若数据量变化存在周期性规律，则可通过资源计划在数据量变化前提前完成集群的扩缩容，避免出现增加或减少资源的延后。

弹性伸缩规则与资源计划均可触发弹性伸缩，两者即可同时配置也可单独配置。资源计划与基于负载的弹性伸缩规则叠加使用可以使得集群节点的弹性更好，足以应对偶尔超出预期的数据峰值出现。

当某些业务场景要求在集群扩缩容之后，根据节点数量的变化对资源分配或业务逻辑进行更改时，手动扩缩容的场景客户可以登录集群节点进行操作。对于弹性伸缩场景，MRS支持通过自定义弹性伸缩自动化脚本来解决。自动化脚本可以在弹性伸缩前后执行相应操作，自动适应业务负载的变化，免去了人工操作。同时，自动化脚本给用户实现个性需求提供了途径，完全自定义的脚本与多个可选的执行时机基本可以满足用户的各项需求，使弹性伸缩更具灵活性。

客户价值

MRS的自动弹性伸缩可以帮助用户实现以下价值。

- 降低使用成本

  部分企业在进行批量分析时，并不是时时刻刻都在进行分析，例如一般都存在数据持续接入，而到了特定时间段（例如凌晨3点）进行批量分析，可能仅需要消耗2小时。

  MRS提供的弹性伸缩能力，可以帮助客户，在晚上的时候，将分析节点扩容到指定规模，而计算完毕后，则自动释放计算节点，尽可能的降低使用成本。

- 平衡突发查询

  大数据集群上，由于有大量的数据，企业会经常面临临时的分析任务，例如支撑企业决策的临时数据报表等，都会导致对于资源的消耗在极短时间内剧增。MRS提供的弹性伸缩能力，可以让突发大数据分析时，可以及时的补充计算节点，避免因为计算能力不足，导致业务宕机，使用户无需购买额外资源，当突发事件结束后，MRS会自动判断缩容时机，自动完成缩容。

- 聚焦核心业务

  大数据作为二次开发平台，开发人员非常难判断具体的资源消耗，因为查询分析的条件复杂性（例如全局排序，过滤，合并等）以及数据的复杂性，例如增量数据的不确定性等，都会导致预估多少计算量是非常困难的行为，而使用弹性伸缩能力，可以让业务人员专注于业务开发，无需分心再做各种资源评估。

1.5.9.4 创建 Task 节点

特性简介

支持创建Task节点，只作为计算节点，不存放持久化的数据，是实现弹性伸缩的基础。

客户价值

在MRS服务只作为计算资源的场景下，使用Task节点可以节省成本，并可以更加方便快捷地对集群节点进行扩缩容，满足用户对集群计算能力随时增减的需求。

用户场景

当集群数据量变化不大而集群业务处理能力需求变化比较大，大的业务处理能力只是临时需要，此时选择添加Task节点。

- 临时业务量增大，如年底报表处理。
- 需要在短时间内处理完原来需要处理很久的任务，如一些紧急分析任务。

1.5.9.5 升级 Master 节点规格

MRS大数据集群采用Manager实现集群的管理，而管理集群的相关服务，如HDFS存储系统的NameNode，Yarn资源管理的ResourceManager，以及MRS的Manager管理服务都部署在集群的Master节点。

随着新业务的上线，集群规模不断扩大，Master节点承担的管理负荷也越来越高，企业用户面临CPU负载过高，内存使用率超过阈值的问题。通常自建大数据集群需要完成数据搬迁，采购升级节点硬件配置实现Master规格提升，而MRS服务借助云服务的优势，实现一键式Master节点升级，并在升级过程中通过Master节点的主备HA保证已有业务的不间断，方便快捷帮助用户解决主节点规格升级问题。

Master节点具体升级操作请参见**升级Master节点规格**。

1.5.9.6 隔离主机

用户发现某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。在为集群安装补丁的场景中，也支持排除指定节点不安装补丁。隔离主机仅支持隔离非管理节点。

主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。另外，主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。

1.5.9.7 标签管理

标签是集群的标识，为集群添加标签，可以方便用户识别和管理拥有的集群资源。MRS服务通过与标签管理服务（TMS）关联，可以让拥有大量云资源的用户，通过给云资源打标签，快速查找具有同一标签属性的云资源，进行统一检视、修改、删除等管理操作，方便用户对大数据集群及其他相关云资源的统一管理。

您可以在创建集群时添加标签，也可以在集群创建完成后，在集群的详情页添加标签，您最多可以给集群添加10个标签。

1.5.10 集群运维

告警管理

MRS可以实时监控大数据集群，通过告警和事件可以识别系统健康状态。同时MRS也支持用户自定义配置监控与告警阈值用于关注各指标的健康情况，当监控数据达到告警阈值，系统将会触发一条告警信息。

MRS还可以与消息通知服务(SMN)的消息服务系统对接，将告警信息通过短信或者邮件等形式推送给用户。具体介绍请参见**消息通知**。

补丁管理

MRS集群支持补丁操作，会及时发布开源大数据组件的补丁。用户能够在MRS集群管理页面上查看到运行集群相关的补丁发布信息，包括其修复问题的详细说明及影响场景，客户可以根据业务运行情况自行选择是否安装补丁。补丁安装过程是一键式操作，无需人工干预，通过滚动安装，补丁升级不会停止业务，保障用户集群长期可用。

MRS服务可以展示详细的补丁安装过程，补丁管理也支持补丁的卸载和失败回滚。

说明

MRS 3.x及之后版本暂不支持在管理控制台执行补丁管理操作。

运维支撑

MRS提供的集群的资源是完全属于用户的，通常情况下，当集群出现问题，需要运维人员支撑时，运维人员是无法直接访问的。为了更好的服务客户，MRS提供两种方式来减少定位问题时的信息传递：

- 日志共享：用户可以在MRS 页面发起日志共享，选择日志范围共享给运维人员，以便运维人员在不接触集群的情况下帮助定位问题。
- 运维授权：MRS服务提供运维授权功能，用户在使用MRS集群过程中，发生问题可以在MRS页面发起运维授权，由运维人员帮助客户快速定位问题，用户可以随时收回该授权。

健康检查

MRS为用户提供界面化的系统运行环境自动检查服务，帮助用户实现一键式系统运行健康度巡检和审计，保障系统的正常运行，降低系统运维成本。用户查看检查结果后，还可导出检查报告用于存档及问题分析。

1.5.11 消息通知

特性简介

大数据集群运行过程中经常会进行如下操作：

- 大数据集群经常会发生变更，比如扩容、缩容集群。
- 业务数据量突然变化，集群触发弹性伸缩。
- 相关业务结束，需要终止大数据集群等。

用户想要及时得知这些操作是否成功了，以及当集群出现大数据服务不可用，或节点故障时，用户希望不用隔段时间就登录集群查看，而是可以及时地收到告警通知。MRS联合消息通知服务(SMN)，可以将以上信息主动地通知到用户的手机及邮箱，让维护更加省心省力。

客户价值

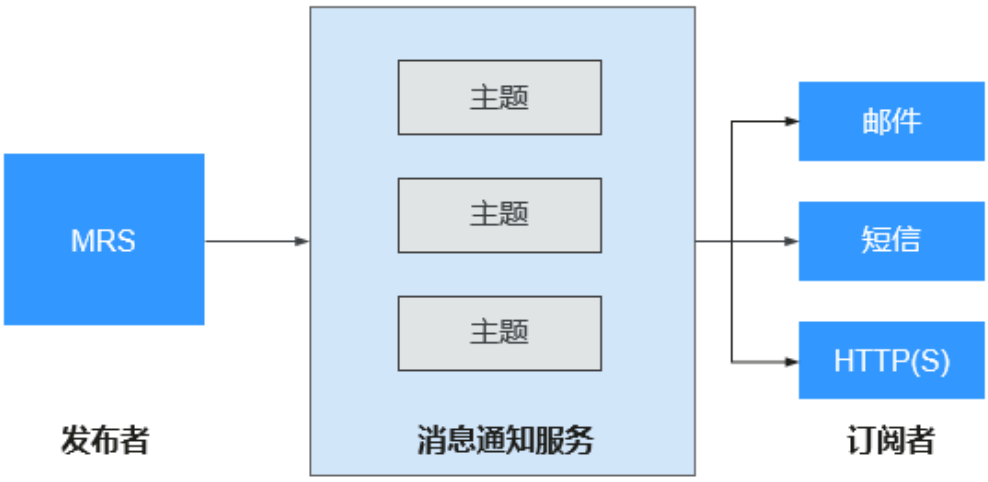
配置消息通知后，可以实时给用户发送MRS集群健康状态，用户可以通过手机短信或邮箱实时接收到MRS集群变更及组件告警信息。MRS可以帮助用户轻松运维，实时监控，实时发送告警，操作灵活，大数据业务部署更加省心省力。

特性描述

MRS联合消息通知服务(SMN)，采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式。

首先，作为主题拥有者，可以先创建一个主题，并对主题设置访问控制权限来决定哪些发布者和订阅者可以通过该主题进行交流。MRS将集群消息发送至您有权限发布消息的主题，然后所有订阅了该主题的订阅者（可以是手机短信、邮箱等）都将收到集群变更以及组件告警的消息。

图 1-146 实现过程



1.6 约束与限制

使用MRS前，您需要认真阅读并了解以下使用限制。

- MRS集群必须创建在VPC子网内。
- 建议使用支持的浏览器登录MRS。
  - Google Chrome：36.0及更高版本
  - Internet Explorer：9.0及更高版本

    当使用Internet Explorer 9.0时可能无法登录MRS管理控制台，原因是某些Windows系统例如Win7旗舰版，默认禁止Administrator用户，Internet Explorer在安装时自动选择其他用户如System用户安装，从而导致Internet Explorer无法打开登录页面。请使用管理员身份重新安装Internet Explorer 9.0或更高版本（建议），或尝试使用管理员身份运行Internet Explorer 9.0。
- 创建MRS集群时，支持自动创建安全组，也可从下拉框中选择已有的安全组。集群创建完成后，请勿随意删除或更改已使用的安全组。否则可以能导致集群异常，影响MRS集群的使用。
- MRS集群使用的安全组请勿随意放开权限，避免被恶意访问。
- 请勿随意执行如下操作，避免集群进入异常状态，影响MRS集群的使用。
  - 在ECS中对MRS集群的节点进行关机、重启、删除、变更OS、重装OS和修改规格等操作。
  - 删除集群节点上已有的进程、安装的应用程序和文件。
- 集群处于非人为异常状态时，可以联系技术支持人员，技术支持人员征得您同意后会请您提供密码，登录MRS集群进行问题排查。
- 请根据业务需要规划集群节点的磁盘，如果需要存储大量业务数据，请增加云硬盘数量或存储空间。以防止存储空间不足影响节点正常运行。
- 集群节点仅用于存储用户业务数据，非业务数据建议保存在对象存储服务或其他弹性云服务器中。
- 集群节点仅用于运行MRS集群，其他客户端应用程序、用户业务程序建议申请独立弹性云服务器部署。

- MRS集群的节点（包含Master，Core和Task节点）扩充存储容量时，不支持通过扩容原有磁盘的方式实现，需要通过新购磁盘再挂载的方式实现。
- 当关闭集群中的某一个Master节点后仍然使用集群执行任务或修改配置，且任务执行或配置修改后未启动被关闭的Master节点就关闭集群中其他Master节点时，存在由于主备倒换导致的数据丢失风险。在该场景下，请在任务执行或配置修改后先启动被关闭的Master节点，再关闭全部节点。若集群中节点已经被全部关闭，请按照节点关机顺序的倒序启动集群节点。
- 当使用MRS集群过程中进行Capacity和Superior调度器切换时只完成调度器的切换，不保证配置同步。如果您需要配置同步，请基于新的调度器重新配置。

1.7 技术支持

云原生数据湖MRS 服务（MapReduce Service）是租户完全可控的半托管云服务，为用户提供一站式企业级大数据平台，用户可以在MRS集群上轻松运行Hadoop、Hive、Spark、HBase、Kafka、Flink等大数据组件，帮助企业快速构建海量数据信息处理系统，并通过对海量信息数据实时与非实时的分析挖掘，发现全新价值点和企业商机。

维护策略声明

MRS租户集群资源归属于用户，MRS提供基于该资源的半托管云服务能力，用户拥有对集群使用的完全控制权，默认情况下，云服务无权限对客户集群进行操作，集群日常运维管理由用户负责，如果在大数据集群运维过程中遇到了相关技术问题，可以联系技术支持团队获得帮助，该技术支持仅协助分析处理MRS云服务相关求助，不包含云服务以外的求助，例如用户基于大数据平台构建的应用系统等。

技术支持范围

- MRS云服务管理控制台提供的相关功能：
  - 集群的创删扩缩
  - 集群作业管理
  - 集群告警管理
  - 集群补丁管理
  - IAM用户委托管理
  - 对外API接口管理
- 支持客户进行MRS服务相关开源组件漏洞分析，如影响分析、修复建议，由用户负责评估对应的业务影响和负责最终实施。

1.8 权限管理

如果您需要对云上创建的MapReduce服务资源，给企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制云资源的访问。

通过IAM，您可以在云帐号中给员工创建IAM用户，并授权控制他们对资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有MapReduce服务的使用权限，但是不希望他们拥有删除MRS集群等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用MRS，但是不允许删除MRS集群的权限策略，控制他们对MRS集群资源的使用范围。

如果云帐号已经能满足您的要求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用MRS服务的其它功能。

IAM是云提供权限管理的基础服务，无需付费即可使用。

MRS 权限说明

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

MRS部署时通过物理区域划分，为项目级服务。授权时，“作用范围”需要选择“区域级项目”，然后在指定区域对应的项目中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问MRS时，需要先切换至授权区域。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对MRS服务，管理员能够控制IAM用户仅能对集群进行指定的管理操作。如不允许某用户组删除集群，仅允许操作MRS集群基本操作，如创建集群、查询集群列表等。多数细粒度策略以API接口为粒度进行权限拆分。

如表1-29所示，包括了MRS的所有系统策略。

表 1-29 MRS 系统策略

策略名称	描述	策略类别
MRS FullAccess	MRS管理员权限，拥有该权限的用户可以拥有MRS所有权限。	细粒度策略
MRS CommonOperations	MRS服务普通用户权限，拥有该权限的用户可以拥有MRS服务使用权限，无新增、删除资源权限。	细粒度策略
MRS ReadOnlyAccess	MRS服务只读权限，拥有该权限的用户仅能查看MRS的资源。	细粒度策略
MRS Administrator	操作权限： • 对MRS服务的所有执行权限。 • 拥有该权限的用户必须同时拥有Tenant Guest、Server Administrator和BSS Administrator权限。	RBAC策略

表1-30列出了MRS常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 1-30 常用操作与系统策略的授权关系

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
创建集群	√	x	x	√
调整集群	√	x	x	√
升级节点规格	√	x	x	√
删除集群	√	x	x	√
查询集群详情	√	√	√	√
查询集群列表	√	√	√	√
设置弹性伸缩策略	√	x	x	√
查询主机列表	√	√	√	√
查询操作日志	√	√	√	√
创建并执行作业	√	√	x	√
停止作业	√	√	x	√
删除单个作业	√	√	x	√
批量删除作业	√	√	x	√
查询作业详情	√	√	√	√
查询作业列表	√	√	√	√
新建文件夹	√	√	x	√
删除文件	√	√	x	√
查询文件列表	√	√	√	√

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
批量操作集群标签	√	√	x	√
创建单个集群标签	√	√	x	√
删除单个集群标签	√	√	x	√
按照标签查询资源列表	√	√	√	√
查询集群标签	√	√	√	√
访问Manager页面	√	√	x	√
查询补丁列表	√	√	√	√
安装补丁	√	√	x	√
卸载补丁	√	√	x	√
运维通道授权	√	√	x	√
运维通道日志共享	√	√	x	√
查询告警列表	√	√	√	√
订阅告警消息提醒	√	√	x	√
提交SQL语句	√	√	x	√
查询SQL结果	√	√	x	√
取消SQL执行任务	√	√	x	√

1.9 与其他云服务的关系

MRS 服务与其他服务的关系

表 1-31 MRS 服务与其他服务的关系

服务名称	MRS服务与其他服务的关系
虚拟私有云（Virtual Private Cloud）	MRS集群创建在虚拟私有云（VPC）的子网内，VPC通过逻辑方式进行网络隔离，为用户的MRS集群提供安全、隔离的网络环境。
对象存储服务（Object Storage Service）	对象存储服务（OBS）用于存储用户数据，包括MRS作业输入数据和作业输出数据： ● MRS作业输入数据：用户程序和数据文件 ● MRS作业输出数据：作业输出的结果文件和日志文件 MRS中HDFS、Hive、MapReduce、YARN、Spark、Flume和Loader支持从OBS导入、导出数据。 MRS使用OBS的并行文件系统提供服务。
弹性云服务器（Elastic Cloud Server）	MRS服务使用弹性云服务器（Elastic Cloud Server，简称ECS）作为集群的节点，每个弹性云服务器是集群中的一个节点。
关系型数据库（Relational Database Service）	关系型数据库（RDS）用于存储MRS系统运行数据，包括MRS集群元数据等。
统一身份认证服务（Identity and Access Management）	统一身份认证服务（IAM）为MRS提供了鉴权功能。
消息通知服务（SMN）	MRS联合消息通知服务（SMN），采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式。
云审计服务（Cloud Trace Service）	云审计服务（CTS）为用户提供MRS资源操作请求及请求结果的操作记录，供用户查询、审计和回溯使用。

表 1-32 云审计支持的 MRS 操作列表

操作名称	资源类型	事件名称
创建集群	cluster_mrs	createCluster
删除集群	cluster_mrs	deleteCluster
集群扩容	cluster_mrs	scaleOutCluster
集群缩容	cluster_mrs	scaleInCluster

在您开启了云审计服务后，系统开始记录云服务资源的操作。云审计服务管理控制台保存最近7天的操作记录。详细操作步骤请参考“云审计服务（CTS）> 快速入门 > 查看追踪事件”。

1.10 常见概念

HBase 表

HBase的表是三个维度排序的映射。从行主键、列主键和时间戳映射为单元格的值。所有的数据存储在HBase的表单元格中。

列

HBase表的一个维度。列名称的格式为“<family>:<label>”，<family>和<label>为任意字符组合。表由<family>的集合组成（<family>又称为列族）。HBase表中的每个列都归属于某个列族。

列族

列族是预定义的列集合，存储在HBase Schema中。如果需要在列族下创建一些列，首先需创建列族。列族将HBase中具有相同性质的数据进行重组，且没有类型的限制。同一列族的每行数据存储在同一个服务器中。每个列族像一个属性，如压缩包、时间戳、数据块缓存等等。

MemStore

MemStore是HBase存储的核心，当WAL中数据存储达到一定量时，加载到MemStore进行排序存储。

RegionServer

RegionServer是HBase集群运行在每个工作节点上的服务。一方面维护Region的状态，提供对于Region的管理和服务；另一方面，上传Region的负载信息，参与Master的分布式协调管理。

时间戳

用于索引同一份数据的不同版本，时间戳的类型是64位整型。时间戳可以由HBase在数据写入时自动赋值或者由客户显式赋值。

Store

HBase存储的核心，一个Store拥有一个MemStore和多个StoreFile，一个Store对应一个分区中表的列族。

索引

一种数据结构，提高了对数据库表中的数据检索效率。可以使用一个数据库表中的一列或多列，提供了快速随机查找和有效访问有序记录的基础。

协处理器

HBase提供的在RegionServer执行的计算逻辑的接口。协处理器分两种类型，系统协处理器可以全局导入RegionServer上的所有数据表，表协处理器即是用户可以指定一张表使用协处理器。

Block Pool

Block Pool是隶属于单个Namespace的块的集合。DataNode存储来自集群中所有块池的块。每个块池都是独立管理的。这就允许一个Namespace为新块生成块ID，而不需要和其他Namespace合作。一个NameNode失效，不会影响DataNode为集群中其他NameNode提供服务。

DataNode

HDFS集群的工作节点。根据客户端或者是元数据节点的调度存储和检索数据，定期向元数据及客户端发送所存储的文件块的列表。

文件块

HDFS中存储的最小逻辑单元。每个HDFS文件由一个或多个文件块存储。所有的文件块存储在DataNode中。

文件块副本

一个副本是存储在HDFS中的一些文件块拷贝件。同一个文件块存储多个拷贝件主要用于系统的可用性和容错。

Namespace Volume

一种独立的（自给自足的）管理单元。一个Namespace和它的Block Pool，合称为“Namespace Volume”。当一个NameNode/Namespace被删除后，DataNode上的相关块池也会被删除。在集群升级时，每个命名空间卷将作为一个整体被升级。

NodeManager

负责执行应用程序的容器，同时监控应用程序的资源使用情况（CPU、内存、硬盘、网络）并且向ResourceManager汇报。

ResourceManager

集群的资源管理器，基于应用程序对资源的需求进行调度。资源管理器提供一个调度策略的插件，它负责将集群资源分配给多个队列和应用程序。调度插件可以基于现有的能力调度和公平调度模型。

分区

每一个Topic可以被分为多个Partition，每个Partition对应一个可持续追加的有序且不可变的log文件。

跟随者

跟随者（Follower）负责处理读请求的模块，配合Leader一起进行写请求处理。也可作为Leader的储备，当Leader故障时从Follower当中选举出Leader，避免出现单点故障。

观察者

观察者（Observer）不参与选举和写请求的投票，只负责处理读请求、并向Leader转发写请求，避免系统处理能力浪费。

领导者

作为ZooKeeper集群的领导者，由各Follower通过Zab协议选举产生。主要负责接受和协调所有写请求，并把写入的信息同步到Follower和Observer。

CarbonData

基于Spark SQL开放架构，将自研的MOLAP引擎和Spark深度集成，快速构建基于Spark的分布式多维分析引擎，使Spark分析性能从分钟级提升到秒级，增强了Spark的多维分析能力。

离散流

Spark Streaming提供的抽象概念。表示一个连续的数据流，是从数据源获取或者通过输入流转换生成的数据流。从本质上说，一个DStream表示一系列连续的RDD。

堆内存（Heap memory）

堆是JVM运行时数据区域，所有类实例和数组的内存均从此处分配。初始堆内存根据JVM启动参数-Xms控制，最大堆内存通过JVM启动参数的-Xmx进行控制。

- 最大堆内存（Maximum Heap memory）：系统可以分配给程序的最大堆内存，JVM启动参数的-xmx指定。
- 分配的堆内存（Committed Heap memory）：为保证程序运行的系统堆分配的堆内存总量，Committed heap memory在程序运行期间根据使用情况，会在初始堆内存和最大堆内存之间波动变化。
- 使用的堆内存（Used heap memory）：当前程序运行时已经使用的堆内存，这个内存小于Committed heap memory。
- 非堆内存：在JVM中堆之外的内存称为非堆内存（Non-heap memory），JVM自身运行时所需要的内存区域，非堆内存有多个内存池，通常包括以下3个部分：
  - 代码缓存区（Code Cache）主要用于存放JIT所编译的代码。默认限制240M，可以通过JVM启动参数-XX:InitialCodeCacheSize -XX:ReservedCodeCacheSize进行设置。
  - 类指针压缩空间（Compressed Class Space）存储类指针的元数据，默认限制1024M，通过JVM启动参数-XX:CompressedClassSpaceSize进行设置。
  - 元空间（Metaspace）用于存放元数据，通过JVM启动参数-XX:MetaspaceSize -XX:MaxMetaspaceSize进行设置。
- 最大非堆内存（Maximum Non Heap Memory）：系统可以分配给程序的最大非堆内存。其值为代码缓存区（Code Cache），类指针压缩空间（Compressed Class Space），元空间（Metaspace）最大值之和。

- 分配的非堆内存（Committed Non Heap Memory）：为保证程序运行的系统非堆内存总量，Committed Non Heap Memory在程序运行期间根据使用的非堆内存情况，会在初始非堆内存和最大非堆内存之间波动变化。
- 使用非堆内存（Used Non Heap memory）：当前程序运行时已经使用的非堆内存，这个值小于分配的非堆内存（Committed Non Heap Memory）。

Hadoop

一个分布式系统框架。用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用了集群的高速运算和存储。 Hadoop能够对大量数据以可靠的、高效的、可伸缩的方式进行分布式处理。Hadoop是可靠的，因为它假设计算单元和存储会失败，因此维护多个工作数据副本，确保对失败节点重新分布处理；Hadoop是高效的，因为它以并行的方式工作，从而加快处理速度；Hadoop是可伸缩的，能够处理PB级数据。 Hadoop主要由HDFS、MapReduce、HBase和Hive组成。

角色

角色是服务的组成要素，每个服务由一个或多个角色组成。服务通过角色安装到主机（即服务器）上，保证服务正常运行。

集群

将多个服务器集中起来使它们能够像一台服务器一样提供服务的计算机技术。采用集群通常是为了提高系统的稳定性、可靠性、数据处理能力或服务能力。例如，可以减少单点故障、共享存储资源、负荷分担或提高系统性能等。

实例

当一个服务的角色安装到主机上，即形成一个实例。每个服务有各自对应的角色实例。

元数据

元数据（Metadata），又称中介数据、中继数据，为描述数据的数据（data about data），主要是描述数据属性（property）的信息，用来支持如指示存储位置、历史数据、资源查找、文件纪录等功能。

2 入门

2.1 如何使用 MRS

MRS是一个在云上部署和管理Hadoop系统的服务，一键即可部署Hadoop集群。MRS提供租户完全可控的企业级大数据集群云服务，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件。

MRS使用简单，通过使用在集群中连接在一起的多台计算机，您可以运行各种任务，处理或者存储（PB级）巨量数据。MRS的基本使用流程如下：

1. 上传程序和数据文件到对象存储服务（OBS）中，用户需要先将本地的程序和数据文件上传至OBS中。
2. **创建自定义集群**，用户可以指定集群类型用于离线数据分析和流处理任务，指定集群中预置的弹性云服务器实例规格、实例数量、数据盘类型（普通IO、高 IO、超高 IO）、要安装的组件（Hadoop、Spark、HBase、Hive、Kafka、Storm等）。用户可以使用**引导操作**在集群启动前（或后）在指定的节点上执行脚本，安装其他第三方软件或修改集群运行环境等自定义操作。
3. **管理作业**，MRS为用户提供程序执行平台，程序由用户自身开发，MRS负责程序的提交、执行和监控。
4. **管理集群**，MRS为用户提供企业级的大数据集群的统一管理平台，帮助用户快速掌握服务及主机的健康状态，通过图形化的指标监控及定制及时的获取系统的关键信息，根据实际业务的性能需求修改服务属性的配置，对集群、服务、角色实例等实现一键启停等操作。
5. **删除集群**，如果作业执行结束后不需要集群，可以删除MRS集群。

2.2 创建集群

使用MRS的首要操作就是集群，本章节为您介绍如何在MRS管理控制台创建一个新的集群。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 。

> **说明**
>
> 创建集群时需要注意配额提醒。当资源配额不足时，建议按照提示申请足够的资源，再创建集群。

步骤3 在集群页面，选择“自定义创建”页签。

步骤4 配置集群软件信息。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：默认最新版本即可。
- 集群类型：默认选择“分析集群”即可。
- 组件选择：分析集群勾选Spark2x、HBase和Hive等组件。流式集群勾选Kafka和Storm等组件。混合集群可同时勾选分析集群流式集群的组件。
- 元数据：默认即可。

> **说明**
>
> 针对MRS 3.x之前版本，分析集群勾选Spark、HBase和Hive等组件。

步骤5 单击“下一步”。

- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 安全组：选择“自动创建”。
- 弹性公网IP：选择“暂不绑定”。
- 企业项目：默认即可。
- 实例规格：Master和Core节点都选择“通用计算型S3->8核 16GB(s3.2xlarge.2)”。
- 系统盘：存储类型选择“普通IO”，存储空间默认即可。
- 数据盘：存储类型选择“普通IO”，存储空间默认即可，数据盘数量默认即可。
- 实例数量：Master节点数量默认为2，Core节点数量配置为3。

步骤6 单击“下一步”进入高级配置页签，配置参数，其他参数保持默认。

- Kerberos认证：
  - Kerberos认证：关闭Kerberos认证。
  - 用户名：Manager管理员用户，目前默认为**admin**用户。
  - 密码：Manager管理员用户的密码。
- 登录方式：选择登录ECS节点的登录方式。
  - 密码：设置登录ECS节点的登录密码。
  - 密钥对：从下拉框中选择密钥对，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。

- 通信安全授权：勾选确认授权。

步骤7 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤8 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

----结束

2.3 上传示例数据和程序

用户通过“文件管理”页面可以在分析集群进行文件夹创建、删除，文件导入、导出、删除操作。

背景信息

MRS集群处理的数据源来源于OBS或HDFS，OBS为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用。

导入数据

MRS目前只支持将OBS上的数据导入至HDFS中。上传文件速率会随着文件大小的增大而变慢，适合数据量小的场景下使用。

支持导入文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。

   “bd_app1”目录仅为示例，可以是界面上的任何目录，也可以通过“新建”创建新的文件夹。

   新建文件夹时需要满足以下要求：

   - 文件夹名称小于等于255字符。
   - 不允许为空。
   - 不能包含：/:*?"<>|\;&,'`!{}[]$%+特殊字符。
   - 不能以“.”开头或结尾。
   - 开头和末尾的空格会被忽略。
6. 单击“导入数据”，正确配置HDFS和OBS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。
   - OBS路径

- 必须以“obs://”开头。
- 不支持导入KMS加密的文件或程序。
- 不支持导入空的文件夹。
- 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- OBS全路径长度小于等于255字符。

– HDFS路径

- 默认以“/user”开头。
- 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- HDFS全路径长度小于等于255字符。

7. 单击“确定”。

   文件上传进度可在“文件操作记录”中查看。MRS将数据导入操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

导出数据

数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。

支持导出文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群基本信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。
6. 单击“导出数据”，配置OBS和HDFS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。

   – OBS路径

   - 必须以“obs://”开头。
   - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
   - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
   - OBS全路径长度小于等于255字符。

   – HDFS路径

- 默认以“/user”开头。
- 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- HDFS全路径长度小于等于255字符。

说明

当导出文件夹到OBS系统时，在OBS路径下，将增加一个标签文件，文件命名为“folder name_$folder$”。请确保导出的文件夹为非空文件夹，如果导出的文件夹为空文件夹，OBS无法显示该文件夹，仅生成一个命名为“folder name_$folder$”的文件。

7. 单击“确定”。

文件上传进度可在“文件操作记录”中查看。MRS将数据导出操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

2.4 添加作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。

本章节以MapReduce作业为例指导您在MRS集群页面如何提交一个新的作业。MapReduce作业用于提交jar程序快速并行处理大量数据，是一种分布式数据处理模式和执行环境。

若在集群详情页面不支持“作业管理”和“文件管理”功能，请通过后台功能来提交作业。

用户创建作业前需要将本地数据上传至OBS系统用于计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，并使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

说明

IAM用户名存在空格时（如admin 01），不支持添加作业。

步骤6 “作业类型”选择“MapReduce”，并配置其他作业信息。

表 2-1 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。
执行程序参数	可选参数，程序执行的关键参数。多个参数间使用空格隔开。 配置方法：*程序类名 数据输入路径 数据输出路径* ● 程序类名：由用户程序内的函数指定，MRS只负责参数的传入。 ● 数据输入路径：通过单击“HDFS”或者“OBS”选择或者直接手动输入正确路径。 ● 数据输出路径：输出路径请手动输入一个不存在的目录。最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。

参数	参数说明
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表2-2**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 2-2 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在“节点管理”页签中单击某一Master节点名称，进入弹性云服务器管理控制台。

步骤4 单击页面右上角的“远程登录”。

步骤5 根据界面提示，输入Master节点的用户名和密码，用户名、密码分别为root和创建集群时设置的密码。

步骤6 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤7 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证，则无需执行该步骤。

kinit *MRS集群用户*

例如, **kinit admin**

步骤8 执行如下命令拷贝OBS文件系统中的程序到集群的Master节点。

hadoop fs -Dfs.obs.access.key=AK -Dfs.obs.secret.key=SK -copyToLocal source_path.jar target_path.jar

例如：**hadoop fs -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX -copyToLocal "obs://mrs-word/program/hadoop-mapreduce-examples-XXX.jar" "/home/omm/hadoop-mapreduce-examples-XXX.jar"**

AK/SK可登录OBS控制台，请在集群控制台页面右上角的用户名下拉框中选择“我的凭证 > 访问密钥”页面获取。

步骤9 执行如下命令提交wordcount作业，如需从OBS读取或向OBS输出数据，需要增加AK/SK参数。

source /opt/Bigdata/client/bigdata_env;hadoop jar execute_jar wordcount input_path output_path

例如：**source /opt/Bigdata/client/bigdata_env;hadoop jar /home/omm/hadoop-mapreduce-examples-XXX.jar wordcount -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX "obs://mrs-word/input/*" "obs://mrs-word/output/"**

input_path为OBS上存放作业输入文件的路径。output_path为OBS上存放作业输出文件地址，请设置为一个不存在的目录。

----结束

2.5 快速使用 Kerberos 认证集群

本章节提供从零开始使用安全集群并执行MapReduce程序、Spark程序和Hive程序的操作指导。

MRS 3.x版本Presto组件暂不支持开启Kerberos认证。

本指导的基本内容如下所示：

1. **创建安全集群并登录其Manager**
2. **创建角色和用户**
3. **执行MapReduce程序**
4. **执行Spark程序**
5. **执行Hive程序**

创建安全集群并登录其 Manager

步骤1 创建安全集群，请参见**创建自定义集群**页面，开启“Kerberos认证”参数开关，并配置“密码”、“确认密码”参数。该密码用于登录Manager，请妥善保管。

步骤2 登录MRS管理控制台页面。

步骤3 单击“集群列表”，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤4 单击“集群管理页面”后的“前往Manager”，打开Manager页面。

- 若用户创建集群时已经绑定弹性公网IP。
  a. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问Manager服务。

b. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

- 若用户创建集群时暂未绑定弹性公网IP。

a. 在弹性公网IP下拉框中选择可用的弹性公网IP或单击“管理弹性公网IP”创建弹性公网IP。

b. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请“管理安全组规则”。

说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问Manager服务。

c. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

步骤5 单击“确定”，进入Manager登录页面，如需给其他用户开通访问Manager的权限，请参见**访问MRS Manager（MRS 2.x及之前版本）**章节，添加对应用户访问公网的IP地址为可信范围。

步骤6 输入创建集群时默认的用户名“admin”及设置的密码，单击“登录”进入Manager页面。

----结束

创建角色和用户

开启Kerberos认证的集群，必须通过以下步骤创建一个用户并分配相应权限来允许用户执行程序。

步骤1 在Manager界面选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，详情请参见**创建角色**。

填写如下信息：

- 填写角色的名称，例如mrrole。
- 在“配置资源权限”选择待操作的集群，然后选择“Yarn > 调度队列 > root”，勾选“权限”列中的“提交”和“管理”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。
- 选择“HBase > HBase Scope”，勾选global的“权限”列的“创建”、“读”、“写”和“执行”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。
- 选择“HDFS > 文件系统 > hdfs://hacluster/”，勾选“权限”列的“读”、“写”和“执行”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。

- 选择“Hive > Hive读写权限”，勾选“权限”列的“查询”、“删除”、“插入”和“建表”，单击“确定”，完成角色的创建。

步骤3 选择“系统 > 权限 > 用户组 > 添加用户组”，为样例工程创建一个用户组，例如mrgroup，详情请参见**创建用户组**。

步骤4 选择“系统 > 权限 > 用户 > 添加用户”，为样例工程创建一个用户，详情请参见**创建用户**。

- 填写用户名，例如**test**，当需要执行Hive程序时，请设置用户名为“hiveuser”。
- 用户类型为“人机”用户。
- 输入密码（特别注意该密码在后面运行程序时要用到）。
- 加入用户组mrgroup和supergroup。
- 设置其“主组”为supergroup，并绑定角色mrrole取得权限。

  单击“确定”完成用户创建。

步骤5 选择“系统 > 权限 > 用户”，选择新建用户**test**，选择“更多 > 下载认证凭据”，保存后解压得到用户的keytab文件与krb5.conf文件。

----结束

执行 MapReduce 程序

本小节提供执行MapReduce程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如mapreduce-examples-1.0.jar、input_data1.txt和input_data2.txt。

操作步骤

步骤1 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

步骤2 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

步骤3 使用上传工具（比如：WinScp）将mapreduce-examples-1.0.jar、input_data1.txt和input_data2.txt复制到test目录下，将“创建角色和用户”中的步骤**步骤5**获得的keytab文件和krb5.conf文件复制到conf目录。

步骤4 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回（首次登录需按照系统提示修改密码），则完成了用户的kerberos认证。

步骤5 执行如下命令将数据导入到HDFS中：

```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

步骤6 执行如下命令运行程序：

```
yarn jar mapreduce-examples-1.0.jar com.xxx.bigdata.mapreduce.examples.FemaleInfoCollector /tmp/
input /tmp/mapreduce_output
```

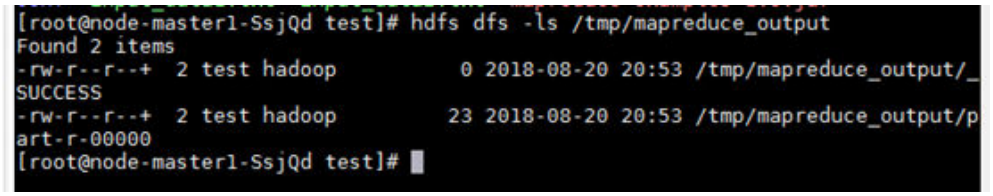
其中：

/tmp/input指HDFS文件系统中input的路径。

/tmp/mapreduce_output指HDFS文件系统中output的路径，该目录必须不存在，否则会报错。

步骤7 程序运行成功后，执行 hdfs dfs -ls /tmp/mapreduce_output会显示如下：

图 2-1 查看程序运行结果



----结束

执行 Spark 程序

本小节提供执行Spark程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如FemaleInfoCollection.jar、input_data1.txt和input_data2.txt。

操作步骤

步骤1 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

步骤2 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

步骤3 使用上传工具（比如：WinScp）将样FemaleInfoCollection.jar、input_data1.txt和input_data2.txt复制到test目录下，将“创建角色和用户”中的步骤**步骤5**获得的keytab文件和krb5.conf文件复制到conf目录。

步骤4 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回，则完成了用户的kerberos认证。

步骤5 执行如下命令将数据导入到HDFS中：

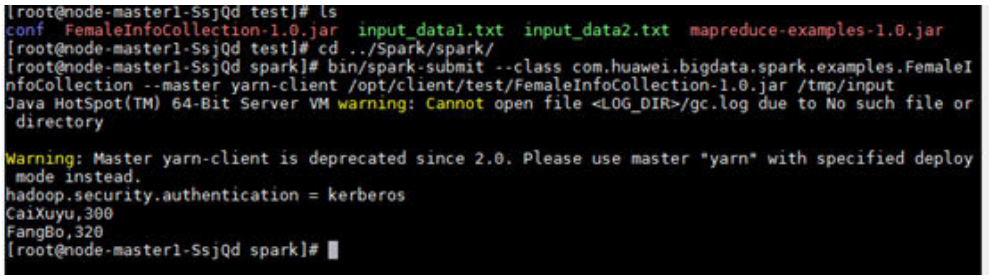
```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

步骤6 执行如下命令运行程序：

```
cd /opt/Bigdata/client/Spark/spark
bin/spark-submit --class com.xxx.bigdata.spark.examples.FemaleInfoCollection --master yarn-client /opt/Bigdata/client/test/FemaleInfoCollection-1.0.jar /tmp/input
```

步骤7 程序运行成功后，会显示如下：

图 2-2 程序运行结果



----结束

执行 Hive 程序

本小节提供执行Hive程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如hive-examples-1.0.jar、input_data1.txt和input_data2.txt。

操作步骤

步骤1 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

步骤2 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

步骤3 使用上传工具（比如：WinScp）将样FemaleInfoCollection.jar、input_data1.txt和input_data2.txt复制到test目录下，将“创建角色和用户”中的步骤**步骤5**获得的keytab文件和krb5.conf文件复制到conf目录。

步骤4 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回，则完成了用户的kerberos认证。

步骤5 执行如下命令运行程序：

```
chmod +x /opt/hive_examples -R   cd /opt/hive_examples   java -cp .:hive-examples-1.0.jar:/opt/
hive_examples/conf:/opt/Bigdata/client/Hive/Beeline/lib/*:/opt/Bigdata/client/HDFS/hadoop/lib/*
com.xxx.bigdata.hive.example.ExampleMain
```

步骤6 程序运行成功后，会显示如下：

图 2-3 程序运行的结果

```
[root@node-master1-iYpxp hive_examples]# java -cp .:hive-examples-mrs-1.7.0.jar:
/opt/hive_examples/conf:/opt/client/Hive/Beeline/lib/*:/opt/client/HDFS/hadoop/l
ib/* com.huawei.bigdata.hive.example.ExampleMain
log4j:WARN No appenders could be found for logger (com.huawei.bigdata.security.L
oginUtil).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Create table success!
_c0
0
Delete table success!
[root@node-master1-iYpxp hive_examples]#
```

----结束

2.6 删除集群

如果作业执行结束后不需要集群， 可以删除MRS集群。

背景信息

一般在数据完成分析和存储后或集群异常无法提供服务时才执行集群删除操作。当MRS集群部署失败时，集群会被自动删除。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“现有集群”。

步骤3 在需要删除的集群对应的“操作”列中，单击“删除”。

集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。

说明

当集群已对接了OBS（存算分离或者冷热分离场景），若需要删除组件或者MRS集群，需要在删除组件或者集群后，手工将OBS上相关的业务数据进行删除。

----结束

3 准备用户

3.1 创建 MRS 操作用户

如果您需要对您所拥有的MapReduce服务（MapReduce Service）进行精细的权限管理，您可以使用统一身份认证服务（Identity and Access Management，简称IAM），通过IAM，您可以：

- 根据企业的业务组织，在您的云帐号中，给企业中不同职能部门的员工创建IAM用户，让员工拥有唯一安全凭证，并使用MRS资源。
- 根据企业用户的职能，设置不同的访问权限，以达到用户之间的权限隔离。
- 将MRS资源委托给更专业、高效的其他云帐号或者云服务，这些帐号或者云服务可以根据权限进行代运维。

如果云帐号已经能满足您的要求，不需要创建独立的IAM用户，您可以跳过本章节，不影响您使用MRS服务的其它功能。

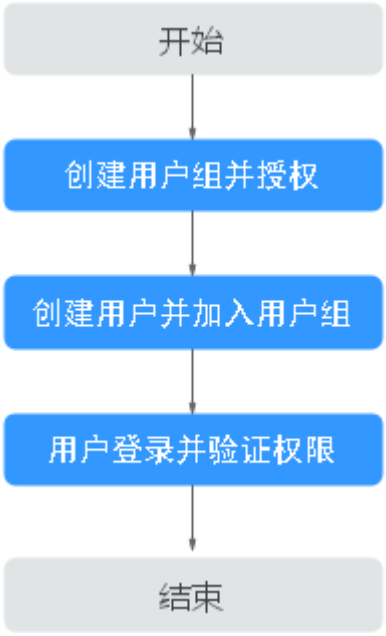
本章节为您介绍对用户授权的方法，操作流程如**图3-1**所示。

前提条件

给用户组授权之前，请您了解用户组可以添加的MRS权限，并结合实际需求进行选择。

示例流程

图 3-1 给用户授权 MRS 权限流程



1. 创建用户组并授权

   在IAM控制台创建用户组，并授予MRS服务对应权限。

2. 创建用户并加入用户组

   在IAM控制台创建用户，并将其加入**1.创建用户组并授权**中创建的用户组。

3. 并验证权限

   新创建的用户登录控制台，切换至授权区域，验证权限：

   - 在“服务列表”中选择MRS服务，进入MRS主界面，单击右上角“创建集群”，尝试MRS集群，如果无法MRS集群（假设当前权限仅包含MRS ReadOnlyAccess），表示“MRS ReadOnlyAccess”已生效。
   - 在“服务列表”中选择除MRS服务外（假设当前策略仅包含MRS ReadOnlyAccess）的任一服务，若提示权限不足，表示“MRS ReadOnlyAccess”已生效。

MRS 权限说明

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

MRS部署时通过物理区域划分，为项目级服务。授权时，“作用范围”需要选择“区域级项目”，然后在指定区域对应的项目中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问MRS时，需要先切换至授权区域。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于各服务之间存在业

务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。

- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对MRS服务，管理员能够控制IAM用户仅能对集群进行指定的管理操作。如不允许某用户组删除集群，仅允许操作MRS集群基本操作，如创建集群、查询集群列表等。多数细粒度策略以API接口为粒度进行权限拆分。

如表3-1所示，包括了MRS的所有系统策略。

表 3-1 MRS 系统策略

策略名称	描述	策略类别
MRS FullAccess	MRS管理员权限，拥有该权限的用户可以拥有MRS所有权限。	细粒度策略
MRS CommonOperations	MRS服务普通用户权限，拥有该权限的用户可以拥有MRS服务使用权限，无新增、删除资源权限。	细粒度策略
MRS ReadOnlyAccess	MRS服务只读权限，拥有该权限的用户仅能查看MRS的资源。	细粒度策略
MRS Administrator	操作权限： • 对MRS服务的所有执行权限。 • 拥有该权限的用户必须同时拥有Tenant Guest、Server Administrator和BSS Administrator权限。	RBAC策略

表3-2列出了MRS常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 3-2 常用操作与系统策略的授权关系

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
创建集群	√	x	x	√
调整集群	√	x	x	√
升级节点规格	√	x	x	√
删除集群	√	x	x	√

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
查询集群详情	√	√	√	√
查询集群列表	√	√	√	√
设置弹性伸缩策略	√	x	x	√
查询主机列表	√	√	√	√
查询操作日志	√	√	√	√
创建并执行作业	√	√	x	√
停止作业	√	√	x	√
删除单个作业	√	√	x	√
批量删除作业	√	√	x	√
查询作业详情	√	√	√	√
查询作业列表	√	√	√	√
新建文件夹	√	√	x	√
删除文件	√	√	x	√
查询文件列表	√	√	√	√
批量操作集群标签	√	√	x	√
创建单个集群标签	√	√	x	√
删除单个集群标签	√	√	x	√
按照标签查询资源列表	√	√	√	√
查询集群标签	√	√	√	√

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
访问Manager页面	√	√	x	√
查询补丁列表	√	√	√	√
安装补丁	√	√	x	√
卸载补丁	√	√	x	√
运维通道授权	√	√	x	√
运维通道日志共享	√	√	x	√
查询告警列表	√	√	√	√
订阅告警消息提醒	√	√	x	√
提交SQL语句	√	√	x	√
查询SQL结果	√	√	x	√
取消SQL执行任务	√	√	x	√

3.2 创建 MRS 自定义策略

如果系统预置的MRS权限，不满足您的授权要求，可以创建自定义策略。

目前支持以下两种方式创建自定义策略：

- 可视化视图创建自定义策略：无需了解策略语法，按可视化视图导航栏选择云服务、操作、资源、条件等策略内容，可自动生成策略。
- JSON视图创建自定义策略：可以在选择策略模板后，根据具体需求编辑策略内容；也可以直接在编辑框内编写JSON格式的策略内容。

本章为您介绍常用的MRS自定义策略样例。

MRS 自定义策略样例

- 示例1：授权用户仅有创建MRS集群的权限

```
{
    "Version": "1.1",
    "Statement": [
```

```
        {
            "Effect": "Allow",
            "Action": [
                "mrs:cluster:create",
                "ecs:*:*",
                "bms:*:*",
                "evs:*:*",
                "vpc:*:*",
                "smn:*:*"
            ]
        }
    ]
}
```

- 示例2：授权用户调整MRS集群

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "mrs:cluster:resize"
            ]
        }
    ]
}
```

- 示例3：授权用户创建集群、创建并执行作业、删除单个作业，但不允许用户删除集群的权限

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "mrs:cluster:create",
                "mrs:job:submit",
                "mrs:job:delete"
            ]
        },
        {
            "Effect": "Deny",
            "Action": [
                "mrs:cluster:delete"
            ]
        }
    ]
}
```

- 示例4：授权用户最小权限，创建ECS规格的集群

  说明

  - 创建集群时如果使用密钥对，增加权限：ecs:serverKeypairs:get和ecs:serverKeypairs:list
  - 创集群时使用数据盘加密，增加权限：kms:cmk:list
  - 创建集群时开启告警功能，增加权限：mrs:alarm:subscribe
  - 创建集群时使用外置数据源，增加权限：rds:instance:list

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "mrs:cluster:create"
            ]
        },
        {
```

```
            "Effect": "Allow",
            "Action": [
                "ecs:cloudServers:updateMetadata",
                "ecs:cloudServerFlavors:get",
                "ecs:cloudServerQuotas:get",
                "ecs:servers:list",
                "ecs:servers:get",
                "ecs:cloudServers:delete",
                "ecs:cloudServers:list",
                "ecs:serverInterfaces:get",
                "ecs:serverGroups:manage",
                "ecs:servers:setMetadata",
                "ecs:cloudServers:get",
                "ecs:cloudServers:create"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "vpc:securityGroups:create",
                "vpc:securityGroupRules:delete",
                "vpc:vpcs:create",
                "vpc:ports:create",
                "vpc:securityGroups:get",
                "vpc:subnets:create",
                "vpc:privateIps:delete",
                "vpc:quotas:list",
                "vpc:networks:get",
                "vpc:publicIps:list",
                "vpc:securityGroups:delete",
                "vpc:securityGroupRules:create",
                "vpc:privateIps:create",
                "vpc:ports:get",
                "vpc:ports:delete",
                "vpc:publicIps:update",
                "vpc:subnets:get",
                "vpc:publicIps:get",
                "vpc:ports:update",
                "vpc:vpcs:list"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "evs:quotas:get",
                "evs:types:get"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "bms:serverFlavors:get"
            ]
        }
    ]
}
```

- 示例5：授权用户最小权限，创建BMS规格的集群

  说明

  - 创建集群时如果使用密钥对，增加权限：ecs:serverKeypairs:get和ecs:serverKeypairs:list
  - 创集群时使用数据盘加密，增加权限：kms:cmk:list
  - 创建集群时开启告警功能，增加权限：mrs:alarm:subscribe
  - 创建集群时使用外置数据源，增加权限：rds:instance:list

```
{
    "Version": "1.1",
```

```
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "mrs:cluster:create"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "ecs:servers:list",
        "ecs:servers:get",
        "ecs:cloudServers:delete",
        "ecs:serverInterfaces:get",
        "ecs:serverGroups:manage",
        "ecs:servers:setMetadata",
        "ecs:cloudServers:create",
        "ecs:cloudServerFlavors:get",
        "ecs:cloudServerQuotas:get"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "vpc:securityGroups:create",
        "vpc:securityGroupRules:delete",
        "vpc:vpcs:create",
        "vpc:ports:create",
        "vpc:securityGroups:get",
        "vpc:subnets:create",
        "vpc:privateIps:delete",
        "vpc:quotas:list",
        "vpc:networks:get",
        "vpc:publicIps:list",
        "vpc:securityGroups:delete",
        "vpc:securityGroupRules:create",
        "vpc:privateIps:create",
        "vpc:ports:get",
        "vpc:ports:delete",
        "vpc:publicIps:update",
        "vpc:subnets:get",
        "vpc:publicIps:get",
        "vpc:ports:update",
        "vpc:vpcs:list"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "evs:quotas:get",
        "evs:types:get"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "bms:servers:get",
        "bms:servers:list",
        "bms:serverQuotas:get",
        "bms:servers:updateMetadata",
        "bms:serverFlavors:get"
      ]
    }
  ]
}
```

- 示例6：授权用户最小权限，创建ECS和BMS混合集群

说明

- 创建集群时如果使用密钥对，增加权限：ecs:serverKeypairs:get和ecs:serverKeypairs:list
- 创集群时使用数据盘加密，增加权限：kms:cmk:list
- 创建集群时开启告警功能，增加权限：mrs:alarm:subscribe
- 创建集群时使用外置数据源，增加权限：rds:instance:list

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "mrs:cluster:create"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "ecs:cloudServers:updateMetadata",
        "ecs:cloudServerFlavors:get",
        "ecs:cloudServerQuotas:get",
        "ecs:servers:list",
        "ecs:servers:get",
        "ecs:cloudServers:delete",
        "ecs:cloudServers:list",
        "ecs:serverInterfaces:get",
        "ecs:serverGroups:manage",
        "ecs:servers:setMetadata",
        "ecs:cloudServers:get",
        "ecs:cloudServers:create"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "vpc:securityGroups:create",
        "vpc:securityGroupRules:delete",
        "vpc:vpcs:create",
        "vpc:ports:create",
        "vpc:securityGroups:get",
        "vpc:subnets:create",
        "vpc:privateIps:delete",
        "vpc:quotas:list",
        "vpc:networks:get",
        "vpc:publicIps:list",
        "vpc:securityGroups:delete",
        "vpc:securityGroupRules:create",
        "vpc:privateIps:create",
        "vpc:ports:get",
        "vpc:ports:delete",
        "vpc:publicIps:update",
        "vpc:subnets:get",
        "vpc:publicIps:get",
        "vpc:ports:update",
        "vpc:vpcs:list"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "evs:quotas:get",
        "evs:types:get"
      ]
    },
    {
      "Effect": "Allow",
```

```
            "Action": [
                "bms:servers:get",
                "bms:servers:list",
                "bms:serverQuotas:get",
                "bms:servers:updateMetadata",
                "bms:serverFlavors:get"
            ]
        }
    ]
}
```

3.3 IAM 用户同步 MRS 说明

IAM用户同步是指将绑定MRS相关策略的IAM用户同步至MRS系统中，创建同用户名、不同密码的帐号，用于集群管理。同步之后，用户可以使用IAM用户名（密码需要Manager的管理员admin重置后方可使用）登录Manager管理集群。也可以在开启Kerberos认证的集群中，通过界面方式提交作业。

IAM用户权限策略及同步MRS后权限对比请参考**表3-3**，Manager对应默认权限说明请参考**MRS集群中的用户与权限**。

表 3-3 IAM 权限策略与 MRS 权限同步映射

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
细粒度	MRS ReadOnlyAccess	Manager_viewer	否	否
	MRS CommonOperations	• Manager_viewer • default • launcher-job	否	是

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
	MRS FullAccess	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	是	是
RBAC	MRS Administrator	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	否	是

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
	Server Administrator、Tenant Guest和MRS Administrator	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	是	是
	Tenant Administrator	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	是	是

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
自定义	Custom policy（自定义策略）	• Manager_viewer • default • launcher-job	• 自定义策略以RBAC策略为模板则参考RBAC策略。 • 自定义策略以细粒度策略为模板则参考细粒度策略，建议使用细粒度策略。	是

说明

为了更方便进行用户权限管理，请尽可能使用细粒度策略，减少RBAC策略的使用，细粒度策略判断action时以deny优先原则。

- 只有具有Tenant Administrator或同时具有Server Administrator、Tenant Guest、MRS Administrator角色才在MRS集群中拥有同步IAM用户的权限。
- 只要拥有**action:mrs:cluster:syncUser**策略就在MRS集群中拥有同步IAM用户的权限。

操作步骤

步骤1 创建用户并授权使用MRS服务，具体请参考**创建MRS操作用户**。

步骤2 登录MRS控制台并创建集群，具体请参考**创建自定义集群**。

步骤3 在左侧导航栏中选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤4 在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

步骤5 同步请求下发后，返回MRS控制台在左侧导航栏中选择“操作日志”页面查看同步是否成功，日志相关说明请参考**查看MRS服务操作日志**。

步骤6 同步成功后，即可使用IAM同步用户进行后续操作。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。
- 单击“IAM用户同步”右侧的“同步”后，集群详情页面会出现短时间空白，这是由于正在进行用户数据同步中，请耐心等待，数据同步完成后，页面将会正常显示。

- 安全集群提交作业：安全集群中用户可通过界面“作业管理”功能提交作业，具体请参考**运行MapReduce作业**。
- 集群详情页面页签显示完整（包含“组件管理”，“租户管理”和“备份恢复”）。
- 登录Manager页面。
  a. 使用admin帐号登录Manager，具体请参考**访问集群Manager**。
  b. 初始化IAM同步用户密码，具体请参考**初始化系统用户密码**。
  c. 修改用户所在用户组绑定的角色，精确控制Manager下用户权限，具体请参考**相关任务**修改用户组绑定的角色，如需创建修改角色请参考**创建角色**。用户所在用户组绑定的组件角色修改后，权限生效需要一定时间，请耐心等待。
  d. 使用IAM同步用户及**步骤6.b**初始化后的密码登录Manager。

  说明

  当IAM用户权限发生变化时，需要执行**步骤4**进行二次同步。对于系统用户，二次同步后用户的权限为IAM系统策略定义的权限和用户在Manager自行添加角色的权限的并集。对于自定义用户，二次同步后用户的权限以Manager配置的权限为准。
  - 系统用户：如果IAM用户所在用户组全部都绑定系统策略（RABC策略和细粒度策略均属于系统策略），则该用户为系统用户。
  - 自定义用户：如果IAM用户所在用户组只要有绑定任何自定义策略，则该用户为自定义用户。

----结束

4 配置集群

4.1 创建方式简介

本节介绍MRS服务的方式。

- **快速创建Hadoop分析集群**：快速Hadoop分析集群为您提高了配置效率，可以在几分钟之内快速创建Hadoop集群，更加方便快捷的进行海量数据分析与查询。
- **快速创建HBase查询集群**：快速HBase查询集群为您提高了配置效率，可以在几分钟之内快速创建HBase集群，更加方便快捷的进行海量数据存储以及分布式计算。
- **快速创建Kafka流式集群**：快速Kafka流式集群为您提高了配置效率，可以在几分钟之内快速创建Kafka集群，更加方便快捷的进行流式数据采集，实时数据处理存储等。
- **快速创建ClickHouse集群**：快速一个ClickHouse集群，ClickHouse是一个用于联机分析的列式数据库管理系统，具有压缩率和极速查询性能。
- **快速创建实时分析集群**：快速一个实时分析集群为您提高了配置效率，可以在几分钟之内快速创建实时分析集群，更加方便快捷的进行海量的数据采集、数据的实时分析和查询。
- **创建自定义集群**：自定义可以灵活地选择配置项，针对不同的应用场景，可以选择不同规格的弹性云服务器，全方位贴合您的业务诉求。

4.2 快速创建集群

4.2.1 快速创建 Hadoop 分析集群

本章节为您介绍如何快速一个Hadoop分析集群，Hadoop完全使用开源Hadoop生态，采用YARN管理集群资源，提供Hive、Spark离线大规模分布式数据存储和计算，SparkStreaming、Flink流式数据计算，Presto交互式查询，Tez有向无环图的分布式计算框架等Hadoop生态圈的组件，进行海量数据分析与查询。

快速创建 Hadoop 分析集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“Hadoop分析集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.2 快速创建 HBase 查询集群

本章节为您介绍如何快速一个HBase查询集群，HBase集群使用Hadoop和HBase组件提供一个稳定可靠，性能卓越、可伸缩、面向列的分布式云存储系统，适用于海量数据存储以及分布式计算的场景，用户可以利用HBase搭建起TB至PB级数据规模的存储系统，对数据轻松进行过滤分析，毫秒级得到响应，快速发现数据价值。

快速创建 HBase 查询集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“HBase查询集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.3 快速创建 Kafka 流式集群

本章节为您介绍如何快速一个Kafka流式集群，Kafka集群使用Kafka和Storm组件提供一个开源高吞吐量，可扩展性的消息系统。广泛用于日志收集、监控数据聚合等场景，实现高效的流式数据采集，实时数据处理存储等。

快速创建 Kafka 流式集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20200321”。
- 集群版本：不同版本集群提供的组件有所不同，请根据需要选择集群版本。
- 组件选择：选择“Kafka流式集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- LVM：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.4 快速创建 ClickHouse 集群

本章节为您介绍如何快速一个ClickHouse集群，ClickHouse是一个用于联机分析的列式数据库管理系统，具有压缩率和极速查询性能。被广泛的应用于互联网广告、App和Web流量、电信、金融、物联网等众多领域。

ClickHouse集群包含的组件：

- MRS 3.1.0版本：ClickHouse 21.3.4.25、ZooKeeper 3.5.6。

CPU架构为鲲鹏计算的ClickHouse集群表引擎不支持使用HDFS和Kafka。

快速创建 ClickHouse 集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20201121”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“ClickHouse集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.x版本无该参数。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.5 快速创建实时分析集群

本章节为您介绍如何快速一个实时分析集群，实时分析集群使用Hadoop、Kafka、Flink和ClickHouse组件提供一个海量的数据采集、数据的实时分析和查询的系统。

集群包含的组件信息实时分析：

- MRS 3.1.0版本：Hadoop 3.1.1、Kafka 2.11-2.4.0、Flink 1.12.0、ClickHouse 21.3.4.25、ZooKeeper 3.5.6、Ranger 2.0.0。

快速创建实时分析集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20201130”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“实时分析集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.3 创建自定义集群

使用MRS的首要操作就是集群，本章节为您介绍如何在MRS管理控制台自定义创建一个新的MRS集群。

注册帐号后，如果需要对云上的资源进行精细管理，请使用IAM服务创建IAM用户及用户组，并授权，以使得IAM用户获得具体的操作权限，具体请参考**创建MRS操作用户**。

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

说明

创建集群时需要注意配额提醒。当资源配额不足时，建议按照提示申请足够的资源，再创建集群。

步骤3 在集群页面，选择“自定义创建”页签。

步骤4 参考**软件配置**配置集群信息后，单击“下一步”。

步骤5 参考**硬件配置**配置集群信息后，单击“下一步”。

步骤6 参考**高级配置（可选）**配置集群信息后，单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

软件配置

表 4-1 MRS 集群软件配置

参数	参数说明
区域	选择区域。 不同区域的云服务产品之间内网互不相通。请就近选择靠近您业务的区域，可减少网络时延，提高访问速度。
集群名称	集群名称不允许重复。 只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 默认名称为mrs_xxxx，xxxx为字母和数字的四位随机组合数，系统自动组合。
集群版本	目前支持MRS 2.1.1、MRS 3.0.5版本。
集群类型	提供几种集群类型： ● 分析集群：用来做离线数据分析，提供的是Hadoop体系的组件。 ● 流式集群：用来做流处理任务，提供的是流式处理组件。 ● 混合集群：既可以用来做离线数据分析，也可以用来做流处理任务，提供的是Hadoop体系的组件和流式处理组件。建议同时需要做离线数据分析和流处理任务时使用混合集群。 ● 自定义：用户可按照业务需求调整集群服务的部署方式，具体请参见**创建自定义拓扑集群**。（目前仅MRS 3.x版本支持） **说明** ● MRS流式集群不支持“作业管理”和“文件管理”功能。 ● 如需在集群中安装全部组件，请选择“自定义”类型集群。

参数	参数说明
组件选择	MRS配套的组件如下： 分析集群组件 ● Presto：开源、分布式SQL查询引擎。 ● Hadoop：分布式系统基础架构。 ● Spark：内存分布式系统框架。（ MRS 3.x版本不支持 ） ● Spark2x：Spark2x是一个对大规模数据处理的快速和通用引擎,基于开源Spark2.x版本开发。（ 仅MRS 3.x版本支持 ） ● Hive：建立在Hadoop上的数据仓库框架。 ● HBase：分布式列数据库。 ● Tez：提供有向无环图的分布式计算框架。 ● Hue：提供Hadoop UI能力，让用户通过浏览器分析处理Hadoop集群数据。 ● Loader：基于开源sqoop 1.99.7开发，专为Apache Hadoop和结构化数据库（如关系型数据库）设计的高效传输大量数据的工具。（ MRS 3.x版本不支持 ） Hadoop为必选组件，且Spark与Hive组件需要配套使用。请根据业务选择搭配组件。 ● Flink：分布式大数据处理引擎，可对有限数据流和无限数据流进行有状态计算。 ● Oozie：Hadoop作业调度系统。（ 仅MRS 3.x版本支持 ） ● Alluxio：一个基于内存的分布式存储系统。 ● Ranger：一个基于Hadoop平台监控和管理数据安全的框架。 ● Impala：一种处理大量数据的SQL查询引擎。 ● ClickHouse：ClickHouse是一个用于联机分析(OLAP)的列式数据库管理系统(DBMS)。CPU架构为鲲鹏计算的ClickHouse集群表引擎不支持使用HDFS和Kafka。 ● Kudu：一种列存储管理器。 流式集群组件 ● Kafka：提供分布式消息订阅的系统。 ● Flume：提供分布式、高可用、高可靠的海量日志采集、聚合和传输系统。
元数据	是否使用外部数据源存储元数据。 ● 本地元数据： 元数据存储在集群本地。 ● 数据连接：使用外部数据源元数据，若集群异常或删除时将不影响元数据，适用于存储计算分离的场景。 支持Hive或Ranger组件的集群支持该功能。
组件名	当“元数据”选择“数据连接”时该参数有效。用于表示可以设置外部数据源的组件类型。MRS 3.x版本暂不支持该功能。 ● Hive ● Ranger

参数	参数说明
数据连接类型	当“元数据”选择“数据连接”时该参数有效。用于表示外部数据源的类型。 ● Hive组件支持的数据连接类型： – RDS服务PostgreSQL数据库 – RDS服务MySQL数据库 – 本地数据库 ● Ranger组件支持的数据连接类型： – RDS服务MySQL数据库 – 本地数据库
数据连接实例	当“数据连接类型”选择“RDS服务PostgreSQL数据库”或“RDS服务MySQL数据库”时，该参数有效。用于表示MRS集群与RDS服务数据库连接的名称，该实例必选先创建才能在此处引用。可单击“创建数据连接”进行创建，具体请参考**配置数据连接**。

硬件配置

表 4-2 MRS 集群硬件配置

参数	参数说明
可用区	选择集群工作区域下关联的可用区。 可用区是使用独立电源和网络资源的物理区域。通过内部网络互联，再以物理方式进行隔离，提高了应用程序的可用性。建议您在不同的可用区下创建集群。
虚拟私有云	VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。 选择需要创建集群的VPC，单击“查看虚拟私有云”进入VPC服务查看已创建的VPC名称和ID。如果没有VPC，需要创建一个新的VPC。
子网	通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。 选择需要创建集群的子网，单击“查看子网”可查看所选子网的详细信息，若VPC下未创建子网，请在VPC服务控制台单击“创建子网”进行创建。网络ACL出规则配置请参考**如何配置网络ACL出规则？** 。 **说明** 创建MRS 集群需要的IP数量和集群节点和组件个数相关，集群类型不影响IP数量。 MRS集群部署默认需要的IP数量为：集群节点数量+2（Manager+DB），如果部署集群时选择Hadoop、Hue、Sqoop或Loader、Presto组件，则每一个组件需要再加一个IP。若单独ClickHouse集群则需要的IP数量为：集群节点数量+1（Manager）。

参数	参数说明
安全组	安全组是一组对弹性云服务器的访问规则的集合，为同一个VPC内具有相同安全保护需求并相互信任的弹性云服务器提供访问策略。 用户创建集群时，可自动创建安全组，也可选择下拉框中已有的安全组。 **说明** 选择用户自己创建的安全组时，请确保入方向规则中有一条全部协议，全部端口，源地址为可信任的IP访问范围的规则，源地址请勿使用0.0.0.0/0，否则会有安全风险。若用户不清楚可信任的IP访问范围，请选择自动创建。
弹性公网IP	通过将弹性公网IP与MRS集群绑定，实现使用弹性公网IP访问Manager的目的。 用户创建集群时，可选择下拉框中已有的弹性公网IP进行绑定。若下拉框中没有可选的弹性公网IP，可以单击“管理弹性公网IP”进入弹性公网IP服务进行。 **说明** 弹性公网IP必须和集群在同一区域。
企业项目	选择集群所属的企业项目，如需使用企业项目，请先通过“企业 > 项目管理”服务创建。 企业项目所在的企业管理控制台以面向企业资源管理为出发点，帮助企业以公司、部门、项目等分级管理方式实现企业云上的人员、资源、权限、财务的管理。

表 4-3 集群节点信息

参数	参数说明
CPU架构	MRS提供的CPU架构类型。 • x86计算：x86 CPU架构采用复杂指令集（CISC），CISC指令集的每个小指令可以执行一些较低阶的硬件操作，指令数目多而且复杂，每条指令的长度并不相同。由于指令执行较为复杂所以每条指令花费的时间较长。 • 鲲鹏计算：鲲鹏CPU架构采用精简指令集（RISC），RISC是一种执行较少类型计算机指令的微处理器，它能够以更快的速度执行操作，使计算机的结构更加简单合理地提高运行速度，相对于x86 CPU架构具有更加均衡的性能功耗比。鲲鹏的优势是高密度低功耗，可以提供更高的性价比。
常用模板	当“集群类型”选择“自定义”时该参数有效，请参考**自定义集群模板说明**选择。

参数	参数说明
节点类型	MRS提供节点类型： ● Master：指集群Master节点，负责管理集群，协调将集群可执行文件分配到Core节点。此外，还会跟踪每个作业的执行状态，监控DataNode的运行状况。 ● Core：指集群Core节点，处理数据并在HDFS中存储过程数据。分析集群将创建分析Core节点，流式集群将创建流式Core节点，混合集群分别创建分析Core和流式Core节点。 ● Task：指集群Task节点，主要用于计算，不存放持久数据。主要安装Yarn、Storm组件。Task节点为可选节点，数目可以是零。分析集群将创建分析Task节点，流式集群将创建流式Task节点，混合集群分别创建分析Task和流式Task节点。 当集群数据量变化不大而集群业务处理能力需求变化比较大，大的业务处理能力只是临时需要，此时选择添加Task节点。 – 临时业务量增大，如年底报表处理。 – 需要在短时间内处理完原来需要处理很久的任务，如一些紧急分析任务。
实例规格	选择主节点和核心节点的实例规格。MRS当前支持主机规格的配型由CPU+内存+Disk共同决定。单击  ，配置集群节点的实例规格、系统盘和数据盘参数。 **说明** ● 节点的实例规格配置越高，数据处理分析能力越强。 ● 当Core节点规格选择非HDD磁盘时，Master节点和Core节点的磁盘类型取决于数据磁盘。 ● 当节点的实例规格选项后标示“已售罄”时，将无法此规格的节点，请选择其他规格节点进行。 ● MRS 3.x及之后版本集群Master节点规格不能小于64GB。
系统盘	节点系统盘的存储类型和存储空间。 存储类型： ● SATA：普通IO ● SAS：高IO ● SSD：超高IO ● GPSSD：通用型SSD

参数	参数说明
数据盘	节点数据磁盘存储空间。为增大数据存储容量，创建集群时可同时添加磁盘，有如下应用场景： ● 数据存储和计算分离，数据存储在OBS中，集群存储成本低，存储量不受限制，并且集群可以随时删除，但计算性能取决于OBS访问性能，相对HDFS有所下降，建议在数据计算不频繁场景下使用。 ● 数据存储和计算不分离，数据存储在HDFS中，集群成本较高，计算性能高，但存储量受磁盘空间限制，删除集群前需将数据导出保存，建议在数据计算频繁场景下使用。 目前的存储类型： ● SATA：普通IO ● SAS：高IO ● SSD：超高IO ● GPSSD：通用型SSD **说明** 创建的节点个数越多，对管理节点（即master节点）的硬盘容量要求越高。为了保证集群能够健康地运行，当创建的节点个数达到300时，建议将master的硬盘容量配置成600GB以上；当创建的节点个数达到500时，建议将master的硬盘容量配置成1TB以上。
实例数量	配置主节点和核心节点的个数。 Master： ● 开启“集群高可用”时，Master实例数量固定为2个。 ● 关闭“集群高可用”时，Master实例数量固定为1个。 Core节点至少存在一个，Core节点和Task节点的数量之和不能超过500个。 Task：单击添加Task节点。单击修改Task节点额实例规格和磁盘配置。单击删除已添加的Task节点。 **说明** ● Core节点默认的最大值为500，如果用户需要的Core节点数大于500，请联系技术支持人员。 ● 过小的节点容量会导致您的集群运行缓慢，而过大的节点容量会产生不必要的成本，请根据您要处理的数据对集群节点数量进行调整。
LVM	仅当创建流式Core节点时，该参数在流式Core节点有效。单击该参数以开启或关闭磁盘LVM管理。MRS 3.x及之后版本不支持该参数。 启用逻辑卷管理(LVM)时，会将节点中所有磁盘以逻辑卷的方式挂载，能够更加合理的规划磁盘，避免磁盘不均匀的问题，提升系统的稳定性。

参数	参数说明
拓扑调整	当常用模板中的部署方式不满足需求，请设置“拓扑调整”为“开启”，然后根据业务需要调整实例部署方式，具体说明请参见**自定义集群拓扑调整说明**。当集群类型为“自定义”时该参数有效。

高级配置（可选）

表 4-4 MRS 集群高级配置拓扑

参数	参数说明
标签	具体请参考**添加集群标签**。
主机名前缀	用作集群中ECS机器主机名的前缀。
弹性伸缩	请在“硬件配置”页签指定Task节点的规格，然后参考**配置弹性伸缩规则**配置。
引导操作	具体请参考**添加引导操作**。MRS 3.x版本暂时不支持该参数。
委托	通过绑定委托，ECS或BMS云服务将有权限来管理您的部分资源，请根据实际业务场景需求确认是否需要配置委托。 例如通过配置ECS委托可自动获取AK/SK访问OBS，具体请参见**配置存算分离集群（委托方式）**。 **MRS_ECS_DEFAULT_AGENCY**委托拥有对象存储服务的OBSOperateAccess权限和在集群所在区域拥有CESFullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。
指标共享	用于采集大数据组件的监控指标，当用户使用集群过程中出现问题时，供支持人员定位问题。MRS 3.x版本暂时没有该参数。
OBS权限控制	开启细粒度权限控制的用户可以通过该功能实现不同的MRS用户对OBS文件系统下的不同目录有不同的权限。具体请参见**配置MRS多用户访问OBS细粒度权限**。MRS 3.x版本暂时没有该参数。
数据盘加密	是否对集群挂载的数据盘中的数据进行加密，默认关闭。如需使用该功能，当前用户必须拥有“Security Administrator”和“KMS Administrator”权限。MRS 3.x版本暂时没有该参数。 加密数据盘使用的密钥由数据加密服务（DEW，Data Encryption Workshop）中的密钥管理（KMS，Key Management Service）功能提供，无需您自行构建和维护密钥管理基础设施，安全便捷。 通过单击“数据盘加密”开启或关闭数据盘加密功能。

参数	参数说明
密钥ID	当“数据盘加密”功能开启时，显示该参数。用于显示已选择的密钥名称对应的密钥ID。MRS 3.x版本暂时没有该参数。
密钥名称	当“数据盘加密”功能开启时，需要配置该参数。选择用来加密数据盘的密钥名称，默认选择密钥名称为“evs/default”的默认主密钥，在下拉框中可以选择其他用户主密钥。MRS 3.x版本暂时没有该参数。 使用用户主密钥加密云硬盘，若对用户主密钥执行禁用、计划删除等操作，将会导致云硬盘不可读写，甚至数据永远无法恢复，请谨慎操作。 单击“查看密钥列表”，进入密钥管理页面可以创建及管理密钥。
告警	开启告警功能可在集群运行异常或系统故障时，及时通知集群维护人员定位问题。
规则名称	用户自定义发送告警消息的规则名称，只能包含数字、英文字符、中划线和下划线。
主题名称	选择已创建的主题，也可以单击“创建主题”重新创建。新创建的主题请参考**向主题添加订阅**向该主题添加订阅者才能接收发布至主题的消息。 主题是发送消息和订阅通知的信道，为发布者和订阅者提供一个可以相互交流的通道。
Kerberos认证	登录Manager管理页面时是否启用Kerberos认证。 •  ：“Kerberos认证”关闭时，普通用户可使用MRS集群的所有功能。建议单用户场景下使用。 •  ：“Kerberos认证”开启时，普通用户无权限使用MRS集群的“文件管理”和“作业管理”功能，并且无法查看Hadoop、Spark的作业记录以及集群资源使用情况。如果需要使用集群更多功能，需要找Manager的管理员分配权限。建议在多用户场景下使用。
用户名	Manager管理员用户，目前默认为**admin**用户。

参数	参数说明
密码	配置Manager管理员用户的密码。 需要满足： ● 密码长度应在8～26个字符之间 ● 必须包含如下4种字符的组合 – 至少一个小写字母 – 至少一个大写字母 – 至少一个数字 – 至少一个特殊字符：！?,.: -_{} [ ]@ $% ^ + = / ● 不能和用户名或倒序的用户名相同 安全程度：颜色条红、橙、绿分别表示密码安全强度弱、中、强。
确认密码	再次输入Manager管理员用户的密码。
登录方式	● 密码 使用密码方式登录ECS节点。 密码设置约束如下： 1. 字符串类型，可输入的字符串长度为8~26。 2. 至少包含四种字符组合，如大写字母，小写字母，数字，特殊字符（！?,.: -_{} [ ]@ $% ^ + = /）。 3. 不能与用户名或倒序用户名相同。 ● 密钥对 使用密钥方式登录集群ECS节点。从下拉框中选择密钥对，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。 密钥对即SSH密钥，包含SSH公钥和私钥。您可以新建一个SSH密钥，并下载私钥用于远程登录身份认证。为保证安全，私钥只能下载一次，请妥善保管。 您可以通过以下两种方式中的任意一种使用SSH密钥。 1. 创建SSH密钥：创建SSH密钥，同时会创建公钥和私钥，公钥保存在ECS系统中，私钥保存在用户本机。当登录弹性云服务器时，使用公钥和私钥进行鉴权。 2. 导入SSH密钥：当用户已有公钥和私钥，可以选择将公钥导入系统。当登录弹性云服务器时，使用公钥和私钥进行鉴权。

参数	参数说明
通信安全授权	MRS集群通过管理控制台为用户发放、管理和使用大数据组件，大数据组件部署在用户的VPC内部，MRS管理控制台需要直接访问部署在用户VPC内的大数据组件时需要开通相应的安全组规则，而开通相应的安全组规则需要获取用户授权，此授权过程称为通信安全授权。具体请参考**授权安全通信**。 若不开启通信安全授权，MRS将无法创建集群。

集群创建失败

如果集群创建失败后，失败任务会自动转入“失败任务管理”页面。选择“集群列表 > 现有集群”，单击图4-1中进入“失败任务管理”页面，在“任务状态”列中，将鼠标移动到上可以查看到失败原因。可以参见**查看失败的集群操作任务**章节删除失败任务。

图 4-1 失败任务管理

MRS集群创建失败错误码列表如**表4-5**所示。

表 4-5 错误码

错误码	说明
MRS.101	用户请求配额不足，请联系客服提升配额。
MRS.102	用户Token为空或不合法，请稍后重试或联系客服。
MRS.103	用户请求不合法，请稍后重试或联系客服。
MRS.104	用户资源不足，请稍后重试或联系客服。
MRS.105	现子网IP不足，请稍后重试或联系客服。
MRS.201	因ECS服务导致失败，请稍后重试或联系客服。
MRS.202	因IAM服务导致失败，请稍后重试或联系客服。
MRS.203	因VPC服务导致失败，请稍后重试或联系客服。
MRS.400	MRS内部出错，请稍后重试或联系客服。

4.4 创建自定义拓扑集群

MRS当前提供的“分析集群”、“流式集群”和“混合集群”采用固定模板进行部署集群的进程，无法满足用户自定义部署管理角色和控制角色在集群节点中的需求。如

需自定义集群部署方式，可在创建集群时的“集群类型”选择“自定义”，实现用户自主定义集群的进程实例在集群节点中的部署方式。仅MRS 3.x及之后版本支持创建自定义拓扑集群。

自定义集群可实现以下功能：

- 管控分离部署，管理角色和控制角色分别部署在不同的Master节点中。
- 管控合设部署，管理角色和控制角色共同部署在Master节点中。
- ZooKeeper单独节点部署，增加可靠性。
- 组件分开部署，避免资源争抢。

MRS集群中角色类型：

- 管理角色：Management Node(MN)，安装Manager，即MRS集群的管理系统，提供统一的访问入口。Manager对部署在集群中的节点及服务进行集中管理。
- 控制角色：Control Node(CN)，控制监控数据角色执行存储数据、接收数据、发送进程状态及完成控制节点的公共功能。 MRS的控制节点包括HMaster、HiveServer、ResourceManager、NameNode、JournalNode、SlapdServer等。
- 数据角色：Data Node(DN)，执行管理角色发出的指示，上报任务状态、存储数据，以及执行数据节点的公共功能。MRS的数据节点包括DataNode、RegionServer、NodeManager等。

创建自定义集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“自定义创建”页签。

步骤4 参考下列参数说明配置集群软件信息，参数详细信息请参考**软件配置**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：目前仅MRS 3.x版本支持。
- 集群类型：选择“自定义”并根据需要勾选对应组件。

步骤5 单击“下一步”，并配置硬件信息。

- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 安全组：选择“自动创建”。
- 弹性公网IP：选择“暂不绑定”。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.x版本无该参数。
- 常用模板：具体说明请参见**自定义集群模板说明**。
- 实例规格：单击 配置实例规格、系统盘和数据盘存储类型和存储空间。

- 实例数量：请根据业务量调整集群实例数量。具体可参考**表4-7**。
- 拓扑调整：若常用模板中的部署方式不满足需求或者需要手动安装部分默认安装不部署的实例或者需要手动安装部分实例时，请设置“拓扑调整”为“开启”，然后根据业务需要调整实例部署方式，具体说明请参见**自定义集群拓扑调整说明**。

步骤6 单击“下一步”进入高级配置页签。

参数说明请参见**高级配置（可选）**。

步骤7 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤8 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

----结束

自定义集群模板说明

表 4-6 自定义集群常用模板说明

常用模板	说明	节点数量范围
管控合设	管理角色和控制角色共同部署在Master节点中，数据实例合设在同一节点组。该部署方式适用于100个以下的节点，可以减少成本。	• Master节点数量大于等于3个，小于等于11个。 • 节点组数量总和小于等于10个，非Master节点组中节点数量总和小于等于10000个。
管控分设	管理角色和控制角色分别部署在不同的Master节点中，数据实例合设在同一节点组。该部署方式适用于100-500个节点，在高并发负载情况下表现更好。	• Master节点数量大于等于5个，小于等于11个。 • 节点组数量总和小于等于10个，非Master节点组中节点数量总和小于等于10000个。
数据分设	管理角色和控制角色分别部署在不同的Master节点中，数据实例分设在不同节点组。该部署方式适用于500个以上的节点，可以将各组件进一步分开部署，适用于更大的集群规模。	• Master节点数量大于等于9个，小于等于11个。 • 节点组数量总和小于等于10个，非Master节点组中节点数量总和小于等于10000个。

表 4-7 MRS 自定义集群节点部署方案

节点部署原则		适用场景	组网规则
管理节点、控制节点和数据节点分开部署（此方案至少需要8个节点）	MN × 2 + CN × 9 + DN × n	（推荐）数据节点数500-2000时采用此方案	• 集群节点数超过200时，各节点划分到不同子网，各子网通过核心交换机三层互联，每个子网的节点数控制在200个以内，不同子网中节点数量请保持均衡。 • 集群节点数低于200时，各节点部署在同一子网，集群内通过汇聚交换机二层互联。
	MN × 2 + CN × 5 + DN × n	（推荐）数据节点数100-500时采用此方案	
	MN × 2 + CN × 3 + DN × n	（推荐）数据节点数30-100时采用此方案	
管理节点和控制节点合并部署，数据节点单独部署	(MN+CN) × 3 + DN × n	（推荐）数据节点数3-30时采用此方案	集群内节点部署在同一子网，集群内通过汇聚交换机二层互联。
管理节点、控制节点和数据节点合并部署		• 节点数小于6的集群使用此方案 • 此方案至少需要3个节点 **说明** 生产环境或商用环境不推荐使用此场景： • 管理节点、控制节点和数据节点合并部署时，集群性能和可靠性都会产生较大影响。 • 如节点数量满足需求，建议将数据节点单独部署。 • 如节点数量不满足将数据节点单独部署的要求，必须使用此场景时，需要使用双平面组网方式。将管理网络与业务网络流量隔离，防止业务平面的数据量过大，导致管理操作不能正常下发。	集群内节点部署在同一子网，集群内通过汇聚交换机二层互联。

自定义集群拓扑调整说明

表 4-8 拓扑调整说明

服务名称	依赖关系	角色名称	角色业务部署建议	说明
OMSServer	-	OMSServer	部署在Master节点上，不支持修改。	-
ClickHouse	依赖ZooKeeper	CHS（ClickHouseServer）	所有节点均可部署。 角色实例部署数量范围：偶数个，2~256。	部署了该角色的非Master节点组会被认为是Core节点类型。
		CLB（ClickHouseBalancer）	所有节点均可部署。 角色实例部署数量范围：2~256。	-
ZooKeeper	-	QP(quorumpeer)	只能部署在Master节点上。 角色实例部署数量范围：3~9，步长为2。	-
Hadoop	依赖ZooKeeper	NN(NameNode)	只能部署在Master节点上。 角色实例部署数量范围：2。	NameNode与Zkfc进程共机部署用于集群高可用
		HFS（HttpFS）	只能部署在Master节点上。 角色实例部署数量范围：0~10。	-
		JN(JournalNode)	只能部署在Master节点上。 角色实例部署数量范围：3~60，步长为2。	-
		DN(DataNode)	所有节点均可部署。 角色实例部署数量范围：3~10000。	部署了该角色的非Master节点组会被认为是Core节点类型。
		RM(ResourceManager)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		NM(NodeManager)	所有节点均可部署。 角色实例部署数量范围：3~10000。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
		JHS(JobHistoryServer)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
		TLS(TimelineServer)	只能部署在Master节点上。 角色实例部署数量范围：0~1。	-
Presto	依赖Hive	PCD(Coordinator)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		PWK(Worker)	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
Spark2x	• 依赖Hadoop • 依赖Hive • 依赖ZooKeeper	JS2X(JDBCServer2x)	只能部署在Master节点上。 角色实例部署数量范围：2~10。	-
		JH2X(JobHistory2x)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		SR2X(SparkResource2x)	只能部署在Master节点上。 角色实例部署数量范围：2~50。	-
		IS2X(IndexServer2x)	（可选）只能部署在Master节点上。 角色实例部署数量范围：0~2，步长为2。	-
HBase	依赖Hadoop	HM(HMaster)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		TS(ThriftServer)	所有节点均可部署。 角色实例部署数量范围：0~10000。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
		RT(RESTServer)	所有节点均可部署。 角色实例部署数量范围：0~10000。	-
		RS(RegionServer)	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
		TS1(Thrift1Server)	所有节点均可部署。 角色实例部署数量范围：0~10000。	若集群安装了Hue服务并且需要在Hue WebUI使用HBase，HBase服务需安装此实例。
Hive	• 依赖Hadoop • 依赖DBService	MS(MetaStore)	只能部署在Master节点上。 角色实例部署数量范围：2~10。	-
		WH（WebHCat）	只能部署在Master节点上。 角色实例部署数量范围：1~10。	-
		HS(HiveServer)	只能部署在Master节点上。 角色实例部署数量范围：2~80。	-
Hue	依赖DBService	H(Hue)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
Sqoop	依赖Hadoop	SC(SqoopClient)	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
Kafka	依赖ZooKeeper	B(Broker)	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
Flume	-	MS(MonitorServer)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
		F(Flume)	所有节点均可部署。 角色实例部署数量范围：1~10000。	部署了该角色的非Master节点组会被认为是Core节点类型。

服务名称	依赖关系	角色名称	角色业务部署建议	说明
Tez	• 依赖Hadoop • 依赖DBService • 依赖ZooKeeper	TUI(TezUI)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
Flink	• 依赖ZooKeeper • 依赖Hadoop	FR(FlinkResource)	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
		FS(FlinkServer)	所有节点均可部署。 角色实例部署数量范围：0~2。	-
Oozie	• 依赖Hadoop • 依赖DBService • 依赖ZooKeeper	O(oozie)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
Impala	• 依赖Hadoop • 依赖Hive • 依赖DBService • 依赖ZooKeeper	StateStore	只能部署在Master节点上。 角色实例部署数量范围：1。	-
		Catalog	只能部署在Master节点上。 角色实例部署数量范围：1。	-
		Impalad	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
Kudu	-	KuduMaster	只能部署在Master节点上。 角色实例部署数量范围：3或者5。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
		KuduTserver	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
Ranger	依赖DBservice	RA(RangerAdmin)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
		USC(UserSync)	只能部署在Master节点上。 角色实例部署数量范围：1。	-
		TSC（TagSync）	所有节点均可部署。 角色实例部署数量范围：0~1。	-

4.5 添加集群标签

标签是集群的标识。为集群添加标签，可以方便用户识别和管理拥有的集群资源。

您可以在创建集群时添加标签，也可以在集群创建完成后，在集群的详情页添加标签，您最多可以给集群添加10个标签。

标签共由两部分组成：“标签键”和“标签值”，其中，“标签键”和“标签值”的命名规则如表4-9所示。

表 4-9 标签命名规则

参数	规则	样例
标签键	不能为空。 对于同一个集群，Key值唯一。 长度不超过36个字符。 不能包含“=”，“*”，“<”，“>”，“\”，“,”，“|”，“/”，且首尾字符不能为空格。	Organization
标签值	长度不超过43个字符。 不能包含“=”，“*”，“<”，“>”，“\”，“,”，“|”，“/”，且首尾字符不能为空格。value可以为空。	Apache

为集群增加标签

在申请集群页，为集群增加标签。

1. 登录MRS管理控制台。
2. 单击“创建集群”，进入集群页面。
3. 在集群页面，选择“自定义创建”。
4. 参考**创建自定义集群**配置集群软件配置和硬件配置信息。
5. 在“高级配置”页签的标签栏。

   输入新添加标签的键和值。

   系统支持添加多个标签，最多可添加10个标签，并取各个标签的交集，对目标集群进行搜索。

   说明

   您也可对现有集群增加标签，详见**管理标签**。

搜索目标集群

在现有集群列表页，按标签键或标签值搜索目标集群。

1. 登录MRS管理控制台。
2. 单击现有集群列表右上角的“标签搜索”，展开查询页。
3. 输入待查询集群的标签。

   标签键或标签值可以通过下拉列表中选择，当标签键或标签值全匹配时，系统可以自动查询到目标集群。当有多个标签条件时，会取各个标签的交集，进行集群查询。
4. 单击“搜索”。

   系统根据标签键或标签值搜索目标集群。

管理标签

在现有集群的标签页，执行标签的增、删、改、查操作。

1. 登录MRS管理控制台。
2. 在现有集群列表中，单击待管理标签的集群名称。

   系统跳转至该集群详情页面。
3. 选择“标签管理”页签，对集群的标签执行增、删、改、查。
   - 查看

     在“标签”页，可以查看当前集群的标签详情，包括标签个数，以及每个标签的键和值。
   - 添加

     单击左上角的“添加标签”，在弹出的“添加标签”窗口，输入新添加标签的键和值，并单击“确定”。
   - 修改

     单击标签所在行“操作”列下的“编辑”，在弹出的“编辑标签”窗口，输入修改后标签的值，并单击“确定”。
   - 删除

单击标签所在行“操作”列下的“删除”，如果确认删除，在弹出的“删除标签”窗口，单击“确定”。

说明

MRS标签更新会同步到集群中的每台ECS上，为了使所有ECS标签与MRS标签保持一致，不建议在ECS服务控制台上单独修改MRS集群的ECS标签。当集群中某个ECS的标签数量达到上限时，集群将不能再创建标签。

4.6 授权安全通信

MRS集群通过管理控制台为用户发放、管理和使用大数据组件，大数据组件部署在用户的VPC内部，MRS管理控制台需要直接访问部署在用户VPC内的大数据组件时需要开通相应的安全组规则，而开通相应的安全组规则需要获取用户授权，此授权过程称为通信安全授权。

若不开启通信安全授权，MRS将无法创建集群。集群创建成功后若关闭通信将导致集群状态为“网络通道未授权”且如下功能将受到影响：

- 大数据组件安装、集群扩容、集群缩容、升级Master节点规格功能不可用。
- 集群的运行状态、告警、事件无法监控。
- 集群详情页的节点管理、组件管理、告警管理、文件管理、作业管理、补丁管理、租户管理功能不可用。
- Manager页面、各组件的Web站点无法访问。

再次开启通信安全授权，集群状态会恢复为“运行中”，以上功能将恢复为可用。具体操作请参见**为关闭安全通信的集群开启安全通信**。

当集群中授权的安全组规则不足以支撑MRS集群管理控制台为用户发放、管理和使用大数据组件的操作时，“通信安全授权”右侧出现的提示，请单击“一键修复”按钮进行修复，具体请参考**一键修复**。

创建集群时开启安全通信

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入集群页面。

步骤3 在集群页面，选择“快速创建”或“自定义创建”。

步骤4 参考**创建自定义集群**配置集群信息。

步骤5 在“高级配置”页签的“通信安全授权”栏，勾选“确认授权”。

步骤6 单击“立即创建”创建集群。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

----结束

集群创建成功后关闭安全通信

步骤1 登录MRS管理控制台。

步骤2 在现有集群列表中，单击待关闭安全通信的集群名称。

系统跳转至该集群详情页面。

步骤3 单击“通信安全授权”右侧的开关关闭授权，在弹出窗口单击“确定”。

关闭授权后将导致集群状态变更为“网络通道未授权”，集群部分功能不可用，请谨慎操作。

----结束

为关闭安全通信的集群开启安全通信

步骤1 登录MRS管理控制台。

步骤2 在现有集群列表中，单击待开启安全通信的集群名称。

系统跳转至该集群详情页面。

步骤3 单击“通信安全授权”右侧的开关开启授权。

开启授权后集群状态变更为“运行中”。

----结束

一键修复

当集群中授权的安全组规则不足以支撑MRS集群管理控制台为用户发放、管理和使用大数据组件的操作时，“通信安全授权”右侧出现 的提示，请单击“一键修复”按钮进行修复。

步骤1 登录MRS管理控制台。

步骤2 在现有集群列表中，单击待修复安全通信的集群名称。

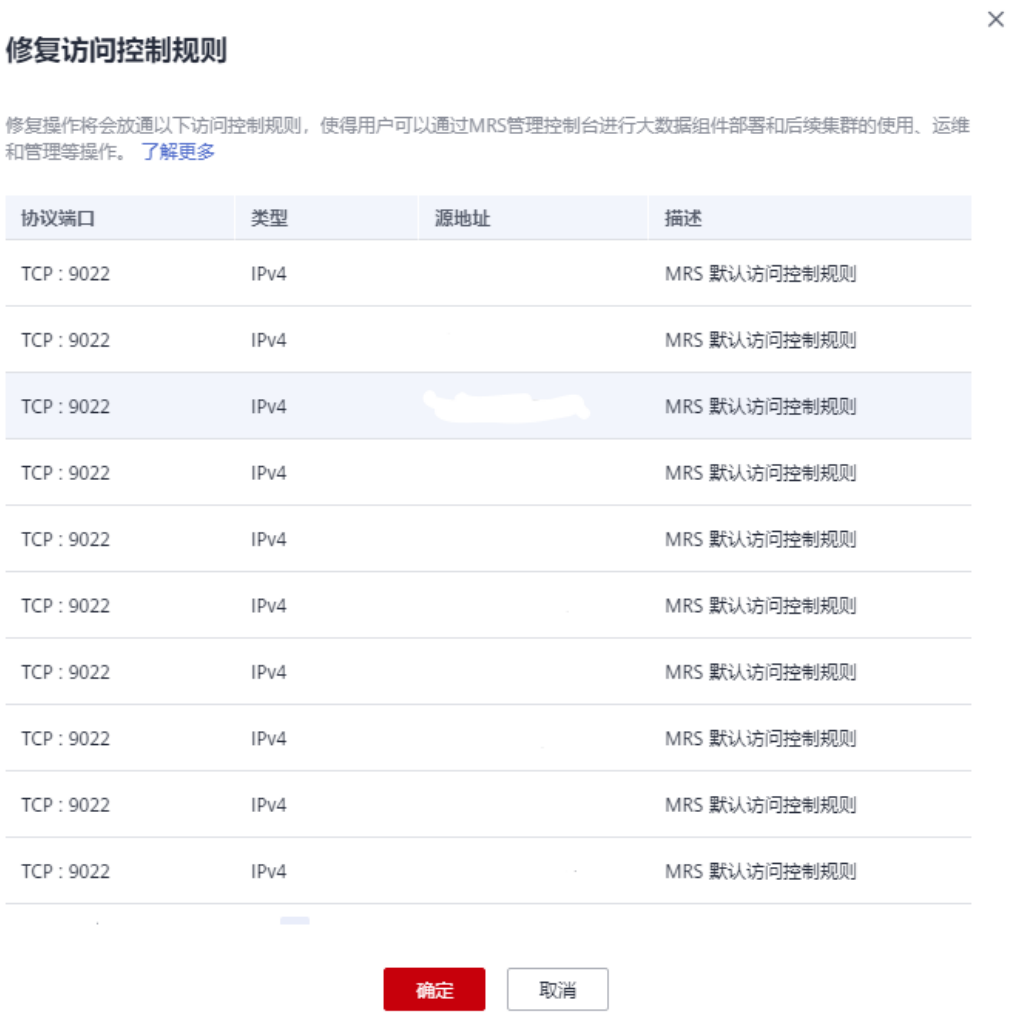
系统跳转至该集群详情页面。

步骤3 单击“通信安全授权”右侧的“一键修复”。

图 4-2 一键修复

步骤4 单击“确定”，完成修复。

图 4-3 修复访问控制策略



----结束

4.7 配置弹性伸缩规则

背景信息

在大数据应用，尤其是实时分析处理数据的场景中，常常需要根据数据量的变化动态调整集群节点数量以增减资源。MRS的弹性伸缩规则功能支持根据集群负载对集群的Task节点进行弹性伸缩。如果数据量是按照周期进行有规律的变化，用户可以按照固定时间段来自动调整Task节点数量范围，从而在数据量变化前提前完成集群的扩缩容。

- 弹性伸缩规则：根据集群实时负载指标对Task节点数量进行调整，数据量变化后触发扩缩容，有一定的延后性。
- 资源计划：按时间段设置Task节点数量范围，若数据量变化存在周期性规律，则可通过资源计划在数据量变化前提前完成集群的扩缩容，避免出现增加或减少资源的延后。

弹性伸缩规则与资源计划均可触发弹性伸缩，两者必须至少配置其中一种，也可以叠加使用。资源计划与基于负载的弹性伸缩规则叠加使用可以使得集群节点的弹性更好，足以应对偶尔超出预期的数据峰值出现。

当某些业务场景要求在集群扩缩容之后，根据节点数量的变化对资源分配或业务逻辑进行更改时，手动扩缩容的场景客户可以登录集群节点进行操作。对于弹性伸缩场景，MRS支持通过自定义弹性伸缩自动化脚本来解决。自动化脚本可以在弹性伸缩前后执行相应操作，自动适应业务负载的变化，免去了人工操作。同时，自动化脚本给用户实现个性需求提供了途径，完全自定义的脚本与多个可选的执行时机基本可以满足用户的各项需求，使弹性伸缩更具灵活性。

- 弹性伸缩规则：
  - 用户对于一个集群，可以同时设置扩容、缩容最多各5条弹性伸缩规则。
  - 系统根据用户的配置顺序从前到后依次判断规则，先扩容，后缩容。请尽量把重要的策略放在前面，以防一次扩容或缩容无法达到预期效果而进行反复触发。
  - 比对因子包括大于、大于等于、小于、小于等于。
  - 集群连续5n（n默认值为1）分钟持续满足配置的指标阈值后才能触发扩容或者缩容。
  - 每次扩容或者缩容后，存在一个冷却时间，冷却时间默认为20分钟，最小值为0。
  - 单次扩容或者缩容的节点数，最小1个节点，最大100个节点。
- 资源计划（按时间段设置Task节点数量范围）：
  - 用户可以按时间段设置集群Task节点的最大数量和最小数量，当集群Task节点数不满足当前时间资源计划节点范围要求时，系统触发扩容或缩容。
  - 用户最多可以为一个集群设置5条资源计划。
  - 资源计划周期以天为单位，起始时间与结束时间可以设置为00:00-23:59之间的任意时间点。起始时间早于结束时间至少30分钟。不同资源计划配置的时间段不可交叉。
  - 资源计划触发扩容或缩容后，存在10分钟的冷却时间，冷却时间内不会再次触发弹性伸缩。
  - 当启用资源计划时，在除配置资源计划配置时间段的其他时间内，集群Task节点数量会被限定在用户配置的默认节点数量范围内。
  - 当不启用资源计划时，集群不会将Task节点数量限制在默认节点数量范围内。
- 自动化脚本：
  - 用户可以设置自定义脚本，当弹性伸缩触发时，在集群节点上自动运行。
  - 用户最多可以为一个集群设置10个自动化脚本。
  - 可以指定自动化脚本某种或多种类型的节点上执行。
  - 脚本执行时机可以是扩容前、扩容后、缩容前或缩容后。
  - 使用自动化脚本前，请先将脚本上传到集群虚拟机或与集群同region的OBS文件系统中。集群虚拟机上的脚本只能在已有节点上执行，若脚本需要在新扩容的节点上执行，请将脚本上传到OBS。

进入弹性伸缩配置界面

弹性伸缩功能可以创建集群时，在高级配置参数中进行配置，也可以集群创建成功后通过管理控制台对集群内的Task节点组配置相关规则。

创建集群时配置弹性伸缩：

步骤1 登录MRS管理控制台。

步骤2 在包含有Task类型节点组件的集群时，参考**创建自定义集群**配置集群软件配置和硬件配置信息后，在“高级配置”页签的弹性伸缩栏，打开对应Task节点类型后的开关按钮，即可进行弹性伸缩规则及资源计划的配置或修改。

您可以参考以下场景进行配置：

- **场景1：单独配置弹性伸缩规则**
- **场景2：单独使用资源计划**
- **场景3：弹性伸缩规则与资源计划叠加使用**

----结束

为已有集群配置弹性伸缩：

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 在“节点管理”页签Task类型节点组的“操作”列单击“弹性伸缩”，即可进入“弹性伸缩”页面。

说明

- 弹性伸缩仅用于Task节点组，当集群内没有Task节点时，先单击“配置Task节点”增加Task节点后再继续配置。
- 对于MRS 3.x及之后版本，“配置Task节点”仅适用于分析集群、流试集群和混合集群。MRS 3.x及之后版本的自定义集群请参考**添加Task节点**添加task类型的节点。

步骤4 打开弹性伸缩开关按钮，即可进行弹性伸缩规则及资源计划的配置或修改。

您可以参考以下场景进行配置：

- **场景1：单独配置弹性伸缩规则**
- **场景2：单独使用资源计划**
- **场景3：弹性伸缩规则与资源计划叠加使用**

----结束

场景 1：单独配置弹性伸缩规则

例如业务场景如下：

需要根据Yarn资源使用情况动态调整节点数，在Yarn可用内存低于20%时扩容5个节点，可用内存高于70%时缩容5个节点。Task节点组最高不超过10个节点，最低不少于1个节点。

步骤1 进入弹性伸缩配置界面后，配置弹性伸缩规则。

- 配置节点默认范围

  输入Task实例弹性伸缩的数量范围，此约束应用于所有扩容缩容规则，取值范围为0～500。

  例如本业务场景中，配置为1~10。

- 配置弹性伸缩规则

  需要配置扩容或者缩容规则，否则弹性伸缩将不会启用。

  a. 规则类型选择“扩容”或者“缩容”。

  b. 单击“添加规则”，进入规则编辑页面。

  c. 配置“规则名称”、“如果”、“持续”、“添加”、“冷却时间”。

  d. 单击“确定”。

     您可以在弹性伸缩页面，扩容或者缩容区域查看、编辑或删除配置的规则。您可以继续添加并配置多条规则。

步骤2 （可选）配置自动化脚本。

在“高级配置”项选择“现在配置 > 添加”或者单击“添加自动化脚本”按钮，进入“自动化脚本”配置页面。

MRS 3.x版本集群不支持该操作。

1. 配置“名称”、“脚本路径”、“执行节点类型”、“参数”、“执行时机”、“失败操作”。参数详情请参考**表4-12**。
2. 单击“确定”保存自动化脚本配置。

步骤3 单击“确定”，完成弹性伸缩规则设置。

说明

如果是为已有集群配置弹性伸缩的场景，需勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

----结束

场景 2：单独使用资源计划

当数据量以天为周期有规律的变化，并且希望在数据量变化前提前完成集群的扩缩容，可以使用MRS的资源计划配置在规定时间内按计划调整Task节点数量。

例如：

某项实时处理业务数据量在每天7:00~13:00出现高峰，其他时间保持平稳低水平。假设使用MRS流式集群来处理该业务数据，在7:00~13:00时，为应对数据量高峰需要5个Task节点的资源，其他时间只需要2个Task节点。

步骤1 进入弹性伸缩配置界面后，配置资源计划。

1. 节点数量范围的“默认范围”设置为“2-2”，表示除资源计划规定时间范围外，其他时间Task节点数量固定为2个。
2. 单击默认范围下方的“配置指定时间段的节点数量范围”或者“添加资源计划”。
3. 配置“时间范围”和“节点数量范围”。

   例如此处“时间范围”设置为“07:00-13:00”，“节点数量范围”设置为“5-5”，表示在该时间范围内，Task节点数量固定为5个。

   参数详情请参考**表4-11**，可以单击“配置指定时间段的节点数量范围”配置多条资源计划。

说明

- 如果没有配置指定时间段的节点数量范围，则节点数量范围以“默认范围”为准。
- 如果配置了指定时间段的节点数量范围，则在这个时间范围内，以配置的“节点数量范围”为准。不在配置的时间范围时，则以“默认范围”为准。

步骤2 （可选）配置自动化脚本。

在“高级配置”项选择“现在配置 > 添加”或者单击“添加自动化脚本”按钮，进入“自动化脚本”配置页面。

MRS 3.x版本集群不支持该操作。

1. 配置“名称”、“脚本路径”、“执行节点类型”、“参数”、“执行时机”、“失败操作”。参数详情请参考**表4-12**。
2. 单击“确定”保存自动化脚本配置。

步骤3 单击“确定”，完成弹性伸缩规则设置。

说明

如果是为已有集群配置弹性伸缩的场景，需勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

----结束

场景 3：弹性伸缩规则与资源计划叠加使用

假如数据量并非非常平稳，有可能出现超出预期的波动，因此并不能保证固定Task节点范围一定可以满足业务场景，此时需要在资源计划的基础上根据实时负载对Task节点数量进行调整。

例如业务场景如下：

某项实时处理业务数据量在每天7:00-13:00出现规律性变化，但是数据量变化并非非常平稳。假设在7:00-13:00期间，需要Task节点的数量范围是5~8个，其他时间需要Task节点数量外围为2~4个。因此可以在资源计划的基础上，设置基于负载的弹性伸缩规则，以实现当数据量超出预期后，Task节点数量可以在资源计划规定的范围内根据负载情况进行浮动，但不会超出该规定范围。资源计划触发时，会以变化最小的方式使节点数量满足计划规定范围，即如果需要扩容则扩容到计划节点数量范围的下限，如果需要缩容则缩容到计划节点数量范围的上限。

步骤1 进入弹性伸缩配置界面后，配置弹性伸缩规则。

- 节点数量范围的默认范围：

  输入Task实例弹性伸缩的数量范围，此约束应用于所有扩容缩容规则。

  例如本场景中，配置为2~4个。

- 伸缩规则：

  需要配置扩容或者缩容，否则弹性伸缩将不会启用。

  a. 规则类型选择“扩容”或者“缩容”。
  b. 单击“添加规则”，进入“添加规则”页面。
  c. 配置“规则名称”、“如果”、“持续”、“添加”、“冷却时间”。
  d. 单击“确定”。

     您可以在弹性伸缩页面，扩容或者缩容区域查看配置的规则。

步骤2 配置资源计划。

1. 单击节点默认范围下方的“配置指定时间段的节点数量范围”或者“添加资源计划”。
2. 配置“时间范围”和“节点数量范围”。

   例如此处“时间范围”设置为“07:00-13:00”，“节点数量范围”设置为“5~8”。

   参数详情请参考**表4-11**，可以单击“配置指定时间段的节点数量范围”或者“添加资源计划”按钮配置多条资源计划。

   说明

   - 如果没有配置指定时间段的节点数量范围，则节点数量范围以“默认范围”为准。
   - 如果配置了指定时间段的节点数量范围，则在这个时间范围内，以配置的“节点数量范围”为准。不在配置的时间范围时，则以“默认范围”为准。

步骤3 （可选）配置自动化脚本。

在“高级配置”项选择“现在配置 > 添加”或者单击“添加自动化脚本”按钮，进入“自动化脚本”配置页面。

MRS 3.x版本集群不支持该操作。

1. 配置“名称”、“脚本路径”、“执行节点类型”、“参数”、“执行时机”、“失败操作”。参数详情请参考**表4-12**。
2. 单击“确定”保存自动化脚本配置。

步骤4 单击“确定”，完成弹性伸缩规则设置。

说明

如果是为已有集群配置弹性伸缩的场景，需勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

----结束

相关信息

在添加规则时，可以参考**表4-10**配置相应的指标。

表 4-10 弹性伸缩指标列表

集群类型	指标名称	数值类型	说明
流式集群	StormSlotAvailable	整型	Storm组件的可用slot数。 取值范围为[0～2147483646]。
	StormSlotAvailablePercentage	百分比	Storm组件可用slot百分比。是可用slot数与总slot数的比值。 取值范围为[0～100]。
	StormSlotUsed	整型	Storm组件的已用slot数。 取值范围为[0～2147483646]。

集群类型	指标名称	数值类型	说明
	StormSlotUsedPercentage	百分比	Storm组件已用slot百分比。是已用slot数与总slot数的比值。 取值范围为[0～100]。
	StormSupervisorMemAverageUsage	整形	Storm组件Supervisor的内存平均使用量。 取值范围为[0～2147483646]。
	StormSupervisorMemAverageUsagePercentage	百分比	Storm组件Supervisor进程使用的内存占系统总内存的平均百分比。 取值范围[0 ~ 100]。
	StormSupervisorCPUAverageUsagePercentage	百分比	Storm组件Supervisor进程使用的CPU占系统总CPU的平均百分比。 取值范围[0 ~ 6000]。
分析集群	YARNAppPending	整型	YARN组件挂起的任务数。 取值范围为[0～2147483646]。
	YARNAppPendingRatio	比率	YARN组件挂起的任务数比例。是YARN挂起的任务数与YARN运行中的任务数比值。 取值范围为[0～2147483646]。
	YARNAppRunning	整型	YARN组件运行中的任务数。 取值范围为[0～2147483646]。
	YARNContainerAllocated	整型	YARN组件中已分配的container个数。 取值范围为[0～2147483646]。
	YARNContainerPending	整型	YARN组件挂起的container个数。 取值范围为[0～2147483646]。
	YARNContainerPendingRatio	比率	YARN组件挂起的container比率。是挂起的container数与运行中的container数的比值。 取值范围为[0～2147483646]。
	YARNCPUAllocated	整型	YARN组件已分配的虚拟CPU核心数。 取值范围为[0～2147483646]。
	YARNCPUAvailable	整型	YARN组件可用的虚拟CPU核心数。 取值范围为[0～2147483646]。
	YARNCPUAvailablePercentage	百分比	YARN组件可用虚拟CPU核心数百分比。是可用虚拟CPU核心数与总虚拟CPU核心数比值。 取值范围为[0～100]。

集群类型	指标名称	数值类型	说明
	YARNCPUPending	整型	YARN组件挂起的虚拟CPU核心数。 取值范围为[0～2147483646]。
	YARNMemoryAllocated	整型	YARN组件已分配内存大小。单位为MB。 取值范围为[0～2147483646]。
	YARNMemoryAvailable	整型	YARN组件可用内存大小。单位为MB。 取值范围为[0～2147483646]。
	YARNMemoryAvailablePercentage	百分比	YARN组件可用内存百分比。是YARN组件可用内存大小与YARN组件总内存大小的比值。 取值范围为[0～100]。
	YARNMemoryPending	整型	YARN组件挂起的内存大小。 取值范围为[0～2147483646]。

说明

- 表4-10中指标数值类型为百分比或比率时，有效数值可精确到百分位。其中百分比类型指标数值为去除百分号（%）后的小数值，如16.80即代表16.80%。
- 混合集群的支持分析集群和流式集群的所有指标。

在添加资源计划时，可以参考表4-11配置相应的参数。

表 4-11 资源计划配置项说明

配置项	说明
时间范围	资源计划的起始时间和结束时间，精确到分钟，取值范围[00:00,23:59]。例如资源计划开始于早上8:00，结束于10:00，则配置为8:00-10:00。结束时间必须晚于开始时间至少30分钟。
节点数量范围	资源计划内的节点数量上下限，取值范围[0,500]，在资源计划时间内，集群Task节点数量小于最小节点数时，弹性伸缩会将集群Task节点一次性扩容到最小节点数。在资源计划时间内，集群Task节点数量大于最大节点数时，弹性伸缩会将集群Task节点一次性缩容到最大节点数。最小节点数必须小于或等于最大节点数。

说明

- 当启用资源计划时，弹性伸缩配置中的“默认节点数量范围”将在资源计划外的时间段内强制生效。例如“默认节点数量范围”配置为1-2，配置资源计划：08:00-10:00之间节点数量范围为4-5，则在一天中的非资源计划时间段（0:00-8:00以及10:00-23:59）内，Task节点会被强制限制在1个到2个中间，若节点数量大于2则触发自动缩容，若节点数量小于1则触发自动扩容。
- 当不启用资源计划时，节点数量范围的“默认范围”会在全部时间范围生效，如果节点数量不在“节点数量范围”的默认范围，主动增减Task节点数量到默认范围内。
- 资源计划间时间段不可交叉，时间段交叉意为某个时间点存在两个生效的资源计划，例如配置资源计划1在08:00-10:00生效，资源计划2在09:00-11:00生效，则两个资源计划存在时间段交叉，交叉时间段09:00-10:00。
- 资源计划不能跨天配置，例如如果要配置23:00至次日01:00的资源计划，请配置时间段为23:00-00:00和00:00-01:00的两个资源计划。

在添加自动化脚本时，可以参考表4-12配置相应参数。

表 4-12 自动化脚本配置说明

配置项	说明
名称	自动化脚本的名称。 只能由数字、英文字符、空格、中划线和下划线组成，且不能以空格开头。 可输入的字符串长度为1～64个字符。 **说明** 同一集群内，不允许配置相同的名称。不同集群之间，可以配置相同的名称。
脚本路径	脚本的路径。路径可以是OBS文件系统的路径或虚拟机本地的路径。 ● OBS文件系统的路径，必须以s3a://开头，以.sh结尾。例如：s3a://mrs-samples/xxx.sh ● 虚拟机本地的路径，脚本所在的路径必须以‘/’开头，以.sh结尾。例如，安装Zepelin的示例脚本路径如下：/opt/bootstrap/zepelin/zepelin_install.sh
执行节点类型	选择自动化脚本所执行的节点类型。 **说明** ● 如果选择Master节点，您可以通过开关选择是否只在Active Master节点执行此脚本。 ● 如果选择开启此功能，表示只在Active Master节点上执行。如果选择关闭，表示在所有Master节点执行。默认关闭。

配置项	说明
参数	自动化脚本参数，支持通过传入以下预定义变量获得弹性伸缩相关信息： ● ${mrs_scale_node_num}：弹性伸缩节点数量，总是正数 ● ${mrs_scale_type}：弹性伸缩类型，扩容为“scale_out”，缩容为“scale_in” ● ${mrs_scale_node_hostnames}：弹性伸缩节点的主机名，多个主机名之间以“,”隔开 ● ${mrs_scale_node_ips}：弹性伸缩节点的IP，多个IP之间以“,”隔开 ● ${mrs_scale_rule_name}：触发弹性伸缩的规则名，如果是资源计划则为“resource_plan”
执行时机	选择自动化脚本执行的时间。支持“扩容前”、“扩容后”、“缩容前”、“缩容后”四种类型。 **说明** 假设执行节点类型中包含Task节点： ● 执行时机为扩容前的脚本不会在将要扩容出的Task节点上执行。 ● 执行时机为扩容后的脚本会在扩容出的Task节点上执行。 ● 执行时机为缩容前的脚本会在即将被删除的Task节点上执行。 ● 执行时机为缩容后的脚本不会在已经被删除的Task节点上执行。
失败操作	该脚本执行失败后，是否继续执行后续脚本和扩缩容操作。 **说明** ● 建议您在调试阶段设置为“继续”，无论此脚本是否执行成功，则集群都能继续扩缩容操作。 ● 若脚本执行失败，请到集群虚拟机机的“/var/log/Bootstrap”路径下查看失败日志。 ● 由于缩容成功不可回滚，缩容后执行的脚本失败操作只能选择“继续”。

说明

自动化脚本只在弹性伸缩时触发，手动调整集群节点时不会运行。

4.8 管理数据连接

4.8.1 配置数据连接

MRS的数据连接是用来管理集群中组件使用的外部源连接，如Hive的元数据使用外部的关系型数据库，可以通过数据连接来关联Hive组件实现。

- 本地元数据：元数据存储于集群内的本地GaussDB中，当集群删除时元数据同时被删除，如需保存元数据，需提前前往数据库手动保存元数据。

- 数据连接：可选择关联与当前集群同一虚拟私有云和子网的RDS服务中的PostgresDB或MySQL数据库，元数据将存储于关联的数据库中，不会随当前集群的删除而删除，多个MRS集群可共享同一份元数据。

说明

不同集群间Hive元数据切换时，MRS当前只对Hive组件自身的元数据数据库中的权限进行同步。这是由于当前MRS上的权限模型是在Manager上维护的，所以不同集群间的Hive元数据切换，不能自动把用户/用户组的权限同步到另一个集群的Manager上。

数据连接前置操作

步骤1 登录RDS管理控制台。

步骤2 选择“实例管理”，单击MRS数据连接使用的RDS实例名称。

步骤3 单击右上角的“登录”，以root用户登录该实例。

步骤4 在实例“首页”即可单击“新建数据库”创建新的数据库。

步骤5 在页面顶部选择“帐号管理 > 用户管理”。

说明

当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**步骤5**-**步骤7**操作。

步骤6 单击“新建用户”，创建一个非root用户。

步骤7 在页面顶部选择“SQL操作 > SQL查询”，在“库名”处切换对应数据库，然后执行如下SQL命令为该数据库用户进行赋权，其中${db_name}与${db_user}为MRS待连接的数据库名和新建的用户名。

```
grant SELECT, INSERT on mysql.* to '${db_user}'@'%' with grant option;
grant all privileges on ${db_name}.* to '${db_user}'@'%' with grant option;
grant reload on *.* to '${db_user}'@'%' with grant option;
flush privileges;
```

步骤8 参考**创建数据连接**创建数据连接。

----结束

创建数据连接

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 单击“新建数据连接 ”。

步骤3 参考**表4-13**配置相关参数。

表 4-13 数据连接

参数	说明
类型	选择外部源连接的类型。 • RDS服务PostgreSQL数据库，MRS 支持Hive组件的集群支持连接该类型数据库。 • RDS服务MySQL数据库，支持Hive或Ranger组件的集群支持连接该类型数据库。
名称	数据连接的名称。
数据库实例	RDS服务数据库实例，该实例需要先在RDS服务创建后在此处引用，且已创建数据库，具体请参考**数据连接前置操作**。单击“查看RDS实例”查看已创建的实例。 **说明** • 为了保证集群和PostgreSQL数据库的网络访问，建议该实例与MRS集群的虚拟私有云和子网一致。 • 该实例的安全组入方向规则需要放通3306端口（可通过在RDS控制台单击实例名称进入实例基本信息页面，在“连接信息”区域单击“内网安全组”名称进入安全组控制台，在入方向规则页签中添加一个“协议端口”为TCP 3306，“源地址”为Hive的MetaStore实例所在的所有节点IP的规则）。 • 当前MRS支持的RDS上Postgres数据库版本号为**PostgreSQL9.5/PostgreSQL9.6**。 • 当前MRS仅支持RDS上MySQL数据库版本为**MySQL 5.7.x**。
数据库	待连接的数据库的名称。
用户名	登录待连接的数据库的用户名。
密码	登录待连接的数据库的密码。

说明

当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**操作。

步骤4 单击“确定”完成创建。

----结束

编辑数据连接

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 在数据连接列表的“操作列”，单击待编辑数据连接所在行的“编辑”。

步骤3 参考**表4-13**修改参数。

如果选择的数据连接已经关联了集群，编辑后会将修改后的配置同步到对应的集群中。

----结束

删除数据连接

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 在数据连接列表的操作列，单击待删除数据连接所在行的“删除 ”。

如果选择的数据连接已经关联了集群，删除动作不会影响对应的集群。

----结束

创建集群时配置数据连接

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“自定义创建”。

步骤4 在软件配置中，参考**表4-14**配置“元数据 ”，其他参数请参考**创建自定义集群**进行配置并创建集群。

表 4-14 数据连接参数说明

参数	参数说明
元数据	是否使用外部数据源存储元数据。 ● 本地元数据： 元数据存储在集群本地。 ● 数据连接：使用外部数据源元数据，若集群异常或删除时将不影响元数据，适用于存储计算分离的场景。 支持Hive或Ranger组件的集群支持该功能。
组件名	当“使用外部数据源存储元数据”功能开启时，该参数有效。用于表示可以设置外部数据源的组件类型。 ● Hive ● Ranger
数据连接类型	当“使用外部数据源存储元数据”功能开启时，该参数有效。用于表示外部数据源的类型。 ● Hive组件支持的数据连接类型： – RDS服务PostgreSQL数据库（1.9.x版本支持） – RDS服务MySQL数据库 – 本地数据库 ● Ranger组件支持的数据连接类型： – RDS服务MySQL数据库 – 本地数据库
数据连接实例	当“数据连接类型”选择“RDS服务PostgreSQL数据库”或“RDS服务MySQL数据库”时，该参数有效。用于表示MRS集群与RDS服务数据库连接的名称，该实例必须先创建才能在此处引用。可单击“创建数据连接”进行创建，具体请参考**数据连接前置操作**和**创建数据连接**进行操作。

----结束

4.8.2 配置 Ranger 数据连接

本指导旨在指导用户将现有集群的Ranger元数据切换为RDS数据库中存储的元数据。该操作可以使多个MRS集群共用同一份元数据，且元数据不随集群的删除而删除。也能够避免集群迁移时Ranger元数据的迁移。

前置条件

已创建RDS服务MySQL数据库的实例，请参考**创建数据连接**。

说明

- 对于MRS 3.x之前版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。
- 对于MRS 3.x及之后版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，数据库用户不允许为root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。

Ranger 元数据外置到 Mysql 前置操作

该前置操作仅在**MRS 3.1.0及之后版本**需要执行。

步骤1 登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。选择“集群 > 服务 > *服务名称*”。

当前MRS 3.1.x集群支持Ranger鉴权的组件为: HDFS、HBase、Hive、Spark、Impala、Storm、Kafka组件。

步骤2 在服务“概览”页面右上角单击“更多 > 停用Ranger鉴权”，如果“停用Ranger鉴权”是灰色，则表示未开启Ranger鉴权无需停用Ranger鉴权，如**图4-4**所示。

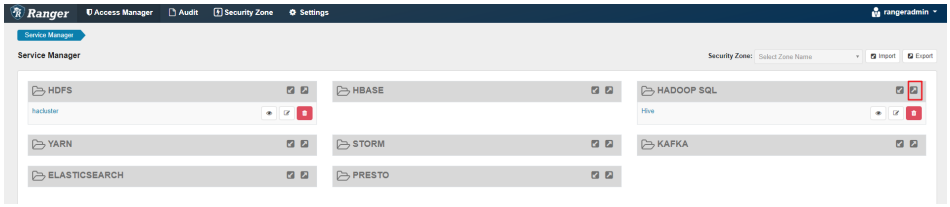
图 4-4 停用 Ranger 鉴权

步骤3 （可选）如需使用已有鉴权策略请执行该步骤在Ranger Web页面导出已有组件的鉴权策略，切换Ranger元数据完成后可重新导入已有的鉴权策略。此处以Hive为例，导出后会生成本地的json格式的策略文件。

1. 登录FusionInsight Manager页面。
2. 选择“集群 > 服务 > Ranger”，进入Ranger服务概览页面。
3. 单击“基本信息”区域中的“RangerAdmin”，进入Ranger WebUI界面。

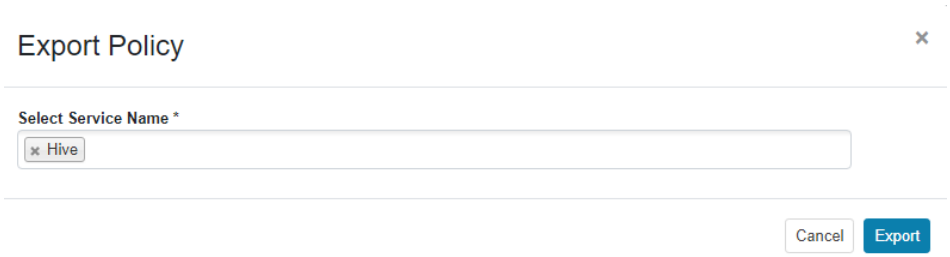
   admin用户在Ranger中的用户类型为“User”，如需查看所有管理页面，可单击右上角用户名后，选择“Log Out”，退出当前用户。
4. 使用**rangeradmin**用户（默认密码为Rangeradmin@123）或者其他具有Ranger管理员权限用户重新登录。
5. 单击Hive组件对应行的导出按钮，导出鉴权策略。

   图 4-5 导出鉴权策略

6. 单击“Export”，导出后会生成本地的json格式的策略文件。

   图 4-6 导出 Hive 鉴权策略

----结束

为 MRS 集群配置数据连接

步骤1 登录MRS控制台。

步骤2 单击集群名称进入集群详情页面。

步骤3 单击“数据连接”右侧的“单击管理”，进入数据连接配置界面。

步骤4 单击“配置数据连接”，并配置相关参数。

- 组件名称：Ranger
- 模块类型：Ranger元数据
- 连接类型：RDS服务MySQL数据库
- 连接实例：请选择已创建的到RDS服务MySQL数据库的实例，如需创建新的数据连接，请参考**创建数据连接**。

步骤5 勾选“我已经阅读上述信息，并了解具体影响。”并单击“测试”。

步骤6 测试成功后，单击“确定”完成数据连接配置。

步骤7 登录FusionInsight Manager页面。

步骤8 选择“集群 > 服务 > Ranger”，进入Ranger服务概览页面。

步骤9 单击“更多 > 重启服务”或“更多 > 滚动重启服务”。

重启服务会造成业务中断，滚动重启可以尽量减少或者不影响业务运行。

重启Ranger组件会影响所有受Ranger控制组件的权限，可能影响业务的正常运行，请在集群空闲或业务量较少时时执行重启。重启Ranger组件前，Ranger中的策略依然生效。

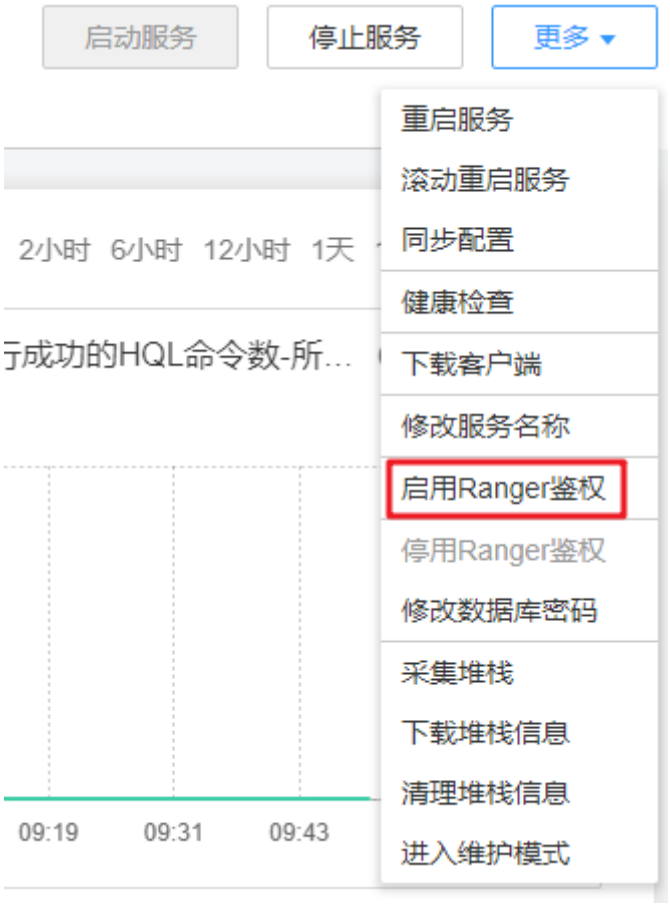
图 4-7 重启服务

步骤10 启用需要鉴权的组件的Ranger鉴权。此处以Hive组件为例。

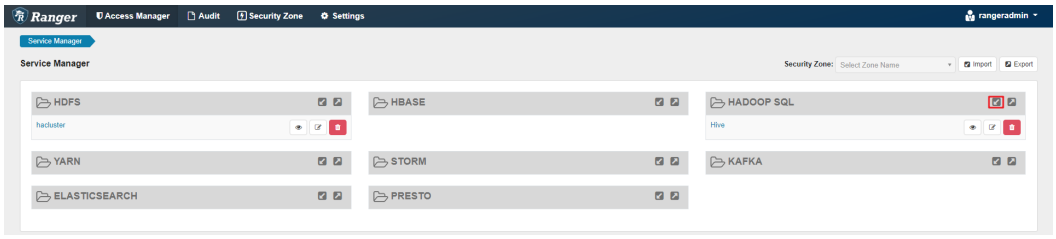
当前MRS 3.1.x集群支持Ranger鉴权的组件为: HDFS、HBase、Hive、Spark、Impala、Storm、Kafka组件。

1. 登录FusionInsight Manager页面，选择“集群 > 服务 > *服务名称*”。
2. 在服务“概览”页面右上角单击“更多 > 启用Ranger鉴权”。

图 4-8 启用 Ranger 鉴权



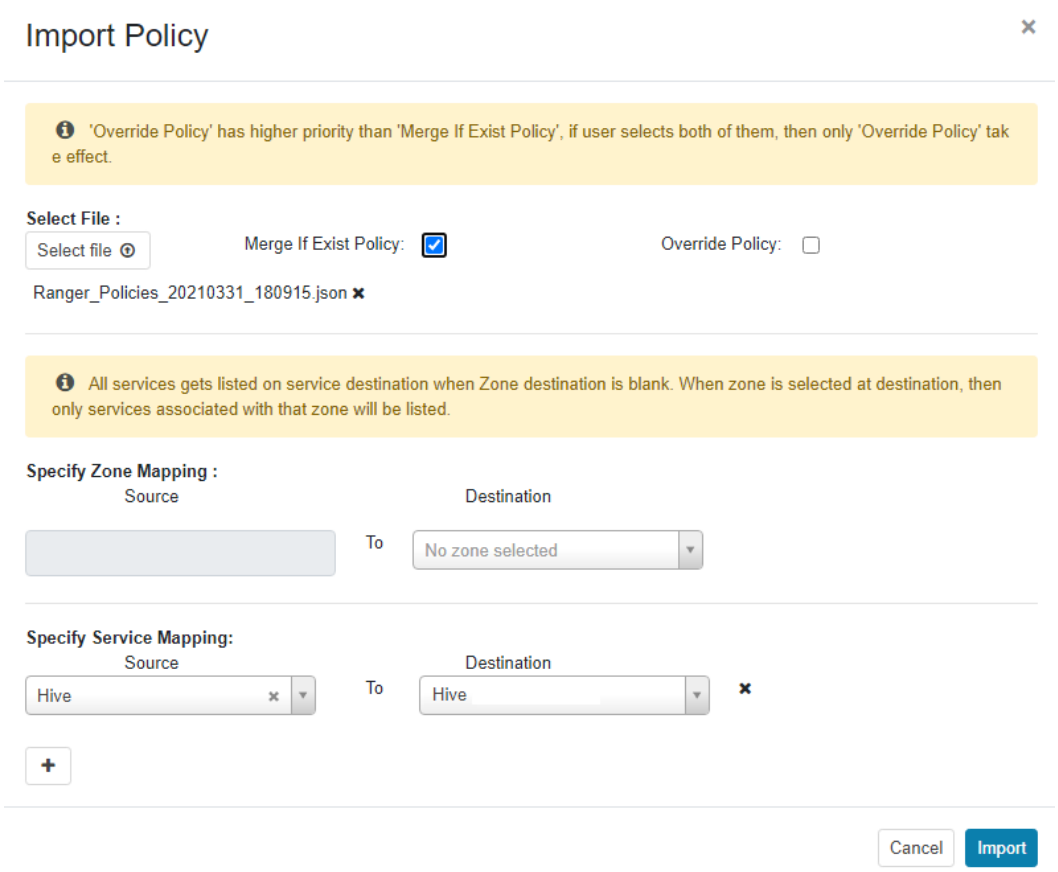
步骤11 登录Ranger WebUI界面，单击Hive组件对应行的导入按钮。



步骤12 配置导入参数。

- Select file：选择**步骤3.6**中下载的鉴权策略文件。
- Merge If Exist Policy：勾选。

图 4-9 导入鉴权策略



步骤13 重启启用Ranger鉴权的组件。

1. 登录FusionInsight Manager页面。
2. 选择“集群 > 服务 > Hive”，进入Hive服务概览页面。
3. 单击“更多 > 重启服务”或“更多 > 滚动重启服务”。
   重启服务会造成业务中断，滚动重启可以尽量减少或者不影响业务运行。

----结束

4.8.3 配置 Hive 数据连接

本章节指导用户在创建后，将现有集群的Hive元数据切换为本地数据库或者RDS数据库中存储的元数据。该操作可以使多个MRS集群共用同一份元数据，且元数据不随集群的删除而删除。也能够避免集群迁移时Hive元数据的迁移。

说明

- 不同集群间Hive元数据切换时，MRS当前只对Hive组件自身的元数据数据库中的权限进行同步。这是由于当前MRS上的权限模型是在Manager上维护的，所以不同集群间的Hive元数据切换，不能自动把用户/用户组的权限同步到另一个集群的Manager上。
- 对于MRS 3.x之前版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。
- 对于MRS 3.x及之后版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，数据库用户不允许为root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。

配置 Hive 数据连接

该功能在MRS 3.0.5版本暂不支持。

步骤1 登录MRS控制台，在导航栏选择“集群列表 > 现有集群”。

步骤2 单击集群名称，进入集群详情页面。

步骤3 在集群详情页的“概览”页签，单击“数据连接”右侧的“单击管理”。

步骤4 在“数据连接”页面显示集群已关联的数据连接，单击“编辑”或“删除”可对数据连接进行编辑或删除。

步骤5 若“数据连接”页面没有关联连接，单击“配置数据连接”进行增加。

说明

一种模块类型只能配置一个数据连接，如在Hive元数据上配置了数据连接后，不能再配置其他的数据连接。当没有可用的模块类型时，“配置数据连接”按钮不可用。

表 4-15 配置 Hive 数据连接

参数	说明
组件名称	Hive
模块类型	Hive元数据
连接类型	• RDS服务PostgreSQL数据库（1.9.x版本支持） • RDS服务MySQL数据库 • 本地数据库
连接实例	当“连接类型”参数选择“RDS服务PostgreSQL数据库”或“RDS服务MySQL数据库”时有效。选择MRS集群与RDS服务数据库连接名称，该连接必须先创建才能在此处引用。可单击“创建数据连接”进行创建，具体请参考**创建数据连接**。

步骤6 单击“测试”，测试此数据连接和集群的连通性。

步骤7 连接成功后单击“确定” 完成配置数据连接。

说明

- 配置了Hive元数据后，请重启Hive服务，Hive会在指定的数据库下创建Hive必须的数据库表（如表已经存在则不会创建）。
- 重启Hive服务前，请确保已安装对应驱动包到所有MetaStore实例所在节点中。
  - Postgres：使用开源驱动包替换集群已有的驱动包。将postgres驱动包postgresql-42.2.5.jar上传至所有MetaStore实例节点${BIGDATA_HOME}/third_lib/Hive目录下。
  - MySQL：进入MySQL官网（https://www.mysql.com/），选择“Downloads > Community > MySQL Connectors > Connector/J”下载对应版本的驱动包，将MySQL对应版本的驱动包上传至所有Metastore实例节点“/opt/Bigdata/FusionInsight_HD_*/install/FusionInsight-Hive-*/hive-*/lib/”目录下。

----结束

4.9 通过引导操作安装第三方软件

4.9.1 引导操作简介

引导操作是指启动集群组件前（或后）在指定的节点上执行脚本。您可以通过引导操作来完成安装其他第三方软件，修改集群运行环境等自定义操作。

如果集群扩容，选择执行引导操作，则引导操作也会以相同方式在新增节点上执行。如果集群开启弹性伸缩功能，可以在配置资源计划的同时添加自动化脚本，则自动化脚本会在弹性伸缩的节点上执行，实现用户自定义操作。

MRS会使用root用户执行您指定的脚本，脚本内部您可以通过**su - XXX**命令切换用户。

说明

引导操作脚本以root身份执行，使用不当可能会对集群可用性造成影响，请谨慎操作。

MRS通过引导操作脚本返回码来判断结果，如果返回零，则代表脚本执行成功，非零代表执行失败。一个节点上执行某个引导脚本失败，则会导致相应引导脚本失败，您可以通过“失败后操作”来选择是否继续执行后续脚本。举例1：创建集群指定所有脚本的“失败后操作”都选择“继续”，则不管这些脚本实际执行成功或失败，都会全部执行，并完成启动流程。举例2：如果一个脚本执行失败，且“失败后操作”选择“终止”，则不会执行后续脚本，集群创建或扩容也随之失败。

您最多可以添加18个引导操作，它们会按照您指定的顺序在集群组件启动前（或后）执行。组件启动前（或后）执行的引导操作，必须在60分钟内完成，否则会引起集群创建或扩容失败。

4.9.2 准备引导操作脚本

引导操作目前仅支持linux shell脚本，脚本文件需以.sh结尾。

上传所需安装包等文件至 OBS 文件系统

正式编写脚本前，您需要将所需安装包、配置包的所有相关文件都上传到同region的OBS文件系统中。因为不同region间有网络隔离，MRS虚拟机无法下载其他region上的OBS文件。

脚本中如何从 OBS 文件系统下载文件

您可以在脚本中指定从OBS下载需要的文件。如果将文件上传到私有文件系统，需要用**hadoop fs**下载，下面的例子会将 obs://yourbucket/myfile.tar.gz 这个文件下载到本地，并解压到 /your-dir 目录下：

```
#!/bin/bash
source /opt/Bigdata/client/bigdata_env;hadoop fs -D fs.obs.endpoint=<obs-endpoint> -D
fs.obs.access.key=<your-ak> -D fs.obs.secret.key=<your-sk> -copyToLocal obs://yourbucket/
myfile.tar.gz ./
mkdir -p /<your-dir>
tar -zxvf myfile.tar.gz -C /<your-dir>
```

说明

- MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。
- Hadoop客户端已预安装在MRS节点上，**hadoop fs**命令可对OBS做下载、上传等操作。
- 获取各region下obs-endpoint。

上传脚本至 OBS 文件系统

脚本完成后上传到同region的OBS文件系统中。在您选定的时机，集群各节点会从OBS将脚本下载下来并以root用户执行。

4.9.3 查看执行记录

您可以在集群详情页选择“引导操作” 页签查看引导操作的执行结果。

查看执行结果

1. 登录MRS管理控制台。
2. 在“集群列表 > 现有集群”中单击需要查询的集群名称。

   系统跳转至该集群详情页面。
3. 在集群详情页面选择“引导操作”页签。系统显示创建集群时所添加的引导操作信息。

   说明

   - 可以通过选择右上角的“组件首次启动前”或者“组件首次启动后”查询相关的引导操作信息。
   - 这里列出的是上次执行结果。对于新创建的集群，则列出的是创建时执行引导操作的记录；如果集群被扩容了，则列出的是上次扩容对新增节点执行引导操作的记录。

查看执行日志

如果需要查看引导操作的执行日志，请在添加引导操作时将“失败操作”配置为“继续”，然后登录到各个节点上查看运行日志，运行日志在/var/log/Bootstrap目录下。如果您对组件启动前后都添加了引导操作，可通过时间戳前后关系来区分两个阶段引导操作的日志。

建议您在脚本中尽量详细地打印日志，以方便查看运行结果。MRS将脚本的标准输出和错误输出都重定向到了引导操作日志目录下。

4.9.4 添加引导操作

该操作适用于MRS 3.x之前版本集群。

MRS 3.x版本暂不支持在创建集群时添加引导操作。

在创建集群时添加引导操作

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在集群页面，选择“自定义创建”。

步骤4 参考**创建自定义集群**配置集群软件配置和硬件配置信息。

步骤5 在“高级配置”区域的引导操作栏，单击“添加”。

表 4-16 参数描述

参数	说明
名称	引导操作脚本的名称。 只能由数字、英文字符、空格、中划线和下划线组成，且不能以空格开头。 可输入的字符串长度为1～64个字符。 **说明** 同一集群内，不允许配置相同的名称。不同集群之间，可以配置相同的名称。
脚本路径	脚本的路径。路径可以是OBS文件系统的路径或虚拟机本地的路径。 • OBS文件系统的路径，必须以s3a://开头，以.sh结尾。例如：s3a://mrs-samples/xxx.sh • 虚拟机本地的路径，脚本所在的路径必须以‘/’开头，以.sh结尾。
参数	引导操作脚本参数。
执行节点	选择引导操作脚本所执行的节点类型。
执行时机	选择引导操作脚本执行的时间。 • 组件首次启动前 • 组件首次启动后
失败操作	该脚本执行失败后，是否继续执行后续脚本和创建集群。 **说明** 建议您设置为“继续”，无论此引导操作是否执行成功，则集群都能继续创建。

步骤6 单击“确定”。

添加成功后，可以通过“操作”列进行编辑、克隆和删除。

----结束

在弹性伸缩集群页面添加自动化脚本

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 在“节点管理”页签Task节点组的“操作”列单击“弹性伸缩”，进入“弹性伸缩”页面。

当集群没有Task节点时，先单击“配置Task节点”增加Task节点，再执行该步骤。

说明

对于MRS 3.x及之后版本，“配置Task节点”仅适用于分析集群、流试集群和混合集群。

步骤4 配置资源计划。

配置操作：

1. 在弹性伸缩页面，开启弹性伸缩功能。
2. 节点数量范围的“默认范围”设置为“2-2”，表示除资源计划规定时间范围外，其他时间Task节点数量固定为2个。
3. 单击默认范围下方的“配置指定时间段的节点数量范围”。
4. 配置“时间范围”和“节点数量范围”。此处“时间范围”设置为“07:00-13:00”，“节点数量范围”设置为“5-5”，表示在该时间范围内，Task节点数量固定为5个。参数详情请参考**表4-11**。

   可以单击“配置指定时间段的节点数量范围”配置多条资源计划。

步骤5 （可选）配置自动化脚本。

1. 在“高级配置”项选择“现在配置”。
2. 单击“添加”，进入“自动化脚本”配置页面。
3. 配置“名称”、“脚本路径”、“执行节点类型”、“参数”、“执行时机”、“失败操作”。参数详情请参考**表4-12**。
4. 单击“确定”保存自动化脚本配置。

步骤6 勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

步骤7 单击“确定”，完成弹性伸缩集群设置。

----结束

4.10 查看失败的集群操作任务

本章节介绍如何查看并删除失败的MRS任务。

背景信息

当集群创建失败、集群删除失败、集群扩容失败和集群缩容失败后，失败任务会转入“失败任务管理”页面，其中仅集群删除失败的任务会同步转入“历史集群”页面。当不需要失败的任务时，可以删除。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”。

步骤3 单击“失败任务管理”右侧的或数字，进入“失败任务管理”页面。

步骤4 在需要删除的任务对应的“操作”列中，单击“删除”。

此处只能删除单个失败的任务。

步骤5 单击任务列表左上方的“删除所有失败任务”可以删除全部任务。

----结束

4.11 查看历史集群信息

选择“集群列表 > 历史集群”，选中一集群并单击集群名，进入集群基本信息页面。用户可查看集群的配置信息、部署的节点信息。

参考下列表格查看集群信息参数说明。

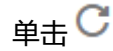
表 4-17 集群基本信息

参数	参数说明
集群名称	集群的名称，创建集群时设置。
集群状态	集群状态信息。
集群版本	集群的版本信息。
集群类型	创建集群时的集群类型。
集群ID	集群的唯一标识，创建集群时系统自动赋值，不需要用户设置。
创建时间	显示集群创建的时间。
可用区	集群工作区域下的可用区，创建集群时设置。
默认生效子网	子网信息，创建集群时所选。 通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。
虚拟私有云	VPC信息，创建集群时所选。 VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。
OBS权限控制	单击“单击管理”，修改MRS用户与OBS权限的映射关系，具体请参考**配置MRS多用户访问OBS细粒度权限**。
数据连接	单击“单击管理”，查看集群关联的数据连接类型，具体请参考**配置数据连接**。
委托	单击“管理委托”，为集群绑定或修改委托。 通过绑定委托，您可以将部分资源共享给ECS或BMS云服务来管理，例如通过配置ECS委托可自动获取AK/SK访问OBS，具体请参见**配置存算分离集群（委托方式）**。 **MRS_ECS_DEFAULT_AGENCY**委托拥有对象存储服务的OBSOperateAccess权限和在集群所在区域拥有CESFullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。
密钥对	密钥对名称，创建集群时设置。 如果创建集群时设置的登录方式为密码，则不显示。

参数	参数说明
Keberos认证	登录Manager管理页面时是否启用Kerberos认证。
企业项目	集群所属的企业项目，仅现有集群列表支持单击企业项目名称进入对应项目的企业项目管理页面。
安全组	集群的安全组名称。
流式Core节点LVM管理	流式Core节点的LVM管理功能是否开启。
数据盘密钥名称	用于加密数据盘的密钥名称。如需对已使用的密钥进行管理，请登录密钥管理控制台进行操作。
数据盘密钥ID	用于加密数据盘的密钥ID。
组件版本	集群安装各组件的版本信息。
License版本	集群的License版本信息。
委托	通过绑定委托，ECS或BMS云服务将有权限来管理您的部分资源。

返回到历史集群列表页面，用户可使用如下按钮进行操作，参考下列表格查看按钮说明。

表 4-18 按钮说明

按钮	说明
	单击  ，手动刷新节点信息。
	在搜索框中输入集群名称或ID，单击  ，搜索集群。

5 管理集群

5.1 登录集群

5.1.1 MRS 集群节点简介快速创建 Hadoop 分析集群

介绍远程登录的概念、MRS集群的节点类型和节点功能。

MRS集群节点支持用户远程登录，远程登录包含界面登录和SSH登录两种方式：

- 界面登录：直接通过弹性云服务器管理控制台提供的远程登录功能，登录到集群Master节点的Linux界面。
- SSH登录：仅适用于Linux弹性云服务器。您可以使用远程登录工具（例如PuTTY），登录弹性云服务器。此时，需要该弹性云服务器绑定弹性IP地址。

  Master节点申请和绑定弹性IP，请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

  可以使用密钥方式也可以使用密码方式登录Linux弹性云服务器。

  须知

  当您使用密钥方式访问集群节点，需要以root用户登录，详细步骤请参见**登录弹性云服务器（SSH密钥方式）**。

  当您使用密码方式访问集群节点，详细步骤请参见**登录弹性云服务器（SSH密码方式）**。

MRS集群中每个节点即为一台弹性云服务器，节点类型及节点功能如**表5-1**所示。

表 5-1 集群节点分类

节点类型	功能
Master节点	MRS集群管理节点，负责管理和监控集群。在MRS管理控制台选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群信息页面。在“节点管理”中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。 Master节点可以通过弹性云服务器界面的VNC方式登录，也可以通过SSH方式登录，并且Master节点可以免密码登录到Core节点。 系统自动将Master节点标记为主备管理节点，并支持MRS集群管理的高可用特性。如果主管理节点无法提供服务，则备管理节点会自动切换为主管理节点并继续提供服务。 查看Master1节点是否为主管理节点，请参见**如何确认Manager的主备管理节点**。
Core节点	MRS集群工作节点，负责处理和分析数据，并存储过程数据。
Task节点	计算节点，用于弹性伸缩，集群计算资源不足时扩容至集群中。

5.1.2 登录集群节点

本章节介绍如何使用弹性云服务器管理控制台上提供的远程登录（VNC方式）和如何使用密钥或密码方式（SSH方式）登录MRS集群中的节点，远程登录主要用于紧急运维场景，远程登录弹性云服务器进行相关维护操作。其他场景下，建议用户采用SSH方式登录。

说明

如果需要使用SSH方式登录集群节点，需要在集群的安全组规则中手动添加入方向规则：其中源地址为“客户端IPV4地址/32(或者客户端IPV6地址/128)”，端口为22，具体请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

登录弹性云服务器（VNC 方式）

步骤1 登录MapReduce服务管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群基本信息页面。

步骤3 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。

步骤4 单击右上角的“远程登录”。

步骤5 根据界面提示，输入Master节点的用户名和密码。

1. 创建集群时登录方式选择了“密码”。此时，你需要输入的用户名、密码分别是root和创建集群时设置的密码。

2. 创建集群时登录方式选择了密钥对，则使用如下方式登录：
   a. 创建集群成功后，参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”为集群的Master节点绑定一个弹性IP地址。
   b. 使用root用户名和密钥文件，SSH方式远程登录Master节点。
   c. 执行**passwd root**命令，设置root用户密码。
   d. 设置成功后，返回界面登录方式，输入root用户名和**步骤5.2.c**设置的密码，登录节点。

----结束

登录弹性云服务器（SSH 密钥方式）

本地使用Windows操作系统

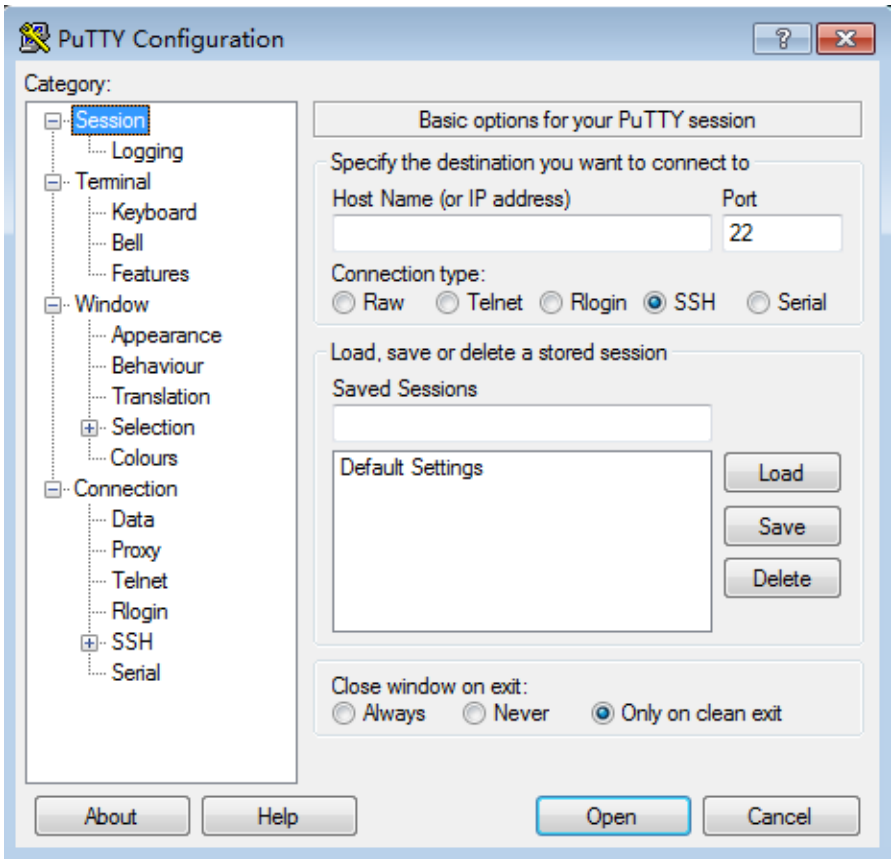
如果您本地使用Windows操作系统登录Linux弹性云服务器，可以按照下面方式登录弹性云服务器。下面步骤以PuTTY为例。

1. 登录MapReduce服务管理控制台。
2. 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群基本信息页面。
3. 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。
4. 选择“弹性公网IP”页签，单击“绑定弹性IP”为弹性云服务器绑定一个弹性公网IP并记录该IP地址，若已绑定弹性公网IP请跳过该步骤。
5. 判断私钥文件是否为.ppk格式。
   - 是，执行**10**。
   - 否，执行**6**。
6. 运行PuTTY。
7. 在“Actions”区域，单击“Load”，并导入创建弹性云服务器时使用的密钥对的私钥文件。
   导入时注意确保导入的格式要求为“All files (*.*)”。
8. 单击“Save private key”。
9. 保存转化后的私钥到本地。例如：kp-123.ppk。
10. 运行PuTTY。
11. 选择“Connection > data”，在Auto-login username处输入镜像的用户名。

    说明

    集群节点镜像的用户名是root。
12. 选择“Connection > SSH > Auth”，在最下面一个配置项“Private key file for authentication”中，单击“Browse”，选择**9**转化的密钥。
13. 单击“Session”。
    a. Host Name (or IP address)：输入弹性云服务器所绑定的弹性公网IP。
    b. Port：输入 22。
    c. Connection Type：选择 SSH。
    d. Saved Sessions：任务名称，在下一次使用putty时就可以单击保存的任务名称，即可打开远程连接。

图 5-1 单击“Session”



14. 单击“Open”登录云服务器。

    如果首次登录云服务器，PuTTY会显示安全警告对话框，询问是否接受服务器的安全证书。单击“是”将证书保存到本地注册表中。

本地使用Linux操作系统

如果您本地使用Linux操作系统登录Linux弹性云服务器，可以按照下面方式登录。下面步骤以私钥文件以kp-123.pem为例进行介绍。

1. 在您的linux计算机的命令行中执行如下命令，变更权限。

   chmod 400 /*path*/kp-123.pem

   说明

   上述令的path为密钥文件的存放路径。

2. 执行如下命令，登录弹性云服务器。

   ssh -i /*path*/kp-123.pem*默认用户名@弹性公网IP*

   假设Linux弹性云服务器的默认用户名是root，弹性公网IP为123.123.123.123，则命令如下：

   ssh -i /*path*/kp-123.pem root@123.123.123.123

   说明

   - path为密钥文件的存放路径。
   - 弹性公网IP地址为弹性云服务器绑定的弹性公网IP地址。
   - 集群节点镜像的用户名是root。

登录弹性云服务器（SSH 密码方式）

本地使用Windows操作系统

如果本地主机为Windows操作系统，可以按照下面方式登录弹性云服务器。下面步骤以PuTTY为例。

步骤1 登录MapReduce服务管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群基本信息页面。

步骤3 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。

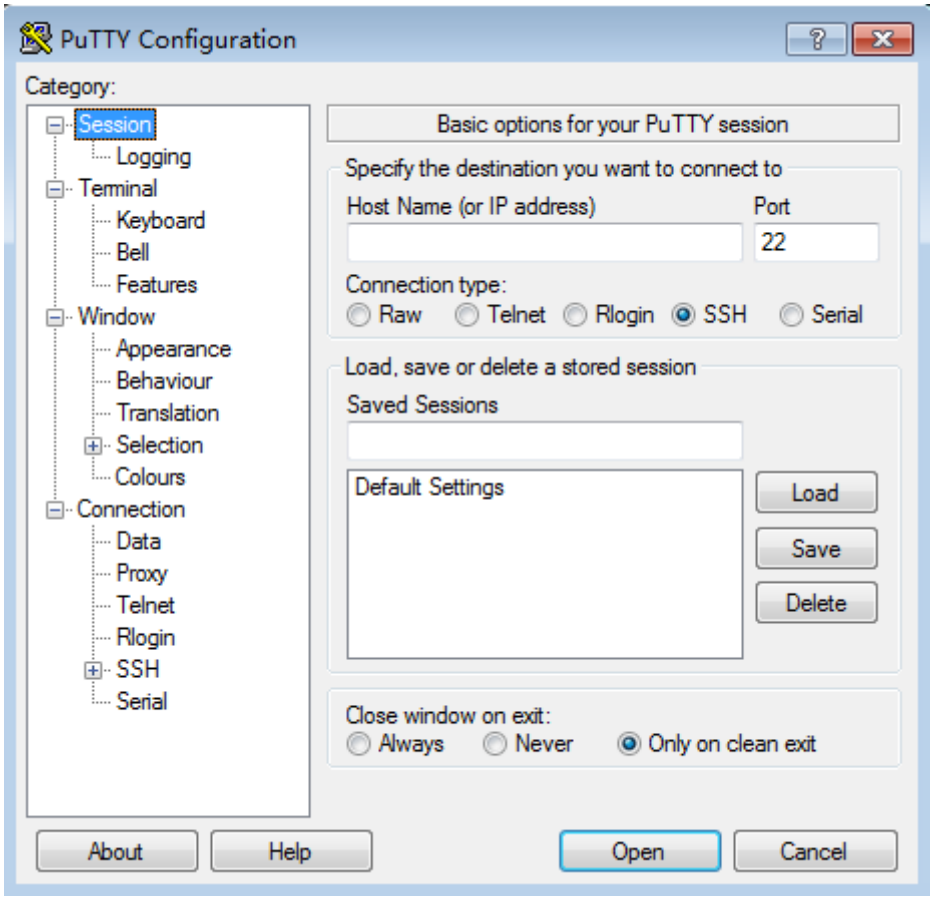
步骤4 选择“弹性公网IP”页签，单击“绑定弹性IP”为弹性云服务器绑定一个弹性公网IP并记录该IP地址，若已绑定弹性公网IP请跳过该步骤。

步骤5 运行PuTTY。

步骤6 单击“Session”。

1. Host Name (or IP address)：输入弹性云服务器所绑定的弹性公网IP。
2. Port：输入 22。
3. Connection Type：选择 SSH。
4. Saved Sessions：任务名称，在下一次使用PuTTY时就可以单击保存的任务名称，即可打开远程连接。

图 5-2 单击 Session



步骤7 单击“Window”，在“Translation”下的“Remote character set:”选择“UTF-8”。

步骤8 单击“Open”登录云服务器。

如果首次登录云服务器，PuTTY会显示安全警告对话框，询问是否接受服务器的安全证书。单击“是”将证书保存到本地注册表中。

步骤9 建立到云服务器的SSH连接后，根据提示输入用户名和密码登录弹性云服务器。

说明

用户名、密码分别是root和创建集群时设置的密码。

----结束

本地使用Linux操作系统

如果本地主机为Linux操作系统，您可以参考**步骤1**~**步骤4**为弹性云服务器绑定弹性公网IP后，在计算机的命令行中运行如下命令登录弹性云服务器：**ssh *弹性云服务器绑定的弹性公网IP***

5.1.3 如何确认 Manager 的主备管理节点

介绍如何在Master1节点中确认Manager的主备管理节点。

背景信息

用户可以在Master节点登录到集群中的其他节点，同时登录Master节点后，可以确认Manager的主备管理节点，并在对应的管理节点中执行命令。

在主备模式下，由于Master1和Master2之间会切换，Master1节点不一定是Manager的主管理节点。

操作步骤

步骤1 确认MRS集群的Master节点。

1. 登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群信息页面。查看指定的集群信息。
2. 在“节点管理”中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。

步骤2 确认Manager的主备管理节点。

1. 远程登录Master1节点，请参见**登录集群节点**。

   Master节点支持Cloud-Init特性，Cloud-init预配置的用户名“root”，密码为创建集群时设置的密码。
2. 执行以下命令切换用户。

   sudo su - root

   su - omm
3. 执行以下命令确认主备管理节点：

   MRS 3.x之前版本集群执行命令：sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh

   MRS 3.x及之后版本集群执行命令：sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh

回显信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“mgtomsdat-sh-3-01-1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“mgtomsdat-sh-3-01-2”为备管理节点）。

```
Ha mode
double
NodeName            HostName                HAVersion       StartTime           HAActive
HAAllResOK          HARunPhase
192-168-0-30        mgtomsdat-sh-3-01-1     V100R001C01     2014-11-18 23:43:02
active              normal                  Actived
192-168-0-24        mgtomsdat-sh-3-01-2     V100R001C01     2014-11-21 07:14:02
standby             normal                  Deactived
```

说明

如果当前登录的Master1节点是备管理节点，且需要登录到主管理节点时，请执行以下命令：

ssh *Master2节点IP地址*

----结束

5.2 集群概览

5.2.1 集群列表简介

通过概览信息可以快速查看所有集群和作业的状态，您也可以通过MRS控制台左侧导航的“帮助”中获取MRS相关帮助文档。

MRS用于海量数据的管理和分析，MRS使用简单，用户创建好一个集群，在集群内可新增MapReduce、Spark和Hive等作业，对用户数据进行分析和处理。数据处理完成后，采用SSL加密传输数据至OBS，保证数据的完整性和机密性。

集群状态

登录MRS管理控制台后，MRS所有集群包含的状态如表5-2所示。

表 5-2 集群状态说明

状态	说明
启动中	集群正在创建，则其状态为“启动中”。
运行中	集群创建成功且集群中所有组件状态均正常，则其状态为“运行中”。
扩容中	集群Core节点或者Task节点正在扩容，则其状态为“扩容中”。 **说明** 如果集群扩容失败，用户可重新进行扩容操作。
缩容中	当对集群节点进行关机、删除、变更OS、重装OS和修改规格的操作时，被变更的集群节点正在删除，则其状态为“缩容中”。
异常	集群中部分组件状态异常，导致集群异常，则其状态为“异常”。
删除中	集群节点正在删除中，则其状态为“删除中”。

状态	说明
已删除	集群已经删除，仅“历史集群”会显示此参数。

作业状态

登录MRS的管理控制台后，用户在MRS集群中执行的作业包含的状态如表5-3所示。

表 5-3 作业状态说明

状态	说明
已接受	作业提交成功后的初始状态。
运行中	作业执行过程中，则其状态为“运行中”。
已完成	作业执行完成，并且执行成功，则其状态为“已完成”。
已终止	作业执行过程中，停止执行，则其状态为“已终止”。
异常	作业执行过程中报错，或者作业执行完成，但执行失败，则其状态为“异常”。

5.2.2 查看集群状态

集群列表显示MRS所有的集群，集群数量较多时，可采用翻页显示，您可以查看任何状态下的集群。

MRS作为一个海量数据管理和分析平台，数据处理能力在PB级以上。MRS支持创建多个集群，集群数量受弹性云服务器数量限制。

集群列表默认按时间顺序排列，时间最近的集群显示在最前端。集群列表参数说明如表5-4所示。

- 现有集群：包括除了“失败”和“已删除”状态以外的所有集群。
- 历史集群：仅包含“已删除”的集群，目前界面只显示6个月内创建且已删除的集群，若需要查看6个月以前删除的集群，请联系支持人员。
- 失败任务管理：仅包含“失败”状态的任务。
  - 集群创建失败的任务
  - 集群删除失败的任务
  - 集群扩容失败的任务
  - 集群缩容失败的任务
  - 集群安装补丁失败的任务（仅MRS 3.x之前版本支持）
  - 集群卸载补丁失败的任务（仅MRS 3.x之前版本支持）
  - 集群升级规格失败的任务

表 5-4 集群列表参数

参数	参数说明
名称/ID	集群的名称，创建集群时设置。集群的ID是集群的唯一标识，创建集群时系统自动赋值，不需要用户设置。 ●  ：修改集群名称。 ●  ：复制集群ID。
集群版本	集群的版本号。
节点数	集群部署的节点个数，创建集群时设置。
状态	集群状态、进度信息。 创建集群进度包括： ● Verifying cluster parameters：校验集群参数中 ● Applying for cluster resources：申请集群资源中 ● Creating VMs：创建虚拟机中 ● Initializing VMs：初始化虚拟机中 ● Installing MRS Manager：安装MRS Manager中 ● Deploying the cluster：部署集群中 ● Cluster installation failed：集群安装失败 扩容集群进度包括： ● Preparing for cluster expansion：准备扩容中 ● Creating VM：创建虚拟机中 ● Initializing VM：初始化虚拟机中 ● Adding node to the cluster：节点加入集群中 ● Cluster expansion failed：集群扩容失败 缩容集群进度包括： ● Preparing for cluster shrink：正在准备缩容 ● Decommissioning instance：实例退服中 ● Deleting VM：删除虚拟机中 ● Deleting node from the cluster：从集群删除节点中 ● Cluster shrink failed：集群缩容失败 集群安装、扩容、缩容失败，会显示失败的原因，详情请参见**表 4-5**。
创建时间	集群节点创建成功。
删除时间	集群节点停止时间，也是集群节点开始删除时间。仅“历史集群”会显示此参数。
可用区	集群工作区域下的可用区，创建集群时设置。
企业项目	集群所属的企业项目。

参数	参数说明
操作	删除：如果作业执行结束后不需要集群，可以单击“删除”，集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。当MRS集群部署失败时，集群会被自动删除。 仅“现有集群”会显示此参数。 **说明** 一般在数据完成分析和存储后或集群异常无法提供服务时才执行删除操作。如果数据没有完成处理分析，删除集群会导致数据丢失，请谨慎操作。

表 5-5 按钮说明

按钮	说明
	在下拉框中选择企业项目，筛选对应集群。
	在下拉框中选择集群状态，筛选现有集群。 ● 现有集群 – 所有状态：表示筛选所有的现有集群 – 启动中：表示筛选“启动中”状态的现有集群 – 运行中：表示筛选“运行中”状态的现有集群 – 扩容中：表示筛选“扩容中”状态的现有集群。 – 缩容中：表示筛选“缩容中”状态的现有集群。 – 异常：表示筛选“异常”状态的现有集群 – 删除中：表示筛选“删除中”状态的现有集群
	选择“集群列表 > 现有集群”，单击  进入“失败任务管理”页面。  *Num*：表示“失败”状态的任务数。
	在搜索框中输入集群名称或ID，单击  ，搜索集群。
标签搜索	单击“标签搜索”输入待查询集群的标签，然后单击“搜索”搜索对应集群。 标签键或标签值可以通过下拉列表中选择，当标签键或标签值全匹配时，系统可以自动查询到目标集群。当有多个标签条件时，会取各个标签的交集，进行集群查询。
	单击  ，手动刷新现有集群列表。

5.2.3 查看集群基本信息

集群创建完成后，可对集群进行监控和管理。选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群详情页面，查看集群的基本配置信息、部署的节点信息。

说明

ECS集群和BMS集群在管理控制台操作基本一致，本文档主要以ECS集群描述为例，如有操作区别则分开描述。

在集群详情页面选择“概览”，参考**表5-6**查看集群详情概览信息参数说明。

表 5-6 集群基本信息

参数	参数说明
集群名称	集群的名称，创建集群时设置。单击 可对集群名称进行修改。 当MRS集群为MRS 3.x之前版本时，修改集群名称后仅MRS管理控制台界面显示的集群名称修改，MRS Manager中集群名称不会同步修改。
集群状态	集群状态信息，请参见**表5-2**。
集群管理页面	Manager页面入口。 • 针对MRS 3.x及以后版本，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）** • 针对MRS 3.x之前版本，需要根据提示绑定弹性公网IP及添加安全组规则后才能进入MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**。
集群版本	MRS版本信息。
集群类型	支持以下集群类型： • 分析集群：用来做离线数据分析，提供的是Hadoop体系的组件。 • 流式集群：用来做流处理任务，提供的是流式处理组件。 • 混合集群：既可以用来做离线数据分析，也可以用来做流处理任务，提供的是Hadoop体系的组件和流式处理组件。 • 自定义：全量自定义组件组合的MRS集群，MRS 3.x及之后版本支持此类型。
集群ID	集群的唯一标识，创建集群时系统自动赋值，不需要用户设置。
创建时间	显示集群创建的时间。
可用区	集群工作区域下的可用区，创建集群时设置。

参数	参数说明
默认生效子网	子网信息，创建集群时所选。 当子网IP不足时，单击“切换子网”切换到当前集群相同VPC下的其他子网，实现可用子网IP的扩充。切换子网不会影响当前已有节点的IP地址和子网。 通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。
虚拟私有云	VPC信息，创建集群时所选。 VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。
弹性公网IP	通过将弹性公网IP与MRS集群绑定，实现使用弹性公网IP访问Manager的目的。
OBS权限控制	单击“单击管理”，修改MRS用户与OBS权限的映射关系，具体请参考**配置MRS多用户访问OBS细粒度权限**。
数据连接	单击“单击管理”，查看集群关联的数据连接类型，具体请参考**配置数据连接**。
委托	单击“管理委托”，为集群绑定或修改委托。 通过绑定委托，您可以将部分资源共享给ECS或BMS云服务来管理，例如通过配置ECS委托可自动获取AK/SK访问OBS，具体请参见**配置存算分离集群（委托方式）**。 **MRS_ECS_DEFAULT_AGENCY**委托拥有对象存储服务的OBS OperateAccess权限和在集群所在区域拥有CES FullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。
密钥对	密钥对名称，创建集群时设置。 如果创建集群时设置的登录方式为密码，则不显示。
Kerberos认证	登录Manager管理页面时是否启用Kerberos认证。
日志记录	用于收集集群创建失败及扩缩容失败的日志。
企业项目	集群所属的企业项目，仅现有集群列表支持单击企业项目名称进入对应项目的企业项目管理页面。
安全组	集群的安全组名称。
流式Core节点LVM管理	流式Core节点的LVM管理功能是否开启。
数据盘密钥名称	用于加密数据盘的密钥名称。如需对已使用的密钥进行管理，请登录密钥管理控制台进行操作。
数据盘密钥ID	用于加密数据盘的密钥ID。

参数	参数说明
IAM用户同步	可以将IAM侧用户信息同步至MRS集群，用于集群管理。具体请参见**IAM用户同步MRS说明**。 **说明** 集群详情页的“组件管理”、“租户管理”和“备份恢复”页签需要同步用户后方可使用。MRS 3.x版本集群同步后可使用“组件管理”。
通讯安全授权	展示安全授权状态，通过  可关闭和开启安全授权。关闭安全授权属于高危操作，请谨慎处理。详细信息请参考**授权安全通信**。

表 5-7 组件版本

参数	参数说明
Hadoop 版本	显示Hadoop组件的版本信息。
Spark 版本	显示Spark组件的版本信息，MRS 3.x之前版本集群支持。
HBase 版本	显示HBase组件的版本信息。
Hive 版本	显示Hive组件的版本信息。
Hue 版本	显示Hue组件的版本信息。
Loader 版本	显示Loader组件的版本信息。
Kafka 版本	显示Kafka组件的版本信息。
Storm 版本	显示Storm组件的版本信息。
Flume 版本	显示Flume组件的版本信息。
Tez 版本	显示Tez组件的版本信息。
Presto 版本	显示Presto组件的版本信息。
KafkaManager 版本	显示KafkaManager组件的版本信息。
Flink 版本	显示Flink组件的版本信息。
Alluxio版本	显示Alluxio组件的版本信息。
Ranger 版本	显示Ranger组件的版本信息。
Impala版本	显示Impala组件的版本信息。
Kudu版本	显示Kudu组件的版本信息。
Spark2x 版本	显示Spark2x组件的版本信息。仅MRS 3.x及之后版本集群支持。
Oozie 版本	显示Oozie组件的版本信息。仅MRS 3.x及之后版本集群支持。
ClickHouse版本	显示ClickHouse组件的版本信息。仅MRS 3.x及之后版本集群支持。

在集群详情页面选择“节点管理”，参考表5-8查看集群节点信息参数说明。

表 5-8 节点信息

参数	参数说明
配置Task节点	用于增加Task节点，请参见**添加Task节点**的相关任务。 对于3.x及之后版本，该操作仅适用于分析集群、流试集群和混合集群。
新增节点组	仅适用于3.x及之后版本，用于增加节点组，请参见**添加节点组**，仅适用自定义集群。
节点组名称	集群节点组名称。
节点类型	节点类型： • Master：集群主节点，负责管理集群，协调将MapReduce可执行文件分配到核心节点。此外，还会跟踪每个作业的执行状态，监控DataNode的运行状况。 • Task类型节点组是指仅部署了不存储数据的数据角色的节点组，主要包含：NodeManager、ThriftServer、Thrift1Server、RESTServer、Supervisor、Logviewer、HBaseIndexer、TagSync。 • 如果节点组内除以上角色外还部署了其他角色，则该节点组为Core类型节点组。 单击节点组名称前方的 ，显示该节点组包含的节点，单击节点名称，使用创建集群时配置的密码或者密钥对远程登录弹性云服务器。节点参数说明请参见**管理组件和主机监控**。
节点数	对应节点组中包含的节点数量。
操作	• 扩容：请参见**扩容集群**。 • 缩容：请参见**缩容集群**。 • 弹性伸缩：请参见**配置弹性伸缩规则**。 • 查看角色信息：可查看所在节点组部署的角色信息。仅适用于3.x及之后版本的自定义集群。

5.2.4 查看集群补丁信息

查看集群组件的补丁信息。如果集群组件，如Hadoop或Spark等出现了异常，可下载补丁版本，选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群详情页面升级组件，修复问题。

说明

MRS 3.x版本无补丁版本信息，不涉及此章节。

- 补丁名称：补丁包的名称。
- 发布时间：补丁包发布的时间。

- 状态：展示补丁的状态。
- 补丁内容：补丁版本的描述信息。
- 操作：可安装或者卸载补丁。

5.2.5 查看和定制集群监控指标

MRS支持将集群中所有部署角色的节点，按管理节点、控制节点和数据节点进行分类，分别计算关键主机监控指标在每类节点上的变化趋势，并在报表中按用户自定义的周期显示分布曲线图。如果一个主机属于多类节点，那么对应的指标将被统计多次。

该任务指导用户了解MRS集群的概览、及在MRS查看、自定义与导出节点监控指标报表。

方式一：（适用于MRS 3.x之前版本集群）

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 选择“概览”页签，即可在页面下方查看到集群主机健康状态统计。

步骤3 如需查看或导出其它指标的报表，请选择页面左侧基本信息栏的“集群管理页面 > 前往 Manager”登录Manager页面，具体请参见**访问集群Manager**。

步骤4 在Manager页面查看、定制与导出节点监控指标报表，具体请参见**系统概览**。

----结束

方式二：

步骤1 登录MRS控制台。

步骤2 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤3 在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

步骤4 用户同步完成后即可在页面右侧查看到集群的监控指标报表。

步骤5 在时间区间选择需要查看监控数据的时间段。可供选择的选项如下：

- 近1小时
- 近3小时
- 近12小时
- 近24小时
- 近7天
- 近一个月
- 自定义：在时间范围内自行选择需要查看的时间。

步骤6 自定义监控指标报表。

1. 单击“定制”，勾选需要显示的监控指标。

   MRS支持统计的指标共14个，界面最多显示12个定制的监控指标。

   - 集群主机健康状态统计
   - 集群网络读速率统计
   - 主机网络读速率分布

- 主机网络写速率分布
- 集群磁盘写速率统计
- 集群磁盘占用率统计
- 集群磁盘信息
- 主机磁盘占用率统计
- 集群磁盘读速率统计
- 集群内存占用率统计
- 主机内存占用率分布
- 集群网络写速率统计
- 主机CPU占用率分布
- 集群CPU占用率统计

2. 单击“确定”保存并显示所选指标。

说明

单击“清除”可批量取消全部选中的指标项。

步骤7 导出监控指标报表。

1. 选择报表的时间范围。可供选择的选项如下：
   - 近1小时
   - 近3小时
   - 近12小时
   - 近24小时
   - 近7天
   - 近一个月
   - 自定义：在时间范围内自行选择需要查看的时间。
2. 单击“导出”，MRS将生成指定时间范围内、已勾选的集群监控指标报表文件，请选择一个位置保存，并妥善保管该文件。

----结束

方式三：（适用于MRS 3.x及之后版本集群）

步骤1 登录MRS控制台。

步骤2 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤3 在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

步骤4 用户同步完成后即可在页面右侧查看到集群的监控指标报表。

步骤5 在时间区间选择需要查看监控数据的时间段。可供选择的选项如下：

- 近1小时
- 近3小时
- 近12小时
- 近24小时
- 近7天

- 近1个月
- 自定义：在时间范围内自行选择需要查看的时间。

步骤6 自定义监控指标报表。

1. 单击“定制”，勾选需要显示的监控指标。
   界面最多显示12个定制的监控指标。
2. 单击“确定”保存并显示所选指标。

   说明

   单击“清除”可批量取消全部选中的指标项。

----结束

5.2.6 管理组件和主机监控

用户在日常使用中，可以在MRS管理所有组件（含角色实例）和主机的状态及指标信息：

- 状态信息，包括运行、健康、配置及角色实例状态统计。
- 指标信息，各组件的主要监控指标项。
- 导出监控指标（MRS 3.x及之后版本暂不支持）。

说明

- 操作方法请参考**管理服务和主机监控**。
- MRS 3.x及之后版本操作方法请参考**操作方法**。
- 用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。
- 组件管理支持三种参数值：
  - “每30秒刷新一次”：刷新间隔30秒。
  - “每60秒刷新一次”：刷新间隔60秒。
  - “停止”：停止刷新。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作方法

管理组件监控

说明

MRS Manager操作，请参考**管理服务监控**操作。

步骤1 在MRS集群详情页面，单击“组件管理”。

组件列表中标题包含“服务”、“操作状态”、“健康状态”、“配置状态”、“角色数”和“操作”。

- 服务操作状态描述如**表5-9**所示。

表 5-9 服务操作状态

状态	描述
已启动	服务已启动。
已停止	服务已停止。
启动失败	用户启动操作失败。
停止失败	用户停止操作失败。
未知	后台系统重启后，服务的初始状态。

- 服务健康状态如表5-10所示。

表 5-10 服务健康状态

状态	描述
良好	该服务中所有角色实例正常运行。
故障	至少一个角色实例运行状态为“故障”或被依赖的服务状态不正常。
未知	该服务中所有角色实例状态为“未知”。
正在恢复	后台系统正在尝试自动启动服务。
亚健康	该服务所依赖的服务状态不正常，异常服务的相关接口无法被外部调用。

- 服务配置状态如表5-11所示。

表 5-11 服务配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
配置超期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
配置失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
配置中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“服务”列按升序排列，单击**服务**、**操作状态**、**健康状态**或**配置状态**可修改排列方式。

步骤2 单击列表中指定服务名称，查看服务状态及指标信息。

步骤3 定制、查看监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。

----结束

管理角色实例监控

说明

针对MRS 3.x之前版本，请参考**管理角色实例监控**操作。

步骤1 在MRS集群详情页面，单击“组件管理”，在组件列表中单击服务指定名称。

步骤2 单击“实例”，查看角色状态。

角色实例列表中包含实例信息的**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**运行状态**及**配置状态**。

- 角色实例的运行状态如**表5-12**所示。

表 5-12 角色实例运行状态

状态	描述
良好	表示实例当前运行正常。
故障	表示实例当前无法正常工作。
已退服	表示实例处于退服状态。
未启动	表示实例已停止。
未知	表示实例的初始状态信息无法检测。
正在启动	表示实例正在执行启动过程。
正在停止	表示实例正在执行停止过程。
正在恢复	表示实例可能存在异常正在自动修复。
正在退服	表示实例正在执行退服过程。
正在入服	表示实例正在执行入服过程。
启动失败	表示实例启动操作失败。
停止失败	表示实例停止操作失败。

- 角色实例的配置状态如**表5-13**所示。

表 5-13 角色实例配置状态

状态	描述
已同步	系统中最新的配置信息已生效。

状态	描述
配置超期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
配置失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
配置中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“角色”列按升序排列，单击**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**运行状态**或**配置状态**可修改排列方式。

支持在“角色”筛选相同角色的全部实例。

单击“高级搜索”，在角色搜索区域中设置搜索条件，单击“搜索”，查看指定的角色信息。单击“重置”清除输入的搜索条件。支持模糊搜索条件的部分字符。

步骤3 单击列表中指定角色实例名称，查看角色实例状态及指标信息。

步骤4 定制、查看监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。

----结束

管理主机监控

说明

针对MRS 3.x之前版本，请参考**管理主机监控**操作。

步骤1 在MRS集群详情页面，单击“节点管理”并展开节点组信息，查看所有主机状态。

主机列表中包括**节点名称**、**IP**、**机架**、**操作状态**、**健康状态**、**CPU使用率**、**内存使用率**、**磁盘使用率**、**网络速度**、**规格名**、**规格**、**可用区**。

- 主机操作状态如**表5-14**所示。

  表 5-14 主机操作状态

状态	描述
正常	主机及主机上的服务角色正常运行。
已隔离	主机被用户隔离，主机上的服务角色停止运行。

- 主机健康状态描述如**表5-15**所示。

表 5-15 主机健康状态

状态	描述
良好	主机心跳检测正常。
故障	主机心跳超时未上报。
未知	执行添加操作时，主机的初始状态。

默认以“节点名称”列按升序排列，单击**节点名称**、**IP**、**机架**、**操作状态**、**健康状态**、**CPU使用率**、**内存使用率**、**磁盘使用率**、**网络速度**、**规格名**或**规格**可修改排列方式。

步骤2 单击列表中指定的节点名称，查看单个节点状态及指标。

----结束

5.3 集群运维

5.3.1 导入导出数据

用户通过“文件管理”页面可以在分析集群进行文件夹创建、删除，文件导入、导出、删除操作，暂不支持文件创建功能。流式集群暂不支持在界面使用“文件管理”功能。开启Kerberos认证的集群中，根目录下的文件夹有权限限制，如需对其进行读写，请参考**创建角色**内容添加拥有对应文件夹权限的角色，再请参考**相关任务**修改提交作业用户所属的用户组，将新增的组件角色加入到该用户组中。

背景信息

MRS集群处理的数据源来源于OBS或HDFS，HDFS是Hadoop分布式文件系统（Hadoop Distributed File System），OBS即对象存储服务，是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用，同时可以通过REST API接口方式单独或集成到业务程序进行管理和访问数据。

用户创建作业前需要将本地数据上传至OBS系统，MRS使用OBS中的数据进行计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

导入数据

MRS目前只支持将OBS上的数据导入至HDFS中。上传文件速率会随着文件大小的增大而变慢，适合数据量小的场景下使用。

支持导入文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群信息页面。

3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。

   “bd_app1”目录仅为示例，可以是界面上的任何目录，也可以通过“新建”创建新的文件夹。

   新建文件夹时需要满足以下要求：
   - 文件夹名称小于等于255字符。
   - 不允许为空。
   - 不能包含 : /:*?"<>|\;&,'`!{}[]$%+特殊字符。
   - 不能以“.”开头或结尾。
   - 开头和末尾的空格会被忽略。
6. 单击“导入数据”，正确配置HDFS和OBS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。
   - OBS路径
     - 必须以“obs://”开头。
     - 不支持导入KMS加密的文件或程序。
     - 不支持导入空的文件夹。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
     - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
     - OBS全路径长度小于等于255字符。
   - HDFS路径
     - 默认以“/user”开头。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
     - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
     - HDFS全路径长度小于等于255字符。
7. 单击“确定”。

   文件上传进度可在“文件操作记录”中查看。MRS将数据导入操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

导出数据

数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。

支持导出文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群基本信息页面。

3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。
6. 单击“导出数据”，配置OBS和HDFS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。
   - OBS路径
     - 必须以“obs://”开头。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
     - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
     - OBS全路径长度小于等于255字符。
   - HDFS路径
     - 默认以“/user”开头。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
     - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
     - HDFS全路径长度小于等于255字符。

   说明

   当导出文件夹到OBS系统时，在OBS路径下，将增加一个标签文件，文件命名为“folder name_$folder$”。请确保导出的文件夹为非空文件夹，如果导出的文件夹为空文件夹，OBS无法显示该文件夹，仅生成一个命名为“folder name_$folder$”的文件。
7. 单击“确定”。

   文件上传进度可在“文件操作记录”中查看。MRS将数据导出操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

查看文件操作记录

通过MRS管理控制台导入和导出数据时，可在“文件管理 > 文件操作记录”查看数据导入、导出进度。

文件操作记录参数说明如表5-16所示。

表 5-16 文操作记录参数说明

Parameter	Description
提交时间	数据导入或导出操作的开始时间。
源目录	数据的源路径。 • 数据导入时“源目录”为OBS路径 • 数据导出时“源目录”为HDFS路径

Parameter	Description
目标目录	数据的目标路径。 • 数据导入时“目标目录”为HDFS路径 • 数据导出时“目标目录”为OBS路径
状态	数据导入或导出操作的状态。 • 已提交 • 已接受 • 运行中 • 已完成 • 已终止 • 异常
持续时间（分钟）	数据导入或导出操作的总时间。 单位：分钟
执行结果	数据导入或导出操作的结果。 • 成功 • 失败 • 终止 • 未定
操作	查看日志：查看文件操作日志。

5.3.2 切换集群子网

MRS支持当子网IP不足时，切换子网到当前集群相同VPC下的其他子网，实现可用子网IP的扩充。切换子网不会影响当前已有节点的IP地址和子网。

如需对网络ACL出规则进行配置请参考**如何配置网络ACL出规则？** 。

未关联网络 ACL 时切换子网

步骤1 登录MRS控制台。

步骤2 单击集群名称进入集群详情页。

步骤3 在“默认生效子网”右侧单击“切换子网”。

步骤4 选择待切换子网，并单击“确定”完成切换。

若没有可用子网，请单击“创建子网”进入VPC控制台创建子网后，再在此处引用。

----结束

关联网络 ACL 时切换子网

步骤1 登录MRS控制台，单击集群名称进入集群详情页。

步骤2 在MRS集群详情页，查看“虚拟私有云”。

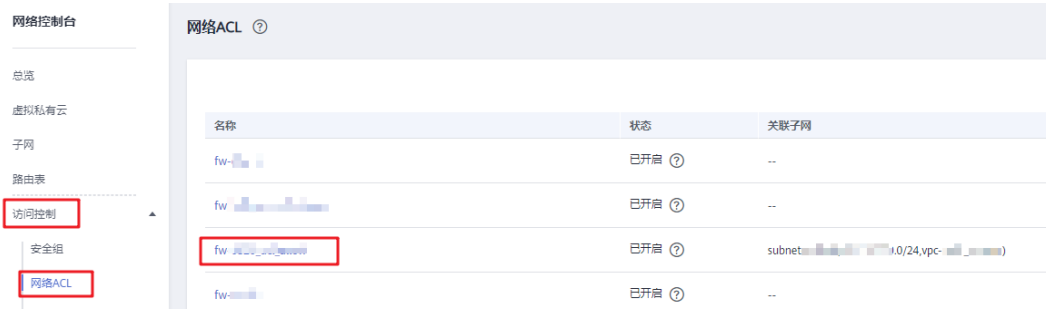
步骤3 进入VPC控制台，在左侧导航处选择“虚拟私有云”，获取**步骤2**中查询的“虚拟私有云”对应的“IPv4网段”。

步骤4 在VPC控制台左侧导航处选择“访问控制 > 网络ACL”，单击MRS集群默认生效子网和切换后子网关联的网络ACL名称，选择“入方向规则”页签。

说明

若集群的默认生效子网和切换后子网均关联网络ACL，则两个子网关联的网络ACL中均需参考**步骤5**-**步骤7**增加入方向规则。

图 5-3 网络 ACL



步骤5 在第一条规则的“操作”列，选择“更多 > 向前插规则”。

步骤6 添加网络ACL规则，其中“策略”选择“允许”，“源地址”填入**步骤3**中获取的“虚拟私有云”对应的“IPv4网段”，其他值保持默认即可。

步骤7 单击“确定”完成网络ACL规则添加。

说明

如果您不想放开VPC对应的全部IPv4网段的规则，请参考**步骤8**-**步骤12**添加切换前后两个子网对应的IPv4网段地址。若已完成VPC对应IPv4网段的规则添加，则无需执行**步骤8**-**步骤12**的操作。

步骤8 登录MRS控制台。

步骤9 单击集群名称进入集群详情页。

步骤10 在“默认生效子网”右侧单击“切换子网”。

步骤11 获取“默认生效子网”和待切换子网对应的IPv4网段。

须知

此时请勿单击切换子网的“确定”按钮，否则默认生效子网将更新为切换后的子网，切换前的子网不易查询，请谨慎操作。

步骤12 参考**步骤4**-**步骤7**添加“默认生效子网”和待切换子网的IPv4网段地址到切换前后子网绑定的网络ACL入方向规则中。

步骤13 登录MRS控制台。

步骤14 单击集群名称进入集群详情页。

步骤15 在“默认生效子网”右侧单击“切换子网”。

步骤16 选择待切换子网，并单击“确定”完成切换。

----结束

如何配置网络 ACL 出规则?

- 方案一：

  放通网络ACL所有出站流量，此方案能保证集群正常创建与使用，优先建议使用此方案。

- 方案二：

  放通保证集群创建成功的最小出规则，此方案可能在后续使用中因出方向规则遗漏导致集群使用问题，不建议使用方案。若出现集群使用问题请联系运维人员支撑处理。

  配置示例：参照方案一中示例，配置策略为“允许”，目的地址为通信安全授权地址、NTP、OBS、Openstack及DNS地址的出方向规则。

5.3.3 配置消息通知

MRS联合消息通知服务(SMN)，采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式（短信和邮件通知）。

操作场景

在MRS管理控制台，按照集群维度，在集群信息页面的告警页签中能够提供选择是否使能通知服务，只有对应集群开关开启以后，才能实现以下场景的功能：

- 在用户订阅了通知服务之后，当集群出现扩容成功/失败、缩容成功/失败、删除成功/失败、弹性升缩成功/失败的场景下，由MRS管理面通过邮件或短信方式通知对应用户。
- 管理面检查大数据集群的告警信息，如果大数据集群的告警信息影响到服务的使用，其告警级别达到致命时，则发送信息通知给对应租户。
- 在用户集群的ECS机器被删除、关机、修改规格、重启、更新OS的行为，会导致大数据集群异常，当检测到用户的虚拟机出现以上状态的时候，发送通知给对应用户。

创建主题

主题是消息发布或客户端订阅通知的特定事件类型。它作为发送消息和订阅通知的信道，为发布者和订阅者提供一个可以相互交流的通道。

1. 登录管理控制台。
2. 单击“服务列表”选择“管理与监管 > 消息通知服务”。

   进入消息通知服务页面。
3. 在左侧导航栏，选择“主题管理 > 主题”。

   进入主题页面。
4. 在主题页面，单击“创建主题”，开始创建主题。

   此时将显示“创建主题”对话框。

5. 在“主题名称”框中，输入主题名称，在“显示名”框中输入相关描述。
6. 在“企业项目”中选择已有的项目，或者单击“新建企业项目”，在“企业项目管理”界面创建好企业项目后再进行添加。
7. 在“标签”填写“标签键”和“标签值”，用于标识云资源，可对云资源进行分类和搜索。

向主题添加订阅

要接收发布至主题的消息，您必须添加一个订阅终端节点到该主题。消息通知服务会发送一条订阅确认的消息到订阅终端，订阅确认的消息将在48小时内有效。如果订阅者在48小时之内确认订阅，将会收到推送至主题的消息。如果订阅者在48小时之内没有确认订阅，则需要再次给订阅者发送订阅确认的消息。

1. 登录管理控制台。
2. 选择“管理与监管 > 消息通知服务”。

   进入消息通知服务页面。
3. 在左侧导航栏，选择“主题管理 > 主题”。

   进入主题页面。
4. 在主题列表中，选择您要向其添加订阅者的主题，在右侧“操作”栏单击“添加订阅”。

   此时将显示“添加订阅”对话框。

   其中：协议参数选项为“短信”、“邮件”、FunctionGraph（函数）HTTP、HTTPS。

   订阅终端参数为订阅的终端地址，短信、邮件终端支持批量输入，批量添加时，每个终端地址占一行。最多可输入10个终端。
5. 单击“确定”。

新增订阅将显示在页面下方的订阅列表中。

向订阅者发送消息

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。
3. 单击“告警管理”。
4. 选择“消息订阅规则 > 添加消息订阅规则”，进入添加消息订阅页面。
5. 配置消息订阅规则相关参数。

表 5-17 消息订阅规则参数说明

参数	说明
规则名称	用户自定义发送订阅消息的规则名称，只能包含数字、英文字符、中划线和下划线。

参数	说明
提醒通知	• 选择开启时，将按照该订阅规则为订阅者发送对应订阅消息。 • 选择关闭时，该规则不会生效，即不会向订阅者发送订阅消息。
主题名称	选择已创建的主题，也可以单击“创建主题”重新创建。
消息类型	选择需要订阅的消息类型。 • 告警
订阅规则	选择需要订阅的消息规则，可根据需要勾选全部或部分规则。 MRS 3.x及之后版本订阅规则： 告警：紧急，重要，次要 MRS 3.x之前版本订阅规则： • 致命 • 严重 • 一般 • 提示

6. 单击“确定”完成消息提醒配置。

5.3.4 健康检查

5.3.4.1 使用前须知

本章节指导用户在MRS控制台执行健康检查管理操作。

在MRS控制台执行健康检查管理操作仅适用于**MRS 1.9.2~MRS 2.1.x版本**集群。

在Manager界面执行健康检查管理操作适用于所有版本，MRS 3.x及之后版本请参考**查看健康检查任务**，MRS 3.x之前版本请参考**执行健康检查**。

5.3.4.2 执行健康检查

操作场景

该任务指导用户在日常运维中完成集群进行健康检查的工作，以保证集群各项参数、配置以及监控没有异常、能够长时间稳定运行。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

操作步骤

- 手动执行所有服务的健康检查

  在集群详情页，单击页面右上角“管理操作 > 启动集群健康检查”。

  说明

  MRS Manager具体请参见**执行健康检查**，MRS 3.x及之后版本FusionInsight Manager操作请参考**集群管理概述**。

  - 集群健康检查包含了Manager、服务与主机状态的检查。
  - 在MRS Manager界面，选择“系统设置 > 健康检查 > 集群健康检查”，也可以执行集群健康检查。
  - 手动执行健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。

- 手动执行单个服务的健康检查

  a. 在集群详情页，单击“组件管理”。

  b. 在服务列表中单击指定服务名称。

  c. 选择“更多 > 启动服务健康检查”启动指定服务健康检查。

- 手动执行主机健康检查

  a. 在集群详情页，单击“节点管理”。

  b. 展开节点组信息，勾选待检查主机前的复选框。

  c. 选择“节点操作 > 启动主机健康检查”启动指定主机健康检查。

5.3.4.3 查看并导出检查报告

操作场景

为了满足对健康检查结果的进一步具体分析，您可以在MRS中查看以及导出健康检查的结果。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

前提条件

已执行健康检查。

操作步骤

步骤1 在集群详情页，单击页面右上角“管理操作 > 查看集群健康检查报告”。

步骤2 在健康检查的报告面板上单击“导出报告”导出健康检查报告，可查看检查项的完整信息。

----结束

5.3.5 远程运维

5.3.5.1 运维授权

当用户使用集群过程中出现问题需要支持人员协助解决时，用户可先联系支持人员，再通过“运维授权”功能授权支持人员访问用户机器的权限用于定位问题。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在页面右上角单击“运维”，选择“运维授权”，选择授权给支持人员访问本机的权限的“截止时间”。在截止时间之前支持人员有临时访问本机的权限。

步骤4 问题解决后，在页面右上角单击“运维”，选择“取消授权”为支持人员取消访问权限。

----结束

5.3.5.2 日志共享

当用户使用集群过程中出现问题需要支持人员协助解决时，用户可先联系支持人员，再通过“日志共享”功能提供特定时间段内的日志给支持人员以便定位问题。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在页面右上角单击“运维”，选择“日志共享”，进入“日志共享”界面。

步骤4 在“起止时间”的输入框单击选择日期和时间。

说明

- “起止时间”按照支持人员的建议选取。
- 结束时间的选择的时间必须大于开始时间选择的时间，否则，无法根据时间筛选日志。

----结束

5.3.6 查看 MRS 服务操作日志

“操作日志”页面记录用户对集群和作业的操作信息。日志信息常用于集群运行异常时的问题定位，帮助用户快速定位问题原因，以便及时解决问题。

操作类型

目前MRS记录以下操作类型的日志信息，可在搜索框中筛选查询：

- 集群操作
  - 创建集群、删除集群、扩容集群和缩容集群等
  - 创建目录、删除目录和删除文件
- 作业操作：创建作业、停止作业和删除作业
- 数据操作：IAM用户任务、新增用户、新增用户组等操作

日志字段

日志列表默认按时间顺序排列，时间最近的日志显示在最前端。

日志信息中的各字段说明如表5-18所示。

表 5-18 日志说明

参数	参数说明
操作类型	记录执行的操作类型，包括： ● 集群操作 ● 作业操作 ● 数据操作
操作IP	记录执行操作的IP地址。 **说明** 当MRS集群部署失败时，集群会被自动删除，并且自动删除集群的操作日志中不包含用户的“操作IP”信息。
操作内容	记录实际操作内容，不超过2048字符。
时间	记录操作的时间。对于已删除的集群，界面只显示6个月内的日志信息，若需要查看6个月之前的日志信息，请联系支持人员。
企业项目	操作的集群所属的企业项目。

表 5-19 按钮说明

按钮	说明
	在下拉框中选择企业项目，筛选日志。

按钮	说明
全部	在下拉框中选择操作类型，筛选日志。 ● 全部：表示筛选所有的日志 ● 集群操作：表示筛选“集群操作”的日志 ● 作业操作：表示筛选“作业操作”的日志 ● 数据操作：表示筛选“数据操作”的日志
	根据时间筛选日志。 1. 单击输入框。 2. 选择日期和时间。 3. 单击“确认”。 左侧框为需要查询的开始时间，右侧框为需要查询的结束时间。右侧的输入框选择的时间必须大于或等于左侧输入框的时间，否则，无法根据时间筛选日志。
	在搜索框中输入“操作内容”中的关键字，单击，搜索日志。
	单击，手动刷新日志列表。

5.3.7 删除集群

如果作业执行结束后不需要集群， 可以删除MRS集群。

背景信息

一般在数据完成分析和存储后或集群异常无法提供服务时才执行集群删除操作。当MRS集群部署失败时，集群会被自动删除。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”。

步骤3 在需要删除的集群对应的“操作”列中，单击“删除”。

集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。集群删除后不再产生费用。

----结束

5.4 节点管理

5.4.1 扩容集群

MRS的扩容不论在存储还是计算能力上，都可以简单地通过增加Core节点或者Task节点来完成，不需要修改系统架构，降低运维成本。集群Core节点不仅可以处理数据，

也可以存储数据。可以在集群中添加Core节点，通过增加节点数量处理峰值负载。集群Task节点主要用于处理数据，不存放持久数据。

背景信息

MRS集群支持Core与Task节点总数最大为500个。如果用户需要的Core/Task节点数大于500，可以联系支持人员或者调用后台接口修改数据库。

目前支持扩容Core节点和Task节点，不支持扩容Master节点。此处扩容的最大Core/Task节点数为（500 - 集群Core/Task节点数）。例如：当前集群Core节点数为3，此处扩容的Core节点数必须小于等于497。如果集群扩容失败，用户可重新进行扩容操作。

如果在创建集群时，没有扩容节点，用户可以在扩容时添加节点个数，但不能指定具体节点扩容。

选择的版本不同，扩容集群的操作也不同。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择 “集群列表 > 现有集群” ，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 选择“节点管理”页签，在需要扩容的节点组的“操作”列单击“扩容”，进入扩容集群页面。

只有运行中的集群才能进行扩容操作。

步骤4 设置“扩容节点数量”、“启动组件”和“执行引导操作”参数，并单击“确定”。

说明

- 若集群中没有Task节点组，请参考**添加Task节点**配置Task节点。
- 如果创建集群时添加了引导操作，则“执行引导操作”参数有效，开启该功能时扩容的节点会把创建集群时添加的引导操作脚本都执行一遍。
- 如果“新节点规格”参数有效，则表示与原有节点相同的规格已售罄或已下架，新扩容的节点将按照“新节点规格”增加。
- 扩容集群前需要检查集群安全组是否配置正确，要确保集群入方向安全组规则中有一条全部协议，全部端口，源地址为可信任的IP访问范围的规则。

步骤5 进入“扩容节点”窗口，单击“确定”。

步骤6 弹出扩容节点提交成功提示框。

集群扩容过程说明如下：

- 扩容中：集群正在扩容时集群状态为“扩容中”。已提交的作业会继续执行，也可以提交新的作业，但不允许继续扩容和删除集群，也不建议重启集群和修改集群配置。
- 扩容成功：集群扩容成功后集群状态为“运行中”。
- 扩容失败：集群扩容失败时集群状态为“运行中”。用户可以执行作业，也可以重新进行扩容操作。

扩容成功后，可以在集群详情的“节点管理”页签查看集群的节点信息。

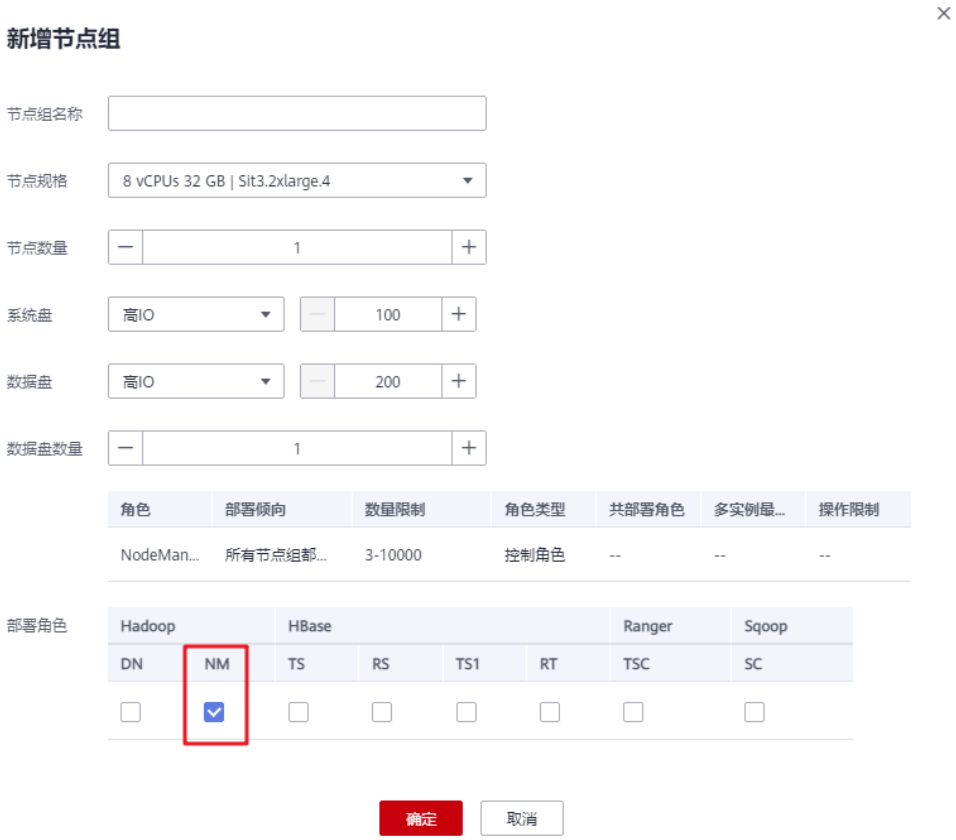
----结束

添加 Task 节点

“自定义”类型集群添加Task节点操作步骤：

1. 在集群详情页面，选择“节点管理”页签，单击“新增节点组”，进入“新增节点组”页面。
2. “部署角色”参数仅选择“NM”部署NodeManager角色，则新增节点组为Task节点组，其他参数根据需要配置。

图 5-4 添加 Task 节点组



非“自定义”类型集群添加Task节点操作步骤：

1. 在集群详情页面，选择“节点管理”页签，单击“配置Task节点”，进入“配置Task节点”页面。
2. 配置“节点类型”、“节点规格”、“节点数量”、“系统盘”，如开启“添加数据盘”后，还需要配置数据盘的存储类型、大小和数量。
3. 单击“确定”。

添加节点组

说明

用于增加节点组，适用MRS 3.x版本的自定义集群。

1. 在集群详情页面，选择“节点管理”页签，单击“新增节点组”，进入“新增节点组”页面。

2. 根据需求配置参数。

表 5-20 新增节点组参数说明

参数名称	描述
节点规格	选择节点组内主机的规格类型。
节点数量	设置新增节点组内的节点数量。
系统盘	设置新增节点的系统盘的规格与容量。
数据盘/数据盘数量	设置新增节点的数据盘的规格与容量及数量。
部署角色	新增节点组内，各节点的实例部署发布，可手动调节。

3. 单击“确定”。

5.4.2 缩容集群

用户可以根据业务需求量，通过简单的缩减Core节点或者Task节点，对集群进行缩容，以使MRS拥有更优的存储、计算能力，降低运维成本。

当集群正在进行主备同步操作时，不允许进行缩容操作。

背景信息

目前支持缩容Core节点和Task节点，不支持缩容Master节点。对集群进行缩容时，只需要在界面调整节点个数，MRS会自动选择缩容节点，完成缩容任务。

自动选择缩容节点的策略如下：

- 不允许缩容安装了基础组件（Zookeeper，DBService，KrbServer，LdapServer等）的节点，MRS不会选择这些节点进行缩容。因为这些基础组件是集群运行的基础。
- Core节点是存放集群业务数据的节点，在缩容时必须保证待缩容节点上的数据被完整迁移到其他节点，即完成各个组件的退服之后，才会执行缩容的后续操作（节点退出Manager和删除ECS等）。在选择Core节点时，会优先选择存储数据量较小，且可退服实例健康状态良好的节点，避免节点退服失败。例如在分析集群上，Core节点安装了DataNode，缩容时会优先选择DataNode存储数据量较小且健康状态良好的节点。

  Core节点在缩容的时候，会对原节点上的数据进行迁移。业务上如果对数据位置做了缓存，客户端自动刷新位置信息可能会影响时延。缩容节点可能会影响部分HBase on HDFS数据的第一次访问响应时长，可以重启HBase或者对相关的表Disable/Enable来避免。
- Task节点本身不存储集群数据，属于计算节点，不存在节点数据迁移的问题。因此在选择Task节点时，优先选择健康状态为故障、未知、亚健康的节点进行缩容。这些节点实例的健康状态信息可以在MRS上的“实例”管理界面查看。

缩容校验策略

缩容节点选择完成后，为了避免组件退服失败，不同组件提供了不同的退服约束规则，只有满足了所有安装组件的退服约束规则才允许缩容。缩容校验策略如表5-21所示。

表 5-21 组件退服约束规则

组件名称	退服约束规则
HDFS/ DataNode	规则：缩容后节点数不小于当前HDFS的副本数且HDFS数据总量不超过缩容后HDFS集群总容量的80%，可以执行缩容操作。 原因：确保缩容后剩余空间足够存放现有数据，并预留一部分空间。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。
HBase/ RegionServer	规则：除缩容节点外，其他节点RegionServer剩余可用内存的总和，大于所选缩容节点RegionServer当前使用内存的1.2倍。 原因：当一个节点退服时，这个节点上的Region会迁移到其他节点，所以其他节点的可用内存必须足够才能负担起退服节点的Region。
Storm/ Supervisor	规则：缩容后集群slot数足够运行当前已提交的任务。 原因：防止缩容后没有充足的资源运行流处理任务。
Flume/ FlumeServer	规则：节点安装了FlumeServer，并且已经配置了Flume任务，则该节点不能删除。 原因：防止误删了已部署的业务程序。

指定数量缩容

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 选择“节点管理”页签，在需要缩容的节点组的“操作”列，单击“缩容”，进入缩容集群页面。

只有运行中的集群且集群中的节点都在运行中才能进行该操作。

步骤4 设置“缩容节点数量”，并单击“确定”。

说明

- 缩容集群前需要检查集群安全组是否配置正确，要确保集群入方向安全组规则中有一条全部协议，全部端口，源地址为可信任的IP访问范围的规则。
- 若HDFS存在损坏的数据块，则缩容集群可能出现失败，请联系支持人员处理。

步骤5 页面右上角弹出缩容节点提交成功提示框。

集群缩容过程说明如下：

- 缩容中：集群正在缩容时集群状态为“缩容中”。已提交的作业会继续执行，也可以提交新的作业，但不允许继续缩容和删除集群，也不建议重启集群和修改集群配置。
- 缩容成功：集群缩容成功后集群状态为“运行中”。
- 缩容失败：集群缩容失败时集群状态为“运行中”。用户可以执行作业，也可以重新进行缩容操作。

缩容成功后，可以在集群详情的“节点管理”页签查看集群的节点信息。

----结束

5.4.3 管理主机（节点）操作

操作场景

当主机（节点）故障异常时，用户可能需要在MRS停止主机（节点）上的所有角色，对主机（节点）进行维护检查。故障清除后，启动主机（节点）上的所有角色恢复主机（节点）业务。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“节点管理”。

步骤2 展开节点组信息，勾选待操作节点前的复选框。

步骤3 选择“节点操作 > 启动所有角色”或“停止所有角色”执行相应操作。

----结束

5.4.4 隔离主机

操作场景

用户发现某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。在为集群安装补丁的场景中，也支持排除指定节点不安装补丁。

该任务指导用户在MRS上根据实际业务或运维规划手工将主机隔离。隔离主机仅支持隔离非管理节点。

对系统的影响

- 主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。
- 主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“节点管理”。

步骤2 展开节点组信息，勾选待隔离主机前的复选框。

步骤3 选择“节点操作 > 隔离主机”。

步骤4 确认待隔离主机信息并单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功隔离，“操作状态”显示为“已隔离”

说明

已隔离的主机，可以取消隔离重新加入集群，请参见**取消隔离主机**。

----结束

5.4.5 取消隔离主机

操作场景

用户已排除主机的异常或故障后，需要将主机隔离状态取消才能正常使用。

该任务指导用户在MRS上取消隔离主机。

前提条件

- 主机状态为“已隔离”。
- 主机的异常或故障已确认修复。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“节点管理”。

步骤2 展开节点组信息，勾选待取消隔离主机前的复选框。

步骤3 选择“节点操作 > 取消隔离主机”。

步骤4 确认待取消隔离主机信息并单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功取消隔离，“操作状态”显示为“正常”。

步骤5 勾选已取消隔离的主机，选择“节点操作 > 启动所有角色”。

----结束

5.4.6 升级 Master 节点规格

随着用户业务的增长，Core节点的扩容，CPU使用率变高，而Master节点规格已经不满足用户需求时，则需要升级Master节点规格。本章节介绍Master节点规格升级的操作流程。

前提条件

确认是否开启了企业主机安全（Host Security Service，简称HSS）服务，如果已开启，升级Master节点规格前需要先暂时关闭HSS服务对MRS集群的监测。

使用限制

- 支持2个及以上Master节点的集群升级Master节点规格 。
- 不支持使用BMS类型规格的集群升级Master节点规格 。

集群 Master 节点规格升级（一键升级）

步骤1 登录MRS管理控制台。

步骤2 选择 “集群列表 > 现有集群” ，选中需要升级Master节点规格的集群并单击集群名，进入集群信息页面。

步骤3 在“节点管理”页签Master节点组的“操作”列选择“升级规格”，进入“升级Master规格”页面。

步骤4 选择升级后的规格，单击“提交”成功提交升级Master规格任务。

节点规格升级过程需要时间，升级成功后集群状态更新为“运行中”，请您耐心等待。

说明

- 升级过程中集群会自动关闭升级的虚拟机，升级完成后自动开启该虚拟机 。
- 因用户对组件使用需求不同，节点规格升级成功后不会自动更新组件内存配置，用户可根据实际使用情况自行调整各组件内存配置。

----结束

5.5 作业管理

5.5.1 MRS 作业简介

MRS作业是MRS为用户提供的程序执行平台，用于处理和分析用户数据。作业创建完成后，所有的作业列表信息展示在“作业管理”页面中，您可以查看所有的作业列表，也可以创建和管理作业。若集群详情页面不支持“作业管理”页签，请通过后台方式提交作业。

MRS集群处理的数据源来源于OBS或HDFS，HDFS是Hadoop分布式文件系统（Hadoop Distributed File System），OBS即对象存储服务，是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用，同时可以通过REST API接口方式单独或集成到业务程序进行管理和访问数据。

用户创建作业前需要将本地数据上传至OBS系统，MRS使用OBS中的数据进行计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

作业分类

目前MRS集群支持创建和管理如下几种类型的作业。如果处于“运行中”状态的集群创建作业失败，请查看集群管理页面中相关组件健康情况。操作方法，请参见**查看和定制集群监控指标**。

- MapReduce：提供快速并行处理大量数据的能力，是一种分布式数据处理模式和执行环境。MRS当前支持提交MapReduce Jar程序。
- Spark：基于内存进行计算的分布式计算框架，MRS当前支持提交SparkSubmit、Spark Script和Spark SQL作业。
  - SparkSubmit：支持提交Spark Jar和Spark python程序，执行Spark application，计算和处理用户数据。
  - SparkScript：支持提交SparkScript脚本，批量执行Spark SQL语句。
  - Spark SQL：运用Spark提供的类似SQL的Spark SQL语言，实时查询和分析用户数据。
- Hive：建立在Hadoop基础上的开源的数据仓库。MRS当前支持提交HiveScript脚本，和执行Hive SQL语句。
- Flink：提供一个分布式大数据处理引擎，可对有限数据流和无限数据流进行有状态计算。

作业列表

作业列表默认按时间顺序排列，时间最近的作业显示在最前端。各类作业列表参数说明如**表 1**所示。

表 5-22 作业列表参数

参数	参数说明
作业名称/ID	作业的名称，新增作业时配置。 ID是作业的唯一标识，作业新增后系统自动赋值。
用户名称	提交作业的用户名称。

参数	参数说明
作业类型	支持的作业类型： ● Distcp：导入、导出数据 ● MapReduce ● Spark ● SparkSubmit ● SparkScript ● Spark SQL ● Hive SQL ● HiveScript ● Flink **说明** ● 在“文件管理”页面进行文件的导入导出操作后，您可以在“作业管理”页面查看Distcp作业。 ● 只有创建集群时选择了Spark、Hive和Flink组件，并且集群处于运行中，才能新增Spark、Hive和Flink类型的作业。
状态	显示作业的状态。 ● 已提交 ● 已接受 ● 运行中 ● 已完成 ● 已终止 ● 异常
执行结果	显示作业执行完成的结果。 ● 未定：正在执行的作业。 ● 成功：执行成功的作业。 ● 终止：执行中被手动终止的作业。 ● 失败：执行失败的作业。 **说明** 作业执行成功或失败后都不能再次执行，只能新增作业，配置作业参数后重新提交作业。
作业提交时间	记录作业提交的开始时间。
作业结束时间	记录作业执行完成或手工停止的时间。

参数	参数说明
操作	• 查看日志：单击“查看日志”，查看运行中的作业执行的实时日志信息。操作方法，请参见**查看作业配置信息和日志**。 • 查看详情：单击“查看详情”，查看作业的详细配置信息。操作方法，请参见**查看作业配置信息和日志**。 • 更多 – 停止：单击“停止”，停止正在运行的作业。操作方法，请参见**停止作业**。 – 删除：单击“删除”，删除一个作业。操作方法，请参见**删除作业**。 – 结果：单击“结果”，查看SparkSql和SparkScript类型的“状态”为“已完成”且“执行结果”为“成功”的作业执行结果。 **说明** • Spark SQL作业不支持停止。 • 作业删除后不可恢复，请谨慎操作。 • 当选择保留作业日志到OBS或HDFS时，系统在作业执行结束后，将日志压缩并存储到对应路径。因此，此类作业运行结束后，作业状态仍然为“运行中”，需等日志存储成功后，状态变更为“已完成”。日志存储花费时间依赖于日志大小，需要数分钟以上。

表 5-23 按钮说明

按钮	说明
	选择提交作业的时间区间，筛选在对应时间区间内提交的作业。
	在下拉框中选择作业执行结果，筛选作业。 • 全部：表示筛选所有的作业。 • 成功：表示筛选执行成功的作业。 • 未定：表示筛选正在执行的作业。 • 终止：表示筛选被手动终止的作业。 • 失败：表示筛选执行失败的作业。

按钮	说明
全部作业类型	在下拉框中选择作业类型，筛选作业。 • 全部作业类型 • MapReduce • HiveScript • Distcp • SparkScript • Spark SQL • Hive SQL • SparkSubmit • Flink
作业名称	在搜索框中根据搜索条件输入对应内容，单击，搜索作业。 • 作业名称 • 作业ID • 用户名称 • 队列名称
	单击，手动刷新作业列表。

作业执行权限说明

对于开启Kerberos认证的安全集群，用户在MRS界面提交作业时，要先执行IAM用户同步操作，同步完成后会在MRS系统中产生同IAM用户名的用户。IAM同步用户是否有提交作业权限，取决于IAM同步时，用户所绑定的IAM策略，提交作业策略请参考**IAM用户同步MRS说明**章节中**表3-3**。

用户提交作业，如果涉及到具体组件的资源使用，如HDFS的目录访问、Hive表的访问等相关组件的权限时，需由admin（Manager管理员）用户进行授权，给提交作业用户赋予相关组件权限。具体操作如下：

步骤1 使用admin用户登录Manager。

步骤2 参考**创建角色**内容，增加用户具体需要的组件权限的角色。

步骤3 参考**相关任务**修改提交作业用户所属的用户组，将新增的组件角色加入到该用户组中。

说明

用户所在用户组绑定的组件角色修改后，权限生效需要一定时间，请耐心等待。

----结束

5.5.2 运行 MapReduce 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节指导您在MRS集群页面如何提交一个新的MapReduce作业。MapReduce作业用于提交jar程序快速并行处理大量数据，是一种分布式数据处理模式和执行环境。

若在集群详情页面不支持“作业管理”和“文件管理”功能，请通过后台功能来提交作业。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 “作业类型”选择“MapReduce”，并配置其他作业信息。

表 5-24 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。

参数	参数说明
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 • SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。
执行程序参数	可选参数，程序执行的关键参数。多个参数间使用空格隔开。 配置方法：*程序类名 数据输入路径 数据输出路径* • 程序类名：由用户程序内的函数指定，MRS只负责参数的传入。 • 数据输入路径：通过单击“HDFS”或者“OBS”选择或者直接手动输入正确路径。 • 数据输出路径：输出路径请手动输入一个不存在的目录。最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-25**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-25 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在“节点管理”页签中单击某一Master节点名称，进入弹性云服务器管理控制台。

步骤4 单击页面右上角的“远程登录”。

步骤5 根据界面提示，输入Master节点的用户名和密码，用户名、密码分别为root和创建集群时设置的密码。

步骤6 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤7 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证，则无需执行该步骤。

kinit *MRS集群用户*

例如, **kinit admin**

步骤8 执行如下命令拷贝OBS文件系统中的程序到集群的Master节点。

hadoop fs -Dfs.obs.access.key=AK -Dfs.obs.secret.key=SK -copyToLocal source_path.jar target_path.jar

例如：**hadoop fs -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX -copyToLocal "obs://mrs-word/program/hadoop-mapreduce-examples-XXX.jar" "/home/omm/hadoop-mapreduce-examples-XXX.jar"**

AK/SK可登录OBS控制台，请在集群控制台页面右上角的用户名下拉框中选择“我的凭证 > 访问密钥”页面获取。

步骤9 执行如下命令提交wordcount作业，如需从OBS读取或向OBS输出数据，需要增加AK/SK参数。

source /opt/Bigdata/client/bigdata_env;hadoop jar execute_jar wordcount input_path output_path

例如：**source /opt/Bigdata/client/bigdata_env;hadoop jar /home/omm/hadoop-mapreduce-examples-XXX.jar wordcount -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX "obs://mrs-word/input/*" "obs://mrs-word/output/"**

input_path为OBS上存放作业输入文件的路径。output_path为OBS上存放作业输出文件地址，请设置为一个不存在的目录。

----结束

5.5.3 运行 SparkSubmit 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在MRS集群页面如何提交一个新的Spark作业。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 配置作业信息。

表 5-26 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。

参数	参数说明
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 • SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行程序参数如**表5-27**。
执行程序参数	可选参数，程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-28**。 **说明** 如需运行长时作业如SparkStreaming等，且需要访问OBS, 需要通过“服务配置参数”传入访问OBS的AK/SK。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-27 运行程序参数

参数	参数说明	取值样例
--conf	添加任务配置项。	spark.executor.memory=2G
--driver-memory	设置driver的运行内存。	2G
--num-executors	设置executor启动数量。	5
--executor-cores	设置executor核数。	2

参数	参数说明	取值样例
--class	设置任务的主类。	org.apache.spark.examples.SparkPi
--files	上传文件给任务，可以是自己定义的配置文件或者某些数据文件。来源可以是OBS或者HDFS。	-
--jars	上传任务额外依赖包，用于给任务添加任务的外部依赖包。	-
--executor-memory	设置executor内存。	2G
--conf spark-yarn.maxAppAttempts	控制AM的重试次数。	设置为0时，不允许重试；设置为1时，允许重试一次。

表 5-28 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 参考**创建用户**页面，创建一个用于提交作业的用户。

本示例创建一个用户开发场景使用的机机用户，并分配了正确的用户组（hadoop、supergroup）、主组（supergroup）和角色权限（System_administrator、default）。

步骤2 下载认证凭据。

- 对于MRS 3.x及之后版本集群，请登录FusionInsight Manager页面选择“系统 > 权限 > 用户”，在新增用户的操作列单击“更多 > 下载认证凭据”。
- 对于MRS 3.x之前版本集群，请登录MRS Manager页面选择“系统设置 > 用户管理”，在新增用户的操作列单击“更多 > 下载认证凭据”。

步骤3 将与作业相关的jar包上传到集群中，本示例使用Spark自带的样例jar包，位置在$SPARK_HOME/examples/jars/下。

步骤4 上传步骤2创建的用户认证凭据到集群的/opt/目录下，并执行如下命令解压

tar –xvf MRSTest _xxxxxx_keytab.tar

您将会得到user.keytab和krb5.conf两个文件。

步骤5 在对集群操作之前首先需要执行：

source /opt/Bigdata/client/bigdata_env

cd $SPARK_HOME

步骤6 提交spark作业，使用的命令如下：

./bin/spark-submit --master yarn --deploy-mode client --conf spark.yarn.principal=MRSTest --conf spark.yarn.keytab=/opt/user.keytab --class org.apache.spark.examples.SparkPi examples/jars/spark-examples_2.11-2.3.2-mrs-2.0.jar 10

参数解释：

1. yarn的计算能力，指定使用client模式提交该作业。
2. Spark作业的配置项，这里是传入了认证文件和用户名。
3. spark.yarn.principal 第一步创建的用户
4. spark.yarn.keytab 认证使用的keytab文件
5. xx.jar 作业的使用的jar。

----结束

5.5.4 运行 HiveSql 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在MRS集群页面如何提交一个新的HiveSql作业。HiveSql作业用于提交SQL语句和SQL脚本文件查询和分析数据，包括SQL语句和Script脚本两种形式，如果SQL语句涉及敏感信息，请使用Script提交。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 配置作业信息。

表 5-29 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
SQL类型	SQL查询语句提交类型。 ● SQL ● Script
SQL语句	“SQL类型”参数为“SQL”时参数有效，请输入待运行的SQL语句，然后单击“检查”来检查SQL语句的正确性，确保输入语句正确。如果同时需要提交多条语句并执行，使用“;”分隔不同语句。
SQL文件	“SQL类型”参数为“Script”时参数有效，待执行SQL文件的路径，需要满足以下要求。 ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。 **说明** 存储在OBS上的文件路径支持以“obs://”开头格式。如需使用该格式提交作业，访问OBS需要配置对应权限。 ● 创建集群时开启“OBS权限控制”功能时，可直接使用“obs://”路径，无需单独配置。 ● 创建集群时未开启或不支持“OBS权限控制”功能时，请参考**访问OBS**页面进行配置。

参数	参数说明
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行参数如**表5-30**。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-31**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-30 运行程序参数

参数	参数说明	取值样例
--hiveconf	设置Hive服务相关配置，例如设置执行引擎为MR。	设置执行引擎为MR：--hiveconf "hive.execution.engine=mr"
--hivevar	设置用户自定义变量，例如设置变量id。	设置变量id：--hivevar id="123" select * from test where id = ${hivevar:id};

表 5-31 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-
hive.execution.engine	选择执行作业的引擎。	• mr • tez

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在“节点管理”页签中单击某一Master节点名称，进入弹性云服务器管理控制台。

步骤4 单击页面右上角的“远程登录”。

步骤5 根据界面提示，输入Master节点的用户名和密码，用户名、密码分别为root和创建集群时设置的密码。

步骤6 执行如下命令初始化环境变量。

source /opt/BigData/client/bigdata_env

说明

- MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。
- 若安装了Hive多实例，在使用客户端连接具体Hive实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载Hive2实例变量：

  source /opt/BigData/client/Hive2/component_env

步骤7 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证(普通模式），则无需执行该步骤。

kinit *MRS集群用户* (用户需要有hive组）

步骤8 执行beeline连接hiveserver，运行任务。

beeline

普通模式，则执行以下命令，如果不指定组件业务用户，则会以当前操作系统用户连接hiveserver。

beeline -n*组件业务用户*

beeline -f sql文件（执行文件里的sql）

----结束

5.5.5 运行 SparkSql 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在MRS集群页面如何提交一个新的SparkSql作业。SparkSQL作业用于查询和分析数据，包括SQL语句和Script脚本两种形式，如果SQL语句涉及敏感信息，请使用Spark Script提交。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 请单击“添加”，进入添加作业页面，“作业类型”选择“SparkSql”，作业参考**表5-32**配置SparkSql作业信息。

表 5-32 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
SQL类型	SQL查询语句提交类型。 ● SQL ● Script
SQL语句	“SQL类型”参数为“SQL”时参数有效，请输入待运行的SQL语句，然后单击“检查”来检查SQL语句的正确性，确保输入语句正确。如果同时需要提交多条语句并执行，使用“;”分隔不同语句。

参数	参数说明
SQL文件	“SQL类型”参数为“Script”时参数有效，待执行SQL文件的路径，需要满足以下要求。 ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。 **说明** 存储在OBS上的文件路径支持以“obs://”开头格式。如需使用该格式提交作业，访问OBS需要配置对应权限。 ● 创建集群时开启“OBS权限控制”功能时，可直接使用“obs://”路径，无需单独配置。 ● 创建集群时未开启或不支持“OBS权限控制”功能时，请参考**访问OBS**页面进行配置。
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行程序参数如**表5-33**。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-34**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-33 运行程序参数

参数	参数说明	取值样例
--conf	添加任务配置项	spark.executor.memory=2G
--driver-memory	设置driver的运行内存	2G
--num-executors	设置executor启动数量	5
--executor-cores	设置executor核数	2

参数	参数说明	取值样例
--jars	上传任务额外依赖包，用于给任务添加任务的外部依赖包	-
--executor-memory	设置executor内存	2G

表 5-34 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤6 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 参考**创建用户**页面，创建一个用于提交作业的用户。

本示例创建一个用户开发场景使用的机机用户，并分配了正确的用户组（hadoop、supergroup）、主组（supergroup）和角色权限（System_administrator、default）。

步骤2 下载认证凭据。

- 对于MRS 3.x及之后版本集群，请登录FusionInsight Manager页面选择“系统 > 权限 > 用户”，在新增用户的操作列单击“更多 > 下载认证凭据”。
- 对于MRS 3.x之前版本集群，请登录MRS Manager页面选择“系统设置 > 用户管理”，在新增用户的操作列单击“更多 > 下载认证凭据”。

步骤3 登录Spark客户端所在节点，上传**2**创建的用户认证凭据到集群的“/opt/”目录下，并执行如下命令解压：

tar –xvf MRSTest _xxxxxx_keytab.tar

得到user.keytab和krb5.conf两个文件。

步骤4 在对集群操作之前首先需要执行：

source /opt/Bigdata/client/bigdata_env

cd $SPARK_HOME

步骤5 打开spark-sql命令行，进入spark-sql命令行后可执行SQL语句，执行命令如下：

./bin/spark-sql --conf spark.yarn.principal=MRSTest --conf spark.yarn.keytab=/opt/user.keytab

若需要执行SQL文件，需要上传SQL文件（如上传到“/opt/”目录），上传文件后执行命令如下：

./bin/spark-sql --conf spark.yarn.principal=MRSTest --conf spark.yarn.keytab=/opt/user.keytab -f /opt/script.sql

----结束

5.5.6 运行 Flink 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节指导用户在MRS集群页面如何提交一个新的Flink作业。Flink作业用于提交jar程序处理流式数据。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 “作业类型”选择“Flink”，参考**表 1**配置Flink作业信息。

表 5-35 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行程序参数如**表5-36**。
执行程序参数	可选参数，程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-37**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-36 运行程序参数

参数	参数说明	取值样例
-ytm	设置每个TaskManager容器的内存（单位可选, 默认单位：MB）。	1024
-yjm	设置JobManager容器内存（单位可选，默认单位：MB）。	1024
-yn	设置分配给应用程序的Yarn容器的数量，该值与TaskManager数量相同。	2

参数	参数说明	取值样例
-ys	设置TaskManager的核数。	2
-ynm	自定义Yarn上应用程序名称。	test
-c	设置程序入口点的类（如“main”或“getPlan()”方法）。该参数仅在JAR文件未指定其清单的类时需要。	com.bigdata.mrs.test

说明

针对MRS 3.x及之后版本，运行程序参数不支持“-yn”。

表 5-37 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录MRS客户端。

步骤2 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤3 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 使用新创建的用户登录Manager页面。
   - MRS 3.x之前版本，登录集群的Manager界面，选择“系统设置 > 用户管理”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
   - MRS 3.x及之后版本，登录集群的Manager界面，选择“系统 > 权限 > 用户”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/Bigdata/client/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。

4. MRS 3.x及之后版本，安全模式下需要将客户端安装节点的业务IP以及Manager的浮动ip追加到“/opt/Bigdata/client/Flink/flink/conf/flink-conf.yaml”文件中的“jobmanager.web.allow-access-address”配置项中，ip之间使用英文逗号分隔。
5. 配置安全认证，在“/opt/Bigdata/client/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/Bigdata/client/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test
6. 在Flink的客户端bin目录下，执行如下命令进行安全加固，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/Bigdata/client/Flink/flink/conf/flink-conf.yaml”中关于SSL的值，针对MRS 3.x之前版本，安全集群默认没有开启外部SSL，用户如果需要启用外部SSL，进行配置后再次运行该脚本即可，配置参数在MRS的Flink默认配置中不存在，用户如果开启外部连接SSL，则需要添加**表5-38**中参数。

表 5-38 参数描述

参数	参数值示例	描述
security.ssl.rest.enabled	true	打开外部SSL开关。
security.ssl.rest.keystore	${path}/flink.keystore	keystore的存放路径。
security.ssl.rest.keystore-password	123456	keystore的password，“123456”表示需要用户输入自定义设置的密码值。
security.ssl.rest.key-password	123456	ssl key的password，“123456”表示需要用户输入自定义设置的密码值。
security.ssl.rest.truststore	${path}/flink.truststore	truststore存放路径。
security.ssl.rest.truststore-password	123456	truststore的password，“123456”表示需要用户输入自定义设置的密码值。

说明

- 针对MRS 3.x之前版本，generate_keystore.sh脚本无需手动生成。
- **认证和加密**会将生成的flink.keystore、flink.truststore、security.cookie自动填充到“flink-conf.yaml”对应配置项中。
- 针对MRS 3.x及之后版本，“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值需要使用Manager明文加密API进行获取：

  curl -k -i -u <*user name*>:<*password*> -X POST -HContent-type:application/json -d '{"plainText":"<*password*>"}' 'https://*x.x.x.x*:28443/web/api/v2/tools/encrypt'；其中<*password*>要与签发证书时使用的密码一致，*x.x.x.x*为集群Manager的浮动IP。

7. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/Bigdata/client/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/Bigdata/client/Flink/flink/conf/”目录下新建目录，例如ssl。
     ii. 移动flink.keystore和flink.truststore文件到“/opt/Bigdata/client/Flink/flink/conf/ssl/”中。
     iii. 针对MRS 3.x及之后版本，修改flink-conf.yaml文件中如下两个参数为相对路径。
     ```
     security.ssl.keystore: ssl/flink.keystore
     security.ssl.truststore: ssl/flink.truststore
     ```
     iv. 针对MRS 3.x之前版本，修改flink-conf.yaml文件中如下两个参数为相对路径。
     ```
     security.ssl.internal.keystore: ssl/flink.keystore
     security.ssl.internal.truststore: ssl/flink.truststore
     ```
8. 如果客户端安装在集群外节点，请在配置文件（如：**“/opt/Bigdata/client/Flink/fink/conf/flink-conf.yaml”**）中增加如下配置值，其中**xx.xx.xxx.xxx**请替换为客户端所在节点的IP。
   ```
   web.access-control-allow-origin: xx.xx.xxx.xxx
   jobmanager.web.allow-access-address: xx.xx.xxx.xxx
   ```

步骤4 运行wordcount作业。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。
    ```
    yarn-session.sh -nm "session-name"
    flink run /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```
  - 执行如下命令在Yarn上提交单个作业。
    ```
    flink run -m yarn-cluster /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```
- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。
      ```
      yarn-session.sh -nm "session-name"
      flink run /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
      ```
    - 执行如下命令在Yarn上提交单个作业。
      ```
      flink run -m yarn-cluster /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
      ```

- flink.keystore和flink.truststore文件路径为相对路径时：
  - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/Bigdata/client/Flink/flink/conf/”，则在“opt/Bigdata/client/Flink/flink/conf/”目录下执行命令。
    ```
    yarn-session.sh -t ssl/ -nm "session-name"
    flink run /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```
  - 执行如下命令在Yarn上提交单个作业。
    ```
    flink run -m yarn-cluster -yt ssl/ /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```

----结束

5.5.7 运行 Kafka 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在Kafka主题中产生和消费消息。

暂不支持通过界面提交Kafka作业，请通过后台功能来提交作业。

通过后台提交作业

先查询ZooKeeper和Kafka的实例地址，再运行Kafka作业。

查询实例地址（3.x版本）

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 请参考**访问FusionInsight Manager（MRS 3.x及之后版本）**，跳转至FusionInsight Manager页面。然后选择“服务 > ZooKeeper > 实例”，查看ZooKeeper角色实例的IP地址。记录ZooKeeper角色实例中任意一个的IP地址即可。

步骤4 选择“服务 > Kafka > 实例”，查看Kafka角色实例的IP地址。记录Kafka角色实例中任意一个的IP地址即可。

----结束

查询实例地址（3.x之前版本）

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在MRS集群详情页面，选择“组件管理 > ZooKeeper > 实例”，查看ZooKeeper角色实例的IP地址。记录ZooKeeper角色实例中任意一个的IP地址即可。

步骤4 选择“组件管理 > Kafka > 实例”，查看Kafka角色实例的IP地址。记录Kafka角色实例中任意一个的IP地址即可。

----结束

运行Kafka作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 在集群信息页面的“节点管理”页签中单击Master2节点名称，进入弹性云服务器管理控制台。

步骤2 单击页面右上角的“远程登录”。

步骤3 根据界面提示，输入Master节点的用户名和密码，用户名、密码分别为root和创建集群时设置的密码。

步骤4 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤5 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证，则无需执行该步骤。

kinit *MRS集群用户*

例如, **kinit admin**

步骤6 执行如下命令，创建kafka topic。

kafka-topics.sh --create --zookeeper <ZooKeeper角色实例IP:2181/kafka> --partitions 2 --replication-factor 2 --topic <Topic名称>

步骤7 在topic test中产生消息。

首先执行命令**kafka-console-producer.sh --broker-list <Kafka角色实例IP:9092> --topic <Topic名称> --producer.config /opt/Bigdata/client/Kafka/kafka/config/producer.properties**。

然后输入指定的内容作为生产者产生的消息，输入完成后按回车发送消息。如果需要结束产生消息，使用“Ctrl + C”退出任务。

步骤8 消费topic test中的消息。

kafka-console-consumer.sh --topic <Topic名称> --bootstrap-server <Kafka角色实例IP:9092> --consumer.config /opt/Bigdata/client/Kafka/kafka/config/consumer.properties

说明

如果集群开启Kerberos认证，则执行如上两个命令时请修改端口号9092为21007，详见**开源组件端口列表**。

----结束

5.5.8 查看作业配置信息和日志

本章节介绍如何查看作业的配置信息和运行日志信息。

背景信息

- 支持查看所有作业的配置信息。
- 只有运行中的作业才能查看运行日志信息。

  由于Spark SQL和Distcp作业在后台无日志，运行中的Spark SQL和Distcp作业不能查看运行日志信息。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群基本信息页面。

步骤3 选择“作业管理”。

步骤4 在需要查看作业对应的“操作”列中，单击“查看详情”。

弹出“查看详情”窗口，显示该作业的配置信息。

步骤5 选择一个运行中的作业，在作业对应的“操作”列中，单击“查看日志”。

弹出一个新页面，显示作业执行的实时日志信息。

每个租户并发提交作业和查看日志的个数均为10。

----结束

5.5.9 停止作业

本章节介绍如何手动停止正在运行的MRS作业。

背景信息

Spark SQL作业不支持停止。作业停止后状态更新为“已终止”，并且该作业不可重新执行。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名。

进入集群基本信息页面。

步骤3 选择“作业管理”。

步骤4 选择一个运行中的作业，在作业对应的“操作”列中，选择“更多 > 停止”。

作业状态由“运行中”更新为“已终止”。

----结束

5.5.10 删除作业

本章节介绍如何删除MRS作业，作业执行完成后，若不需要再查看使用其相关信息，可以选择删除作业。

背景信息

支持删除单个作业和批量删除作业。作业删除后不可恢复，请谨慎操作。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名。

进入集群基本信息页面。

步骤3 选择“作业管理”。

步骤4 在需要删除作业对应的“操作”列中，选择“更多 > 删除”。

此处只能删除单个作业。

步骤5 勾选多个作业，单击作业列表左上方的“删除”。

可以删除一个、多个或者全部作业。

----结束

5.5.11 使用 OBS 加密数据运行作业

支持使用OBS文件系统中加密后的数据来运行作业，同时支持将加密后的作业运行结果存储在OBS文件系统中。目前仅支持通过OBS协议访问数据。

OBS支持使用KMS密钥的加解密方式对数据进行加解密，所有的加解密操作都在OBS完成，同时密钥管理在DEW服务。

如需在MRS中使用OBS加密功能，用户需要有“KMS Administrator”权限，且需要在相应组件进行如下配置。

说明

如果集群同时开启“**OBS权限控制**”功能，此时会使用ECS配置的默认委托“MRS_ECS_DEFAULT_AGENCY”或者用户设置的自定义委托的AK/SK访问OBS服务，同时OBS服务会使用接收到的AK/SK访问数据加密服务获取KMS密钥状态，因此需要在使用的委托上绑定“KMS Administrator”策略，否则在处理加密数据时OBS会返回“403 Forbidden”的错误信息。目前MRS服务会在默认委托“MRS_ECS_DEFAULT_AGENCY”绑定“KMS Administrator”策略，用户使用的自定义委托则需要用户自己绑定。

前提条件

如需使用OBS加密功能，请先配置MRS访问OBS功能，具体请参考**配置存算分离集群（委托方式）**。

Hive 配置

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > Hive > 服务配置”。

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-39 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。

参数	取值	说明
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

----结束

Hadoop 配置

方式一：通过界面配置。

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > HDFS > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-40 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

步骤5 以root用户登录Master节点，密码为用户创建集群时设置的root密码（若集群存在多个Master节点，请分别登录每个Master节点进行**步骤5**~**步骤7**的操作）。

步骤6 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤7 执行以下命令更新客户端配置，并输入用户名和密码，用户名为admin，密码为用户创建集群时设置的admin密码。

./ autoRefreshConfig.sh

----结束

方式二：通过客户端配置文件配置。

在Master节点上的客户端配置文件（例如“/opt/Bigdata/client/HDFS/hadoop/etc/hadoop/core-site.xml”）中的增加如下参数配置（若集群存在多个Master节点，请分别登录每个Master节点进行该操作）。

表 5-41 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 • true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 • false：关闭安全连接。

HBase 配置

方式一：通过界面配置。

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > HBase > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-42 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 • true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 • false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

步骤5 以root用户登录Master节点，密码为用户创建集群时设置的root密码（若集群存在多个Master节点，请分别登录每个Master节点进行**步骤5**~**步骤7**的操作）。

步骤6 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤7 执行以下命令更新客户端配置，并输入用户名和密码，用户名为admin，密码为用户创建集群时设置的admin密码。

./ autoRefreshConfig.sh

----结束

方式二：通过客户端配置文件配置。

在Master节点上的客户端配置文件（例如“/opt/Bigdata/client/HBase/hbase/conf/core-site.xml”）中的增加如下参数配置（若集群存在多个Master节点，请分别登录每个Master节点进行该操作）。

表 5-43 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。

参数	取值	说明
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

Spark 配置

方式一：通过界面配置。

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > Spark > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-44 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

步骤5 以root用户登录Master节点，密码为用户创建集群时设置的root密码（若集群存在多个Master节点，请分别登录每个Master节点进行步骤5~步骤7的操作）。

步骤6 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤7 执行以下命令更新客户端配置，并输入用户名和密码，用户名为admin，密码为用户创建集群时设置的admin密码。

./autoRefreshConfig.sh

----结束

方式二：通过客户端配置文件配置。

在Master节点上的客户端配置文件（例如“/opt/Bigdata/client/Spark/spark/conf/core-site.xml”）中的增加如下参数配置（若集群存在多个Master节点，请分别登录每个Master节点进行该操作）。

表 5-45 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

Presto 配置

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > Presto > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-46 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。

参数	取值	说明
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

----结束

5.5.12 配置作业消息通知

MRS联合消息通知服务（SMN），采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式（短信和邮件通知）。通过配置作业消息通知可以实现您在作业执行成功或作业执行失败时能立即接收到通知。

操作步骤

步骤1 登录管理控制台。

步骤2 单击“服务列表”选择“管理与监管 > 消息通知服务”，进入消息通知服务页面。

步骤3 创建主题并向主题中添加订阅，具体请参考**配置消息通知**。

步骤4 进入MRS管理控制台，单击集群名称进入集群详情页面。

步骤5 选择“告警管理 > 消息订阅规则 > 添加消息订阅规则”。

步骤6 配置向订阅者发送作业执行结果消息的规则。

表 5-47 消息订阅规则参数说明

参数	说明
规则名称	用户自定义发送订阅消息的规则名称，只能包含数字、英文字符、中划线和下划线。
提醒通知	选择开启，将向订阅者发送对应订阅消息。
主题名称	选择已创建的主题，也可以单击“创建主题”重新创建。
消息类型	选择“事件”。

参数	说明
订阅规则	1. 单击“提示”前的。 2. 单击“Manager”前的。 3. 勾选“作业执行成功”和“作业执行失败”。

----结束

5.6 组件管理

5.6.1 对象管理简介

MRS集群包含了各类不同的基本对象，不同对象的描述介绍如表5-48所示：

表 5-48 MRS 基本对象概览

对象	描述	举例
服务	可以完成具体业务的一类功能集合。	例如KrbServer服务和LdapServer服务。
服务实例	服务的具体实例，一般情况下可使用服务表示。	例如KrbServer服务。
服务角色	组成一个完整服务的一类功能实体，一般情况下可使用角色表示。	例如KrbServer由KerberosAdmin角色和KerberosServer角色组成。
角色实例	服务角色在主机节点上运行的具体实例。	例如运行在Host2上的KerberosAdmin，运行在Host3上的KerberosServer。
主机	一个弹性云服务器，可以运行Linux系统。	例如Host1～Host5。
机架	一组包含使用相同交换机的多个主机集合的物理实体。	例如Rack1，包含Host1～Host5。
集群	由多台主机组成的可以提供多种服务的逻辑实体。	例如名为Cluster1的集群由（Host1～Host5）5个主机组成，提供了KrbServer和LdapServer等服务。

5.6.2 查看配置

用户可以在MRS上查看服务（含角色）和角色实例的配置。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

- 查看服务的配置。
  a. 在集群详情页，单击“组件管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“服务配置”。
  d. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
  e. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
     在服务节点下的参数属于服务配置参数，在角色节点下的参数是角色配置参数。
  f. 在“——请选择——”选项中选择“非默认”，界面上显示参数值为非默认值的参数。
- 查看角色实例的配置。
  a. 在集群详情页，单击“组件管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“实例”页签。
  d. 单击角色实例列表中指定的角色实例名称。
  e. 单击“实例配置”。
  f. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
  g. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
  h. 在“——请选择——”选项中选择“非默认”，界面上显示参数值为非默认值的参数。

5.6.3 管理服务操作

用户可以在MRS：

- 启动操作状态为“已停止”、“停止失败”或“启动失败”服务，以使用该服务。
- 停止不再使用或异常服务。
- 重启异常或配置过期的服务，以恢复或生效服务功能。

前提条件

- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

对系统影响

- 在Task节点组无法添加stateful的组件。

启动、停止和重启服务

步骤1 在集群详情页，单击“组件管理”。

步骤2 在指定服务所在行，单击“启动”、“停止”和“重启”执行启动、停止和重启操作。

服务之间存在依赖关系。对某服务执行启动、停止和重启操作时，与该服务存在依赖关系的服务将受到影响。

具体影响如下：

- 启动某服务，该服务依赖的下层服务需先启动，服务功能才可生效。
- 停止某服务，依赖该服务的上层服务将无法提供功能。
- 重启某服务，依赖该服务且启动的上层服务需重启后才可生效。

----结束

5.6.4 配置服务参数

用户可以根据实际业务场景，在MRS中快速查看和修改服务默认的配置，及导出或导入配置。

对系统的影响

- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。
- 集群中只剩下一个DBService角色实例时，不支持修改DBService服务的参数。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

修改服务参数

1. 在集群详情页，单击“组件管理”。
2. 单击服务列表中指定的服务名称。
3. 单击“服务配置”。
4. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
5. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

   修改某个参数的值后需要取消修改，可以单击恢复。
6. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

说明

更新YARN服务队列的配置且不重启服务时，在服务状态页签选择“更多 > 刷新队列”更新队列使配置生效。

5.6.5 配置服务自定义参数

MRS各个组件支持开源的所有参数，在MRS支持修改部分关键使用场景的参数，且部分组件的客户端可能不包含开源特性的所有参数。如果需要修改其他MRS未直接支持的组件参数，用户可以在MRS通过自定义配置项功能为组件添加新参数。添加的新参数最终将保存在组件的配置文件中并在重启后生效。

对系统的影响

- 配置服务属性后，需要重启此服务，重启期间无法访问服务。
- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

前提条件

- 用户已充分了解需要新添加的参数意义、生效的配置文件以及对组件的影响。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“服务配置”。

步骤4 将页面右侧“基础配置”切换为“全部配置”。

步骤5 在左侧导航栏选择“自定义”，MRS将显示当前组件的自定义参数。

“参数文件”显示保存用户新添加的自定义参数的配置文件。每个配置文件中可能支持相同名称的开源参数，设置不同参数值后生效结果由组件加载配置文件的顺序决定。自定义参数支持服务级别与角色级别，请根据业务实际需要选择。不支持单个角色实例添加自定义参数。

步骤6 根据配置文件与参数作用，在对应参数项所在行“参数”列输入组件支持的参数名，在“值”列输入此参数的参数值。

- 支持单击和增加或删除一条自定义参数。第一次单击添加自定义参数后才支持删除操作。
- 修改某个参数的值后需要取消修改，可以单击恢复。

步骤7 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

任务示例

配置Hive自定义参数

Hive依赖于HDFS，默认情况下Hive访问HDFS时是HDFS的客户端，生效的配置参数统一由HDFS控制。例如HDFS参数“ipc.client.rpc.timeout”影响所有客户端连接HDFS服务端的RPC超时时间，如果用户需要单独修改Hive连接HDFS的超时时间，可以使用自定义配置项功能进行设置。在Hive的“core-site.xml”文件增加此参数可被Hive服务识别并代替HDFS的设置。

步骤1 在集群详情页，单击“组件管理”。

步骤2 选择 “Hive > 服务配置”。

步骤3 将页面右侧“基础配置”切换为“全部配置”。

步骤4 在左侧导航栏选择Hive服务级别“自定义”，MRS将显示Hive支持的服务级别自定义参数。

步骤5 在“core-site.xml”对应参数“core.site.customized.configs”的“参数”输入“ipc.client.rpc.timeout”，“值”输入新的参数值，例如“150000”。单位为毫秒。

步骤6 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

5.6.6 同步服务配置

操作场景

当用户发现部分服务的“配置状态”为“配置超期”或“配置失败”时，您可以尝试使用同步配置功能，以恢复配置状态。或者集群中所有服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

对系统的影响

同步服务配置后，需要重启配置过期的服务。重启时对应的服务不可用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 在服务列表中，单击指定服务名称。

步骤3 在服务状态页签，选择“更多 > 同步配置”。

步骤4 在弹出窗口勾选“重启配置过期的服务”，并单击“是”重启配置过期的服务。

----结束

5.6.7 管理角色实例操作

操作场景

用户可以在MRS启动操作状态为“停止”、“停止失败”或“启动失败”角色实例，以使用该角色实例，也可以停止不再使用或异常的角色实例，或者重启异常的角色实例，以恢复角色实例功能。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“实例”页签。

步骤4 勾选待操作角色实例前的复选框。

步骤5 选择“更多 > 启动实例”、“停止实例”、“重启实例”或“滚动重启实例”等，执行相应操作。

----结束

5.6.8 配置角色实例参数

操作场景

用户可以根据实际业务场景，在MRS中快速查看及修改角色实例默认的配置。支持导出或导入配置。

对系统的影响

配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

修改角色实例参数

1. 在集群详情页，单击“组件管理”。
2. 单击服务列表中指定的服务名称。
3. 单击“实例”页签。
4. 单击角色实例列表中指定的角色实例名称。
5. 单击“实例配置”页签。

6. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
7. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

   修改某个参数的值后需要取消修改，可以单击恢复。
8. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”，重启角色实例。

5.6.9 同步角色实例配置

操作场景

当用户发现角色实例的“配置状态”为“配置超期”或“配置失败”时，可以在MRS尝试使用同步配置功能，同步角色实例的配置数据与后台配置数据，以恢复配置状态。

对系统的影响

同步配置角色实例后需要重启配置过期的角色实例。重启时对应的角色实例不可用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 选择服务名称。

步骤3 单击“实例”页签。

步骤4 在角色实例列表中，单击指定角色实例名称。

步骤5 在角色实例状态及指标信息上方，选择“更多 > 同步配置”。

步骤6 在弹出窗口勾选“重启配置过期的服务”，并单击“是”重启角色实例。

----结束

5.6.10 退服和入服角色实例

操作场景

某个Core或Task节点出现问题时，可能导致整个集群状态显示为“异常”。MRS集群支持将数据存储在不同Core节点，用户可以在MRS指定角色实例退服，使退服的角色实例不再提供服务。在排除故障后，可以将已退服的角色实例入服。

支持退服、入服的角色实例包括：

- HDFS的DataNode角色实例
- Yarn的NodeManager角色实例

- HBase的RegionServer角色实例
- ClickHouse的ClickHouseServer角色实例
- Kafka的Broker角色实例

限制：

- 当DataNode数量少于或等于HDFS的副本数时，不能执行退服操作。例如HDFS副本数为3时，则系统中少于4个DataNode，将无法执行退服，MRS在执行退服操作时会等待30分钟后报错并退出执行。
- Kafka Broker数量少于或等于副本数时，不能执行退服。例如Kafka副本数为2时，则系统中少于3个节点，将无法执行退服，MRS执行退服操作时会失败并退出执行。
- 已经退服的角色实例，必须执行入服操作启动该实例，才能重新使用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 单击服务列表中相应服务。

步骤3 单击“实例”页签。

步骤4 勾选指定角色实例名称前的复选框。

步骤5 选择“更多 > 退服”或“入服”执行相应的操作。

说明

实例退服操作未完成时在其他浏览器窗口重启集群中相应服务，可能导致MRS提示停止退服，实例的“操作状态”显示为“已启动”。实际上后台已将该实例退服，请重新执行退服操作同步状态。

----结束

5.6.11 启动及停止集群

集群是包含着服务组件的集合。用户可以启动或者停止集群中所有服务。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

启动及停止集群

在集群详情页，单击页面右上角“管理操作 > 启动所有组件”或“停止所有组件”执行相应的操作。

5.6.12 同步集群配置

操作场景

当MRS显示全部服务或部分服务的“配置状态”为“过期”或“失败”时，用户可以尝试使用同步配置功能，以恢复配置状态。

- 若集群中所有服务的配置状态为“失败”时，同步集群的配置数据与后台配置数据。
- 若集群中某些服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

说明

MRS 3.x版本暂不支持在管理控制台执行本章节操作。

对系统的影响

同步集群配置后，需要重启配置过期的服务。重启时对应的服务不可用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击页面右上角“配置 > 同步配置”。

步骤2 在弹出窗口勾选“重启配置过期的服务或实例。”，并单击“确定”，重启配置过期的服务。

界面提示“操作成功”，单击“完成”，集群成功启动。

----结束

5.6.13 导出集群的配置数据

操作场景

为了满足实际业务的需求，用户可以在MRS中将集群所有配置数据导出，导出文件用于快速更新服务配置。

说明

MRS 3.x版本暂不支持在管理控制台执行本章节操作。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

在集群详情页，单击页面右上角“配置 > 导出集群配置”。

导出文件用于更新服务配置，请参见**配置服务参数**中**导入服务配置参数**。

5.6.14 支持滚动重启

在修改了大数据组件的配置项后，需要重启对应的服务来使得配置生效，使用普通重启方式会并发重启所有服务或实例，可能引起业务断服。为了确保服务重启过程中，尽量减少或者不影响业务运行，可以通过滚动重启来按批次重启服务或实例（对于有主备状态的实例，会先重启备实例，再重启主实例）。滚动重启方式的重启时间比普通重启时间久。

当前MRS集群中，服务和实例是否支持滚动重启如**表5-49**所示。

表 5-49 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是

服务	实例	是否支持滚动重启
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	
	Logviewer	

使用限制

- 请在低业务负载时间段进行滚动重启操作。
  - 例如：在滚动重启kafka服务时候， 如果kafka服务业务吞吐量很高（100M/s以上的情况下），会出现kafka服务滚动重启失败的情况。
  - 例如：在滚动重启HBase服务时候，如果原生界面上每个RegionServer上每秒的请求数超过1W，需要增大handle数来预防重启过程中负载过大导致的RegionServer重启失败。
- 重启前需要观察当前hbase的负载请求数（原生界面上每个rs的请求数如果超过1W，需要增大handle数来预防到时候负载不过来）
- 在集群Core节点个数小于6个的情况下，可能会出现业务短时间受影响的情况。
- 请优先使用滚动重启操作来重启实例或服务，并勾选“仅重启配置过期的实例”。

滚动重启服务

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“组件管理”，选择需要滚动重启的服务，进入服务页面。

步骤3 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤4 弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启服务。

步骤5 滚动重启任务完成后，单击“完成”。

----结束

滚动重启实例

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“组件管理”，选择需要滚动重启的服务，进入服务页面。

步骤3 在“实例”页签，勾选要重启的实例，单击“更多”，选择“滚动重启实例”。

步骤4 弹出“滚动重启实例”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启实例。

步骤5 滚动重启任务完成后，单击“完成”。

----结束

滚动重启集群

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 在页面右上角选择“管理操作 > 滚动重启集群”。

步骤3 弹出“滚动重启集群”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启集群。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启参数说明

滚动重启参数说明如**表5-50**所示。

表 5-50 滚动重启参数说明

参数名称	描述
仅重启配置过期的实例	是否只重启集群内修改过配置的实例。
数据节点滚动重启并发数	采用分批并发滚动重启策略的数据节点实例每一个批次重启的实例数，默认为1，取值范围为1～20。只对数据节点有效。
批次时间间隔	滚动重启实例批次之间的间隔时间，默认为0，取值范围为0～2147483647，单位为秒。 说明：设置批次时间间隔参数可以增加滚动重启期间大数据组件进程的稳定性。建议设置该参数为非默认值，例如10。
批次容错阈值	滚动重启实例批次执行失败容错次数，默认为0，即表示任意一个批次的实例重启失败后，滚动重启任务终止。取值范围为0～2147483647。

典型场景操作步骤

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“组件管理”，选择HBase，进入HBase服务页面。

步骤3 单击“服务配置”页签，修改HBase某个参数并保存配置，在出现如下弹窗后，单击“确定”进行保存。

说明

不要勾选“重新启动受影响的服务或实例”，该处重启是普通重启方式，会并发重启所有服务或实例，引起业务断服。

步骤4 保存配置完成后，单击“完成”。

步骤5 选择“服务状态”页签。

步骤6 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤7 弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启。

步骤8 滚动重启任务完成后，单击“完成”。

----结束

5.7 告警管理

5.7.1 查看告警列表

告警列表显示了MRS集群中的所有告警信息，MRS界面显示需要用户及时处理的“告警”和标志事情发生的“事件。

MRS管理控制台“告警管理”只能查询MRS中未清除告警的基本信息，查看详细信息或管理告警具体请参见**查看与手动清除告警**。

告警列表默认按时间顺序排列，时间最近的告警显示在最前端。

告警信息中的各字段说明如**表5-51**所示。

表 5-51 告警说明

参数	参数说明
告警ID	告警的ID。
告警名	告警的名称。

参数	参数说明
级别	告警级别。 MRS 3.x之前版本集群告警级别为： ● 致命 指集群服务不可用，节点故障、GaussDB主备数据不同步、LdapServer数据同步异常等影响集群正常运行的告警，需要根据告警及时检查集群情况并恢复。 ● 严重 指集群部分功能不可用的告警，包括进程故障、周期备份任务失败、关键文件权限异常等，需要根据告警及时检查报告告警的对象并恢复。 ● 一般 指不影响当前集群主要功能的告警，包括证书文件即将过期、审计日志转储失败、License文件即将过期等告警。 ● 提示 指级别最低的一种告警，起到信息展示或信息提示的作用，标识这件事情的发生，一般包括：停止服务、删除服务、停止实例、删除实例、删除节点、重启服务、重启实例、Manager主备倒换、缩容主机、实例恢复、实例故障、作业执行成功、作业执行失败等。 MRS 3.x及之后版本集群告警级别为： ● 紧急 指集群服务不可用，节点故障、GaussDB主备数据不同步、LdapServer数据同步异常等影响集群正常运行的告警，需要根据告警及时检查集群情况并恢复。 ● 重要 指集群部分功能不可用的告警，包括进程故障、周期备份任务失败、关键文件权限异常等，需要根据告警及时检查报告告警的对象并恢复。 ● 次要 指不影响当前集群主要功能的告警，包括证书文件即将过期、审计日志转储失败、License文件即将过期等告警。 ● 提示 指级别最低的一种告警，起到信息展示或信息提示的作用，标识这件事情的发生，一般包括：停止服务、删除服务、停止实例、删除实例、删除节点、重启服务、重启实例、Manager主备倒换、缩容主机、实例恢复、实例故障、作业执行成功、作业执行失败等。
生成时间	产生告警的时间。
定位信息	告警的详细信息。
操作	当告警可手动清除时，单击“清除告警”进行处理。

表 5-52 按钮说明

按钮	说明
	在下拉框中选择刷新告警列表的周期。 ● 每30s刷新一次 ● 每60s刷新一次 ● 停止刷新
	在下拉框中选择告警级别，筛选告警。 MRS 3.x之前版本集群可筛选告警包括：全部、致命、严重、一般、提示。 MRS 3.x及之后版本集群可筛选告警包括：全部、紧急、重要、次要、提示。
	单击  ，手动刷新告警列表。
高级搜索	单击“高级搜索”显示告警搜索区域，设置查询条件后，单击“搜索”，查看指定的告警信息。单击“重置”清除输入的搜索条件。

5.7.2 查看事件列表

事件列表显示了集群中的所有事件信息，如重启服务、停止服务等。

事件列表默认按时间顺序排列，时间最近的告警显示在最前端。

事件信息中的各字段说明如**表1 事件说明**所示。

表 5-53 事件说明

参数	参数说明
事件ID	事件的ID。
事件级别	事件级别。 MRS 3.x之前版本集群事件级别为： ● 致命 ● 严重 ● 一般 ● 提示 MRS 3.x及之后版本集群事件级别为： ● 紧急 ● 重要 ● 次要 ● 提示

参数	参数说明
事件名称	产生事件的名称。
生成时间	产生事件的时间。
定位信息	定位事件的详细信息。

表 5-54 按钮说明

按钮	说明
	在下拉框中选择刷新事件列表的周期。 ● 每30s刷新一次 ● 每60s刷新一次 ● 停止
	单击  ，手动刷新事件列表。
高级搜索	单击“高级搜索”显示事件搜索区域，设置查询条件后，单击“搜索”，查看指定的事件信息。单击“重置”清除输入的搜索条件。

导出事件

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“告警管理 > 事件”。

步骤3 单击“全部导出”。

步骤4 在弹框内选择保存类型，单击“确定”。

----结束

常见事件列表

表 5-55 常见事件列表

事件ID	事件名称
12019	停止服务
12020	删除服务
12021	停止实例
12022	删除实例
12023	删除节点

事件ID	事件名称
12024	重启服务
12025	重启实例
12026	Manager主备倒换
12065	进程重新启动
12070	作业执行成功
12071	作业执行失败
12072	作业被终止
12086	Agent进程重启
14005	NameNode主备倒换
14028	HDFS磁盘均衡任务
14029	主NameNode进入安全模式并生产新的Fsimage
17001	Oozie工作流执行失败
17002	Oozie定时任务执行失败
18001	ResourceManager主备倒换
18004	JobHistoryServer主备倒换
19001	HMaster主备倒换
20003	Hue发生主备切换
24002	Flume Channel溢出
25001	LdapServer主备倒换
27000	DBServer主备倒换
38003	Topic数据保存周期配置调整
43014	Spark2x数据倾斜
43015	Spark2x SQL超大查询结果
43016	Spark2x SQL执行超时
43024	启动JDBCServer
43025	停止JDBCServer
43026	ZooKeeper连接成功
43027	ZooKeeper连接异常

5.7.3 查看与手动清除告警

操作场景

用户可以在MRS上查看、清除告警。

一般情况下，告警处理后，系统自动清除该条告警记录。当告警不具备自动清除功能且用户已确认该告警对系统无影响时，可手动清除告警。

在MRS界面可查看最近十万条告警（包括未清除的、手动清除的和自动清除的告警）。如果已清除告警超过十万条达到十一万条，系统自动将最早的一万条已清除告警转存，转存路径为：

3.x以前版本，主管理节点的“${BIGDATA_HOME}/OMSV100R001C00x8664/workspace/data”。

3.x及后续版本，主管理节点的“${BIGDATA_HOME}/om-server/OMS/workspace/data”。

第一次转存告警时自动生成目录。

说明

用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止”：停止刷新。

操作步骤

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“告警管理”，在告警列表查看告警信息。

- 告警列表每页默认显示最近的十条告警。
- 默认以“生成时间”列按降序排列。针对MRS 3.x之前版本集群，单击“告警ID”、“级别”、“生成时间”可修改排列方式；针对MRS 3.x及以后版本集群，单击“级别”、“生成时间”可修改排列方式。
- 支持在告警“级别”筛选相同级别的全部告警。结果包含已清除和未清除的告警。
- 针对MRS 3.x之前版本集群分别单击页面右上角 、 、 或 可以快速筛选级别为“致命”、“严重”、“一般”或“提示”的未清除告警。
- 针对MRS 3.x及之后版本集群分别单击页面右上角 、 、 或 可以快速筛选级别为“紧急”、“重要”、“次要”或“提示”的未清除告警。

步骤3 单击“高级搜索”显示告警搜索区域，设置查询条件后，单击“搜索”，查看指定的告警信息。单击“重置”清除输入的搜索条件。

说明

“起止时间”表示时间范围的开始时间和结束时间，可以搜索此时间段内产生的告警。

查看“告警参考”章节告警帮助，按照帮助指导处理告警。如果某些场景中告警由于MRS依赖的其他云服务产生，可能需要联系对应云服务运维人员处理。

步骤4 处理完告警后，若需手动清除，单击“清除告警”，手动清除告警。

说明

如果有多个告警已完成处理，可选中一个或多个待清除的告警，单击“清除告警”，批量清除告警。每次最多批量清除300条告警。

----结束

导出告警

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“告警管理 > 告警”。

步骤3 单击“全部导出”。

步骤4 在弹框内选择“保存类型”，单击“确定”。

----结束

5.8 补丁管理

5.8.1 MRS 3.x 之前版本补丁操作指导

当您通过如下途径获知集群版本补丁信息，请根据您的实际需求进行补丁升级操作。

- 通过消息中心服务推送的消息获知MapReduce服务发布了补丁信息。
- 进入现有集群，查看“补丁信息”页面，呈现补丁信息。

安装补丁前准备

- 请参见**执行健康检查**检查集群状态，确认集群健康状态正常后再安装补丁。
- 您根据“补丁内容”中的补丁信息描述，确认将要安装的目标补丁。

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“安装”，安装目标补丁。

说明

- 对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“卸载”，卸载目标补丁。

说明

- 对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

5.8.2 滚动补丁

滚动补丁是指在补丁安装/卸载时，采用滚动重启服务（按批次重启服务或实例）的方式，在不中断或尽可能短地中断集群各个服务业务的前提下完成对集群中单个或多个服务的补丁安装/卸载操作。集群中的服务根据对滚动补丁的支持程度，分为三种：

- 支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的全部业务或部分业务（因服务而异，不同服务存在差别）不中断。
- 不支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的业务会中断。
- 部分角色支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的部分业务不中断。

说明

MRS 3.x版本暂不支持在管理控制台执行本章节操作。

当前MRS集群中，服务和实例是否支持滚动重启如**表5-56**所示。

表 5-56 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	

服务	实例	是否支持滚动重启
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	
	Logviewer	

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“安装”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动安装补丁功能开启：补丁安装前不会停止服务，补丁安装后滚动重启服务来完成补丁安装，可以减少对集群业务的影响，但相比普通方式安装耗时更久。
- 滚动安装补丁功能关闭：补丁安装前会停止服务，补丁安装后再重新启动服务来完成补丁安装，会造成集群和服务暂时中断，但相比滚动方式安装补丁耗时更短。
- 少于2个Master节点和少于3个Core节点的集群不支持滚动方式安装补丁。

步骤5 单击“是”，安装目标补丁。

步骤6 查看补丁安装进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁安装进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“卸载”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动卸载补丁功能开启：补丁卸载前不会停止服务，补丁卸载后滚动重启服务来完成补丁卸载，可以减少对集群业务的影响，但相比普通方式卸载耗时更久。
- 滚动卸载补丁功能关闭：补丁卸载前会停止所有服务，补丁卸载后再重新启动所有服务来完成补丁卸载，会造成集群和服务暂时中断，但相比滚动方式卸载补丁耗时更短。
- 仅通过滚动方式安装的补丁支持滚动方式卸载补丁。

步骤5 单击“是”，卸载目标补丁。

步骤6 查看补丁卸载进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁卸载进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

5.8.3 修复隔离主机补丁

若集群中存在主机被隔离的情况，集群补丁安装完成后，请参见本节操作对隔离主机进行补丁修复。修复完成后，被隔离的主机节点版本将与其他未被隔离的主机节点一致。

说明

MRS 3.x版本暂不支持在管理控制台执行本章节操作。

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“详情”。

步骤4 在补丁详情界面，选中“Status”是“Isolated”的主机节点。

步骤5 单击“Select and Restore”，修复被隔离的主机节点。

----结束

5.9 租户管理

5.9.1 使用前须知

本章节指导用户在MRS控制台执行租户管理操作。

在控制台界面执行租户管理操作仅适用于**MRS 3.x之前版本**集群。

在Manager界面执行租户管理操作适用于所有版本，MRS 3.x及之后版本请参考**简介**，MRS 3.x之前版本请参考**租户简介**。

5.9.2 租户简介

定义

MRS集群拥有的不同资源和服务支持多个组织、部门或应用共享使用。集群提供了一个逻辑实体来统一使用不同资源和服务，这个逻辑实体就是租户。多个不同的租户统称多租户。当前仅分析集群支持租户。

原理

MRS集群提供多租户的功能，支持层级式的租户模型，支持动态添加和删除租户，实现资源的隔离，可以对租户的计算资源和存储资源进行动态配置和管理。

计算资源指租户Yarn任务队列资源，可以修改任务队列的配额，并查看任务队列的使用状态和使用统计。

存储资源目前支持HDFS存储，可以添加删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额。

租户可以在界面上根据业务需要，在集群中创建租户、管理租户。

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。默认情况下，新的计算资源和存储资源的全部权限将分配给租户的角色。
- 默认情况下，查看当前租户的资源，在当前租户中添加子租户并管理子租户资源的权限将分配给租户的角色。
- 修改租户的计算资源或存储资源，对应的角色关联权限将自动更新。

MRS中最多支持512个租户。系统默认创建的租户包含“default”。和默认租户同处于最上层的租户，可以统称为一级租户。

资源池

YARN任务队列支持一种调度策略，称为标签调度（Label Based Scheduling）。通过此策略，YARN任务队列可以关联带有特定节点标签（Node Label）的NodeManager，使YARN任务在指定的节点运行，实现任务的调度与使用特定硬件资源的需求。例如，需要使用大量内存的YARN任务，可以通过标签关联具有大量内存的节点上运行，避免性能不足影响业务。

在MRS集群中，租户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个资源池。YARN任务队列通过配置队列容量策略，与指定的资源池进行关联，可以更有效地使用资源池中的资源，且互不影响。

MRS中最多支持50个资源池。系统默认包含一个“default”资源池。

5.9.3 添加租户

操作场景

当租户需要根据业务需求指定资源使用情况时，可以在MRS创建租户。

前提条件

- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的完整存储路径在HDFS目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“添加租户”，打开添加租户的配置页面，参见以下表格内容为租户配置属性。

表 5-57 租户参数一览表

参数名	描述
名称	指定当前租户的名称，长度为3到50，可包含数字、字母和下划线。
租户类型	可选参数值为“叶子租户”和“非叶子租户”。当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
动态资源	为当前租户选择动态计算资源。系统将自动在Yarn中以租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量 (%)	配置当前租户在“default”资源池中使用的计算资源百分比。
默认资源池最大容量 (%)	配置当前租户在“default”资源池中使用的最大计算资源百分比。

参数名	描述
储存资源	为当前租户选择存储资源。系统将自动在“/tenant”目录中以租户名称创建文件夹。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。
存储空间配额 (MB)	配置当前租户使用的HDFS存储空间配额。取值范围为“1”到“8796093022208”。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间表示所有副本文件在HDFS中占用的磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
存储路径	配置租户在HDFS中的存储目录。系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“tenant/ta1”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
服务	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
描述	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配Manager_tenant角色以及租户对应的角色。具体操作请参见**创建用户**。

----结束

相关任务

查看已添加的租户

步骤1 在集群详情页，单击“租户管理”。

步骤2 在左侧租户列表，单击已添加租户的名称。

默认在右侧显示“概述”页签。

步骤3 查看当前租户的“基本信息”、“资源配额”和“统计”。

如果HDFS处于“已停止”状态，“资源配额”中“Space”的“可用”和“已使用”会显示为“unknown”。

----结束

5.9.4 添加子租户

操作场景

当租户需要根据业务需求，将当前租户的资源进一步分配时，可以在MRS添加子租户。

前提条件

- 已添加“非叶子租户”。
- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果子租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的存储目录在父租户的存储目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 在左侧租户列表，将光标移动到需要添加子租户的租户节点上，单击“添加子租户”，打开添加子租户的配置页面，参见以下表格内容为租户配置属性。

表 5-58 子租户参数一览表

参数名	描述
父租户	显示上级父租户的名称。
名称	指定当前租户的名称，长度为3到20，可包含数字、字母和下划线。
租户类型	可选参数值为“叶子租户”和“非叶子租户”，当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。

参数名	描述
动态资源	为当前租户选择动态计算资源。系统将自动在Yarn父租户队列中以子租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。如果父租户未选择动态资源，子租户也无法使用动态资源。
默认资源池容量 (%)	配置当前租户使用的资源百分比，基数为父租户的资源总量。
默认资源池最大容量 (%)	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。
储存资源	为当前租户选择存储资源。系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。如果父租户未选择存储资源，子租户也无法使用存储资源。
存储空间配额 (MB)	配置当前租户使用的HDFS存储空间配额。最小值值为“1”，最大值为父租户的全部存储配额。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。若此配额大于父租户的配额，实际存储量受父租户配额影响。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“父租户中分配资源”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
存储路径	配置租户在HDFS中的存储目录。系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“tenant/ta1”，系统默认自动配置此参数值为“tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。支持在父目录中自定义存储路径。存储路径的父目录必需是父租户的存储目录。
服务	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
描述	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成子租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配租户对应的角色。具体操作请参见**创建用户**。

----结束

5.9.5 删除租户

操作场景

当租户需要根据业务需求，将当前不再使用的租户删除时，可以在MRS完成操作。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。该任务对应取消角色与用户的绑定，请参见**修改用户信息**。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 在左侧租户列表，将光标移动到需要删除的租户节点上，单击“删除”。

界面显示删除租户对话框。根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户的数据”，否则将自动删除租户对应的存储空间。

步骤3 单击“是”，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

- 租户删除后，Yarn中对应的租户任务队列不会被删除。
- 删除父租户时选择不保留数据，如果存在子租户且子租户使用了存储资源，则子租户的数据也会被删除。

----结束

5.9.6 管理租户目录

操作场景

用户根据业务需求，可以在MRS对指定租户使用的HDFS存储目录，进行管理操作。支持用户对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

- 已添加关联了HDFS存储资源的租户。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

- 查看租户目录
  a. 在集群详情页，单击“租户管理”。

     说明

     MRS 3.x及之后版本请参考**使用说明**。
  b. 在左侧租户列表，单击目标的租户。
  c. 单击“资源”页签。
  d. 查看“HDFS 存储”表格。
     - 指定租户目录的“文件/目录数上限”列表示文件和目录数量配额。
     - 指定租户目录的“存储空间配额”列表示租户目录的存储空间大小。
- 添加租户目录
  a. 在集群详情页，单击“租户管理”。

     说明

     MRS 3.x及之后版本请参考**使用说明**。
  b. 在左侧租户列表，单击需要添加HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS 存储”表格，单击“添加目录”。
     - “路径”填写租户目录的路径。

       说明

       - 如果当前租户不是子租户，新路径将在HDFS的根目录下创建。
       - 如果当前租户是一个子租户，新路径将在指定的目录下创建。

       完整的HDFS存储目录最多包含1023个字符。HDFS目录名称包含数字、大小写字母、空格和下划线。空格只能在HDFS目录名称的中间使用。
     - “文件/目录数上限”填写文件和目录数量配额。

       “文件/目录数上限”为可选参数，取值范围从1到9223372036854775806。

- “存储空间配额 ”填写租户目录的存储空间大小。
  “存储空间配额 ”的取值范围从1到8796093022208。

  说明

  为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

e. 单击“确定”完成租户目录添加，系统将在HDFS根目录下创建租户的目录。

- 修改租户目录

a. 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

b. 在左侧租户列表，单击需要修改HDFS存储目录的租户。

c. 单击“资源”页签。

d. 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。

- “文件/目录数上限”填写文件和目录数量配额。
  “文件/目录数上限”为可选参数，取值范围从1到9223372036854775806。
- “存储空间配额”填写租户目录的存储空间大小。
  “存储空间配额”的取值范围从1到8796093022208。

  说明

  为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

e. 单击“确定”完成租户目录修改。

- 删除租户目录

a. 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

b. 在左侧租户列表，单击需要删除HDFS存储目录的租户。

c. 单击“资源”页签。

d. 在“HDFS 存储”表格，指定租户目录的“操作”列，单击“删除”。
   创建租户时设置的默认HDFS存储目录不支持删除，仅支持删除新添加的HDFS存储目录。

e. 单击“确认”完成租户目录删除。

5.9.7 恢复租户数据

操作场景

租户的数据默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，需要手动恢复。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“基本信息”左侧的圆圈，绿色表示租户可用，灰色表示租户不可用。
2. 单击“资源”，查看“Yarn”或者“HDFS 存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击“恢复租户数据”。

步骤5 在“恢复租户数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

5.9.8 添加资源池

操作场景

在MRS集群中，用户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个YARN资源池。每个NodeManager只能属于一个资源池。系统中默认包含了一个名为“default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

该任务指导用户通过MRS添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“资源池”页签。

步骤3 单击“添加资源池”。

步骤4 在“添加资源池”设置资源池的属性。

- “名称”：填写资源池的名称。不支持创建名称为“default”的资源池。

  资源池的名称，长度为1到20个字节，可包含数字、字母和下划线，且不能以下划线开头。
- “可用主机”：在界面左边主机列表，勾选指定的主机名称加入资源池。只支持选择本集群中的主机。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

步骤6 完成资源池创建后，用户可以在资源池的列表中查看资源池的“名称”、“成员”、“类型”、“虚拟核数”与“内存”。已加入自定义资源池的主机，不再是“default”资源池的成员。

----结束

5.9.9 修改资源池

操作场景

该任务指导用户通过MRS修改已有资源池中的成员。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“修改”。

步骤4 在“修改资源池”修改“已添加主机”。

- 增加主机：在界面左边主机列表，勾选指定的主机名称加入资源池。
- 删除主机：在界面右边主机列表，单击指定主机后的×将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

5.9.10 删除资源池

操作场景

该任务指导用户通过MRS删除已有资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列配置**。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

在弹出窗口中单击“确定”。

----结束

5.9.11 配置队列

操作场景

用户根据业务需求，可以在MRS修改指定租户的队列配置。

前提条件

- 已添加关联Yarn并分配了动态资源的租户。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“队列配置”页签。

步骤3 在租户队列表格，指定租户队列的“操作”列，单击“修改”。

说明

- 在“租户管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击 也能打开修改队列配置页面。
- 一个队列只能绑定一个非default资源池。

MRS 3.x之前版本：

表 5-59 队列配置参数

参数名	描述
最大应用数量	表示最大应用程序数量。取值范围从“1”到“2147483647”。
AM最大资源百分比	表示集群中可用于运行application master的最大资源占比。取值范围从“0”到“1”。
用户资源最小上限百分比 (%)	表示用户使用的最小资源上限百分比。取值范围从“0”到“100”。
用户资源上限因子	表示用户使用的最大资源限制因子，与当前租户在集群中实际资源百分比相乘，可计算出用户使用的最大资源百分比。最小值为“0”。
状态	表示资源计划当前的状态，“运行”为运行状态，“停止”为停止状态。
默认资源池	表示队列使用的资源池。默认为“default”，如果需要修改为其他资源，需要先配置队列容量，请参见**配置资源池的队列容量策略**。

MRS 3.x及之后版本：

表 5-60 队列配置参数

参数名	描述
AM最多占有资源（%）	表示当前队列内所有Application Master所占的最大资源百分比。
每个YARN容器最多分配核数	表示当前队列内单个YARN容器可分配的最多核数，默认为-1，表示取值范围内不限制。
每个YARN容器最大分配内存（MB）	表示当前队列内单个YARN容器可分配的最大内存，默认为-1，表示取值范围内不限制。
最多运行任务数	表示当前队列最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。

参数名	描述
每个用户最多运行任务数	表示每个用户在当前队列中最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
最多挂起任务数	表示当前队列最多同时可挂起任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可挂起任务。取值范围为-1～2147483647。
资源分配规则	表示单个用户任务间的资源分配规则，包括FIFO和FAIR。 一个用户若在当前队列上提交了多个任务，FIFO规则代表一个任务完成后再执行其他任务，按顺序执行。FAIR规则代表各个任务同时获取到资源并平均分配资源。
默认资源标签	表示在指定资源标签（Label）的节点上执行任务。 **说明** 如果需要使用新的资源池，需要修改默认标签为新的资源池标签。
Active状态	• ACTIVE表示当前队列可接受并执行任务。 • INACTIVE表示当前队列可接受但不执行任务，若提交任务，任务将处于挂起状态。
Open状态	• OPEN表示当前队列处于打开状态。 • CLOSED表示当前队列处于关闭状态，若提交任务，任务直接会被拒绝。

----结束

5.9.12 配置资源池的队列容量策略

操作场景

添加资源池后，需要为YARN任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。每个队列只能配置一个资源池的队列容量策略。用户可以在任何一个资源池中查看队列并配置队列容量策略。配置队列策略后，YARN任务队列与资源池形成关联关系。

该任务指导用户通过MRS配置队列策略。

前提条件

- 已添加资源池。
- 任务队列与其他资源池无关联关系。默认情况下，所有队列与“default”资源池存在关联关系。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“资源分布策略”页签。

步骤3 在“资源池”选择指定的资源池。

“可用资源配额”：表示每个资源池默认所有资源都可分配给队列。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“修改”。

步骤5 在“修改资源分配”窗口设置任务队列在此资源池中的资源容量策略。

- “资源容量 (%)”：表示当前租户计算资源使用的资源百分比。
- “最大资源容量 (%)”：表示当前租户计算资源使用的最大资源百分比。

步骤6 单击“确定”保存配置。

----结束

5.9.13 清除队列配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，用户可以在MRS清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

- 如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

说明

MRS 3.x及之后版本请参考**使用说明**。

步骤2 单击“资源分布策略”页签。

步骤3 在“资源池”选择指定的资源池。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“清除”。

在“清除队列设置”中单击“是”，清除队列在当前资源池的配置。

说明

如果用户未配置队列的资源容量策略，则清除功能默认不可用。

----结束

6 使用 MRS 客户端

6.1 安装客户端

6.1.1 安装客户端（3.x 及之后版本）

操作场景

该操作指导安装工程师安装MRS集群所有服务（不包含Flume）的客户端。Flume客户端安装请参见“组件操作指南 > 使用Flume > 安装Flume客户端”。

客户端可以安装集群内节点，也可以安装在集群外节点，本章节以安装目录“/opt/client”为例进行介绍，请以实际集群版本为准。

在集群外节点安装客户端前提条件

- 已准备一个Linux弹性云服务器，主机操作系统及版本建议参见表6-1。

  表 6-1 参考列表

CPU架构	操作系统	支持的版本号
x86计算	Euler	Euler OS 2.5
	SuSE	SUSE Linux Enterprise Server 12 SP4（SUSE 12.4）
	Red Hat	Red Hat-7.5-x86_64（Red Hat 7.5）
	CentOS	CentOS-7.6版本（CentOS 7.6）
鲲鹏计算(ARM)	Euler	Euler OS 2.8
	CentOS	CentOS-7.6版本（CentOS 7.6）

  同时为弹性云服务分配足够的磁盘空间，例如“40GB”。

- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。
- 弹性云服务器操作系统已安装NTP服务，且NTP服务运行正常。
  若未安装，在配置了**yum**源的情况下，可执行**yum install ntp -y**命令自行安装。
- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式）。

集群内节点安装客户端

1. 获取软件包。
   访问FusionInsight Manager（MRS 3.x及之后版本），在“集群”下拉列表中单击需要操作的集群名称。
   选择“更多 > 下载客户端”，弹出“下载集群客户端”信息提示框。

   图 6-1 下载客户端

   说明

   在只安装单个服务的客户端的场景中，选择“集群 > 服务 > *服务名称* > 更多 > 下载客户端”，弹出“下载客户端”信息提示框。

2. “选择客户端类型”中选择“完整客户端”。
   “仅配置文件”下载的客户端配置文件，适用于应用开发任务中，完整客户端已下载并安装后，管理员通过Manager界面修改了服务端配置，开发人员需要更新客户端配置文件的场景。
   平台类型包括x86_64和aarch64两种：
   - x86_64：可以部署在X86平台的客户端软件包。
   - aarch64：可以部署在TaiShan服务器的客户端软件包。

   说明

   集群支持下载x86_64和aarch64两种类型客户端，但是客户端类型必须与待安装节点的架构匹配，否则客户端会安装失败。

3. 勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件。

   文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。单击“确定”，等待下载完成后，使用**omm**用户或**root**用户将获取的软件包复制到将要安装客户端的服务器文件目录。

   客户端软件包名称格式为：“FusionInsight_Cluster_*<集群ID>*_Services_Client.tar”。

   后续步骤及章节以FusionInsight_Cluster_1_Services_Client.tar进行举例。

   说明

   当用户无法获取**root**用户权限，需要用**omm**用户操作。

   如需安装客户端至集群内其他节点，则执行以下命令复制客户端到待安装客户端的节点：

   scp -p /*tmp/FusionInsight-Client*/**FusionInsight_Cluster_1_Services_Client.tar** *待安装客户端节点的IP地址:/opt/Bigdata/client*

4. 以**user_client**用户登录将要安装客户端的服务器。
5. 解压软件包。

   进入安装包所在目录，例如“/tmp/FusionInsight-Client”。执行如下命令解压安装包到本地目录。

   tar -xvf FusionInsight_Cluster_1_Services_Client.tar

6. 校验软件包。

   执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

   sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256

   ```
   FusionInsight_Cluster_1_Services_ClientConfig.tar: OK
   ```

7. 解压获取的安装文件。

   tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

8. 进入安装包所在目录，执行如下命令安装客户端到指定目录（绝对路径），例如安装到“/opt/client”目录。

   cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig

   执行**./install.sh /opt/client**命令，等待客户端安装完成（以下只显示部分屏显结果）。

   ```
   The component client is installed successfully
   ```

   说明

   - 如果已经安装的全部服务或某个服务的客户端使用了“/opt/client”目录，再安装其他服务的客户端时，需要使用不同的目录。
   - 卸载客户端请删除客户端安装目录。
   - 如果要求安装后的客户端仅能被该安装用户（如“user_client”）使用，请在安装时加“-o”参数，即执行**./install.sh /opt/client -o**命令安装客户端。
   - 由于HBase使用的Ruby语法限制，如果安装的客户端中包含了HBase客户端，建议客户端安装目录路径只包含大写字母、小写字母、数字以及_-?.@+=字符。

使用客户端

1. 在已安装客户端的节点，执行**sudo su - omm**命令切换用户。执行以下命令切换到客户端目录：

cd /opt/client

2. 执行以下命令配置环境变量：

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit*MRS集群用户*

   例如，**kinit admin**。

   说明

   启用Kerberos认证的MRS集群默认创建“admin”用户帐号，用于集群管理员维护集群。

4. 直接执行组件的客户端命令。

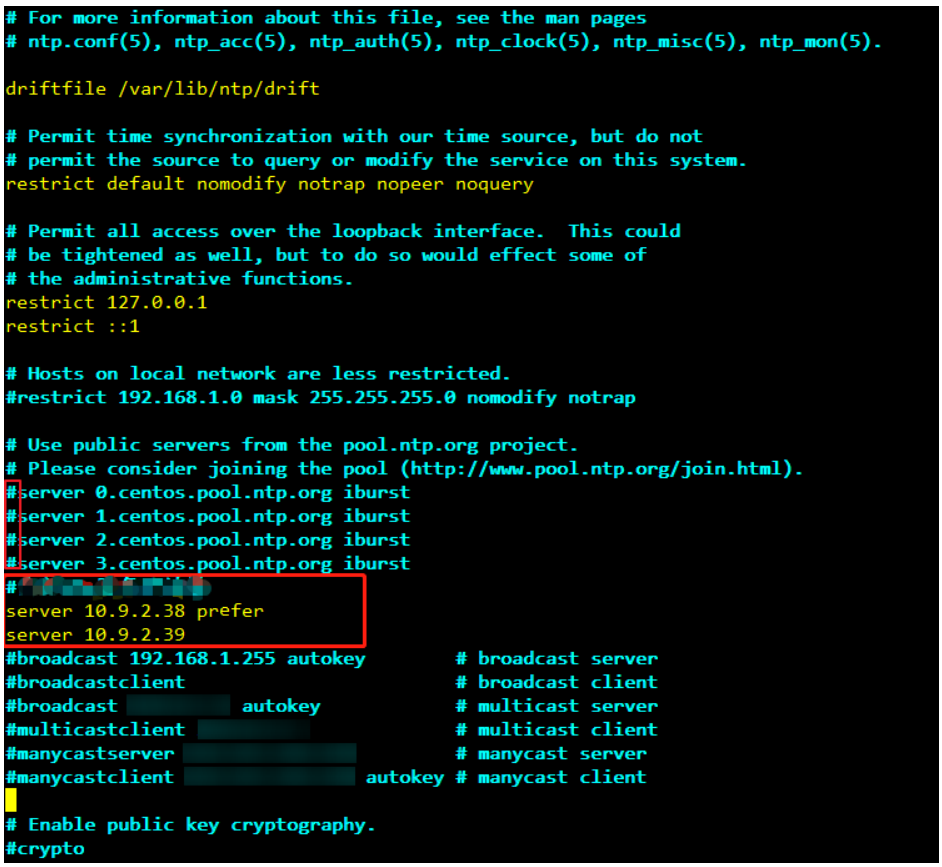
   例如：使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

集群外节点安装客户端

1. 根据**在集群外节点安装客户端前提条件**，创建一个满足要求的弹性云服务器。
2. 执行ntp时间同步，使集群外节点的时间与MRS集群时间同步。
   a. 执行**vi /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

```
server master1_ip prefer
server master2_ip
```

图 6-2 增加 Master 节点的 IP



b. 执行**service ntpd stop**命令关闭NTP服务。
c. 执行**/usr/sbin/ntpdate** *主Master节点的IP地址*命令手动同步一次时间。
d. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。
e. 执行**ntpstat**命令查看时间同步结果。

3. 参考以下步骤，从FusionInsight Manager下载集群客户端软件包并复制到ECS节点后安装客户端。
   a. **访问FusionInsight Manager（MRS 3.x及之后版本）**，参考**集群内节点安装客户端**下载集群客户端到主管理节点的指定目录。
   b. 使用**root**用户登录主管理节点，执行以下命令复制客户端安装包到待安装客户端的节点：

      scp -p /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar *待安装客户端节点的IP地址*:**/tmp**
   c. 使用待安装客户端的用户登录待安装客户端节点。

      执行以下命令安装客户端，如果当前用户无客户端软件包以及客户端安装目录的操作权限，需使用**root**用户进行赋权：

      cd /tmp

      tar -xvf FusionInsight_Cluster_1_Services_Client.tar

      tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

      cd FusionInsight_Cluster_1_Services_ClientConfig

      ./install.sh /opt/client
   d. 执行以下命令，切换到客户端目录并配置环境变量：

      cd /opt/client

      source bigdata_env
   e. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

      kinit *MRS集群用户*

      例如，**kinit admin**。
   f. 直接执行组件的客户端命令。

      例如使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

6.1.2 安装客户端（3.x 之前版本）

操作场景

用户需要使用MRS客户端。MRS集群客户端可以安装在集群内的Master节点或者Core节点，也可以安装在集群外节点上。

MRS 3.x之前版本集群在集群创建后，在主Master节点默认安装有客户端，可以直接使用，安装目录为“/opt/client”。

MRS 3.x及之后版本客户端的安装请参考**安装客户端（3.x及之后版本）**。

说明

如果集群外的节点已安装客户端且只需要更新客户端，请使用安装客户端的用户例如“root”。

在集群外节点安装客户端前提条件

- 已准备一个弹性云服务器，主机操作系统及版本请参见表6-2。

  表 6-2 参考列表

操作系统	支持的版本号
Euler	• 可用：Euler OS 2.2 • 可用：Euler OS 2.3 • 可用：Euler OS 2.5

  例如，用户可以选择操作系统为**Euler**的弹性云服务器准备操作。

  同时为弹性云服务分配足够的磁盘空间，例如“40GB”。
- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。

  如果不同，请修改弹性云服务器安全组或配置弹性云服务器安全组的出入规则允许MRS集群所有安全组的访问。
- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式），请参见弹性云服务器《用户指南》中“实例> 登录Linux弹性云服务器 >SSH密码方式登录”。

在 Core 节点安装客户端

1. 登录MRS Manager页面，选择“服务管理 > 下载客户端”下载客户端安装包至主管理节点。

   说明

   如仅需更新客户端配置文件，请参考**更新客户端（3.x之前版本）**页面的方法二操作。
2. 使用IP地址搜索主管理节点并使用VNC登录主管理节点。
3. 在主管理节点，执行以下命令切换用户。

   sudo su - omm
4. 在MRS管理控制台，查看指定集群“节点管理”页面的“IP”地址。

   记录需使用客户端的Core节点IP地址。
5. 在主管理节点，执行以下命令，将客户端安装包从主管理节点文件拷贝到当前Core节点：

   scp -p /tmp/MRS-client/MRS_Services_Client.tar *Core节点的IP地址*:**/opt/client**
6. 使用“root”登录Core节点。

   Master节点支持Cloud-Init特性，Cloud-init预配置的用户名“root”，密码为创建集群时设置的密码。
7. 执行以下命令，安装客户端：

   cd /opt/client

   tar -xvf MRS_Services_Client.tar

   tar -xvf MRS_Services_ClientConfig.tar

   cd /opt/client/MRS_Services_ClientConfig

./install.sh *客户端安装目录*

例如，执行命令：

./install.sh /opt/client

8. 客户端的使用请参见**使用MRS客户端**。

使用 MRS 客户端

1. 在已安装客户端的节点，执行**sudo su - omm**命令切换用户。执行以下命令切换到客户端目录：

   cd /opt/client

2. 执行以下命令配置环境变量：

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit admin**。

   说明

   启用Kerberos认证的MRS集群默认创建“admin”用户帐号，用于集群管理员维护集群。

4. 直接执行组件的客户端命令。

   例如：使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

在集群外节点上安装客户端

步骤1 根据前提条件，创建一个满足要求的弹性云服务器。

步骤2 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**，然后选择“服务管理”。

步骤3 单击“下载客户端”。

步骤4 在“客户端类型”选择“完整客户端”。

步骤5 在“下载路径”选择“远端主机”。

步骤6 将“主机IP”设置为ECS的IP地址，设置“主机端口”为“22”，并将“存放路径”设置为“/tmp”。

- 如果使用SSH登录ECS的默认端口“22”被修改，请将“主机端口”设置为新端口。
- “存放路径”最多可以包含256个字符。

步骤7 “登录用户”设置为“root”。

如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。

步骤8 在“登录方式”选择“密码”或“SSH私钥”。

- 密码：输入创建集群时设置的root用户密码。
- SSH私钥：选择并上传创建集群时使用的密钥文件。

步骤9 单击“确定”开始生成客户端文件。

若界面显示以下提示信息表示客户端包已经成功保存。单击“关闭”。客户端文件请到下载客户端时设置的远端主机的“存放路径”中获取。

```
下载客户端文件到远端主机成功。
```

若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从**步骤2**执行重新开始下载客户端。

```
连接到服务器失败，请检查网络连接或参数设置。
```

说明

生成客户端会占用大量的磁盘IO，不建议在集群处于安装中、启动中、打补丁中等非稳态场景下载客户端。

步骤10 使用VNC方式，登录弹性云服务器。参见弹性云服务器《用户指南》的**远程登录（VNC方式）**章节（“实例 > 登录Linux弹性云服务器 > 远程登录（VNC方式）”）。

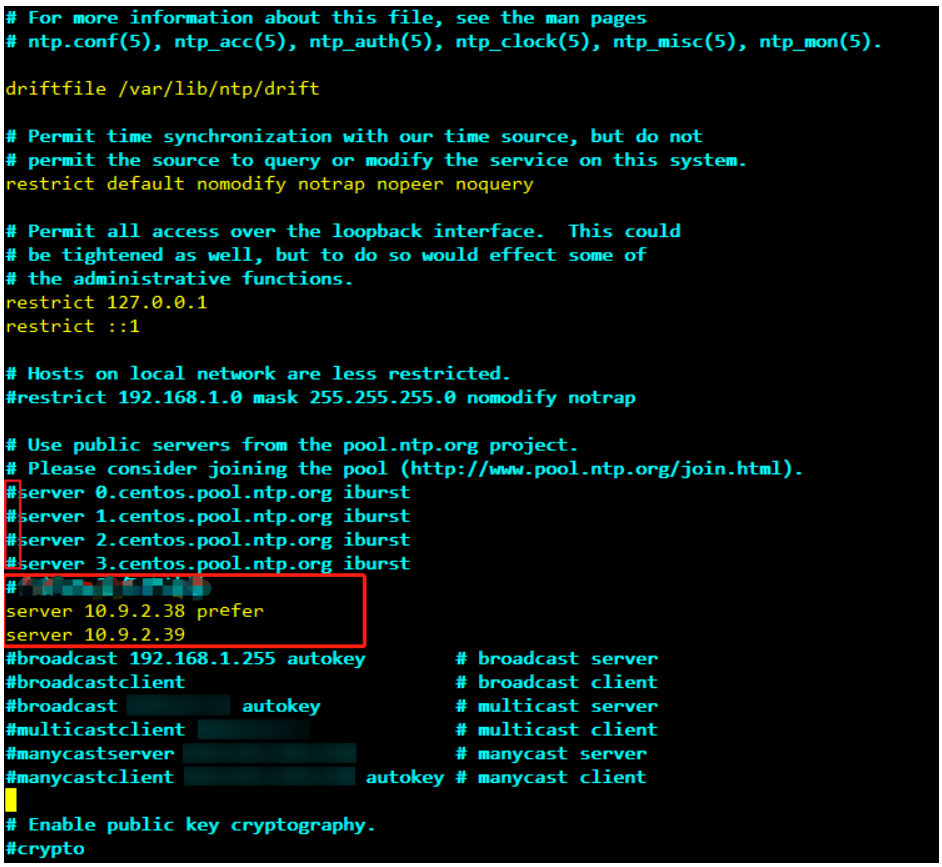
所有镜像均支持Cloud-init特性。Cloud-init预配置的用户名“root”，密码为创建集群时设置的密码。首次登录建议修改。

步骤11 执行ntp时间同步，使集群外节点的时间与MRS集群时间同步。

1. 检查安装NTP服务有没有安装，未安装请执行**yum install ntp -y**命令自行安装。
2. 执行**vim /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

```
server master1_ip prefer
server master2_ip
```

图 6-3 增加 Master 节点的 IP



3. 执行**service ntpd stop**命令关闭NTP服务。
4. 执行**/usr/sbin/ntpdate *主Master节点的IP***命令手动同步一次时间。
5. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。
6. 执行**ntpstat**命令查看时间同步结果。

步骤12 在弹性云服务器，切换到**root**用户，并将**步骤6**中“存放路径”中的安装包复制到目录“/opt”，例如“存放路径”设置为“/tmp”时命令如下。

sudo su - root

cp /tmp/MRS_Services_Client.tar /opt

步骤13 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Services_Client.tar

步骤14 执行以下命令，校验文件包。

sha256sum -c MRS_Services_ClientConfig.tar.sha256

界面显示如下：

```
MRS_Services_ClientConfig.tar: OK
```

步骤15 执行以下命令，解压“MRS_Services_ClientConfig.tar”。

tar -xvf MRS_Services_ClientConfig.tar

步骤16 执行以下命令，安装客户端到新的目录，例如“/opt/Bigdata/client”。安装时自动生成目录。

sh /opt/MRS_Services_ClientConfig/install.sh /opt/Bigdata/client

查看安装输出信息，如有以下结果表示客户端安装成功：

```
Components client installation is complete.
```

步骤17 验证弹性云服务器节点是否与集群Master节点的IP是否连通？

例如，执行以下命令：**ping** *Master节点IP地址*

- 是，执行**步骤18**。
- 否，检查VPC、安全组是否正确，是否与MRS集群在相同VPC和安全组，然后执行**步骤18**。

步骤18 执行以下命令配置环境变量：

source /opt/Bigdata/client/bigdata_env

步骤19 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit*MRS集群用户*

例如, **kinit admin**

步骤20 执行组件的客户端命令。

例如，执行以下命令查看HDFS目录：

hdfs dfs -ls /

----结束

6.2 更新客户端

6.2.1 更新客户端（3.x 及之后版本）

集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户如果在Manager修改了服务配置参数并重启了服务，已安装的客户端需要重新下载并安装，或者使用配置文件更新客户端。

更新客户端配置

方法一：

步骤1 访问FusionInsight Manager（MRS 3.x及之后版本），在“集群”下拉列表中单击需要操作的集群名称。

步骤2 选择“更多 > 下载客户端 > 仅配置文件”。

此时生成的压缩文件包含所有服务的配置文件。

步骤3 是否在集群的节点中生成配置文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。然后执行**步骤4**。
- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，然后执行**步骤4**。

步骤4 使用WinSCP工具，以客户端安装用户将压缩文件保存到客户端安装的目录，例如“/opt/hadoopclient”。

步骤5 解压软件包。

例如下载的客户端文件为“FusionInsight_Cluster_1_Services_Client.tar”执行如下命令进入客户端所在目录，解压文件到本地目录。

cd /opt/hadoopclient

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤6 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar: OK
```

步骤7 解压获取配置文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar

步骤8 在客户端安装目录下执行如下命令，使用配置文件更新客户端。

sh refreshConfig.sh *客户端安装目录 配置文件所在目录*

例如，执行以下命令：

sh refreshConfig.sh /opt/hadoopclient /opt/hadoopclient/FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 以**root**用户登录客户端安装节点。

步骤2 进入客户端安装的目录，例如“/opt/hadoopclient”，执行以下命令更新配置文件：

cd /opt/hadoopclient

sh autoRefreshConfig.sh

步骤3 按照提示输入FusionInsight Manager管理员用户名，密码以及FusionInsight Manager界面浮动IP。

步骤4 输入需要更新配置的组件名，组件名之间使用“,”分隔。如需更新所有组件配置，可直接单击回车键。

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

6.2.2 更新客户端（3.x 之前版本）

说明

本章节适用于MRS 3.x之前版本的集群。MRS 3.x及之后版本，请参考**更新客户端（3.x及之后版本）**。

更新客户端配置文件

操作场景

MRS集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户使用MRS的客户端时，如果在MRS Manager修改了服务配置参数并重启了服务或者重启了服务，需要先下载并更新客户端配置文件。

用户创建集群时，默认在集群所有节点的“/opt/client”目录安装保存了原始客户端。集群创建完成后，仅Master节点的客户端可以直接使用，Core节点客户端在使用前需要更新客户端配置文件。

操作步骤

方法一：

步骤1 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**，然后选择“服务管理”。

步骤2 单击“下载客户端”。

“客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤3 查询并登录主Master节点。

步骤4 若在集群内使用客户端，执行以下命令切换到omm用户，若在集群外使用客户端，请切换到root用户：

sudo su - omm

步骤5 执行以下命令切换客户端目录，例如“**/opt/Bigdata/client**”：

cd /opt/Bigdata/client

步骤6 执行以下命令，更新客户端配置：

sh refreshConfig.sh *客户端安装目录客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/Bigdata/client /tmp/MRS-client/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 集群安装完成之后，执行以下命令切换到omm用户，若在集群外使用客户端，请切换到root用户。

sudo su - omm

步骤2 执行以下命令切换客户端目录，例如“**/opt/Bigdata/client**”。

cd /opt/Bigdata/client

步骤3 执行以下命令并按照提示输入MRS Manager有下载权限的用户名和密码（例如，用户名为admin，密码为创建集群时设置的密码），更新客户端配置。

sh autoRefreshConfig.sh

步骤4 命令执行后显示如下信息，其中XXX表示集群安装的组件名称，如需更新全部组件配置，单击“Enter”键，如需更新部分组件配置，请输入需要更新的组件名称，多个组件名称以逗号相隔。

```
Components "xxx" have been installed in the cluster. Please input the comma-separated names of the components for which you want to update client configurations. If you press Enter without inputting any component name, the client configurations of all components will be updated:
```

界面显示以下信息表示配置更新成功：

```
Succeed to refresh components client config.
```

界面显示以下信息表示用户名或者密码错误：

```
login manager failed,Incorrect username or password.
```

说明

- 该脚本会自动连接到集群并调用refreshConfig.sh脚本下载并刷新客户端配置文件。
- 客户端默认使用安装目录下文件Version中的“wsom=xxx”所配置的浮动IP刷新客户端配置，如需刷新为其他集群的配置文件，请执行本步骤前修改Version文件中“wsom=xxx”的值为对应集群的浮动IP地址。

----结束

全量更新主 Master 节点的原始客户端

场景描述

用户创建集群时，默认在集群所有节点的“/opt/client”目录安装保存了原始客户端。以下操作以“/opt/Bigdata/client”为例进行说明。

- MRS普通集群，在console页面提交作业时，会使用master节点上预置安装的客户端进行作业提交。
- 用户也可使用master节点上预置安装的客户端来连接服务端、查看任务结果或管理数据等

对集群安装补丁后，用户需要重新更新master节点上的客户端，才能保证继续使用内置客户端功能。

操作步骤

步骤1 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**，然后选择“服务管理”。

步骤2 单击“下载客户端”。

“客户端类型”选择“完整客户端”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤3 查询并登录主Master节点。

步骤4 在弹性云服务器，切换到**root**用户，并将安装包复制到目录“/opt”。

sudo su - root

cp /tmp/MRS-client/MRS_Services_Client.tar /opt

步骤5 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Services_Client.tar

步骤6 执行以下命令，校验文件包。

sha256sum -c MRS_Services_ClientConfig.tar.sha256

界面显示如下：

```
MRS_Services_ClientConfig.tar: OK
```

步骤7 执行以下命令，解压“MRS_Services_ClientConfig.tar”。

tar -xvf MRS_Services_ClientConfig.tar

步骤8 执行以下命令，移走原来老的客户端到/opt/Bigdata/client_bak目录下

mv /opt/Bigdata/client /opt/Bigdata/client_bak

步骤9 执行以下命令，安装客户端到新的目录，客户端路径必须为“/opt/Bigdata/client”。

sh /opt/MRS_Services_ClientConfig/install.sh /opt/Bigdata/client

查看安装输出信息，如有以下结果表示客户端安装成功：

```
Components client installation is complete.
```

步骤10 执行以下命令，修改/opt/Bigdata/client目录的所属用户和用户组。

chown omm:wheel /opt/Bigdata/client -R

步骤11 执行以下命令配置环境变量：

source /opt/Bigdata/client/bigdata_env

步骤12 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit*MRS集群用户*

例如, **kinit admin**

步骤13 执行组件的客户端命令。

例如，执行以下命令查看HDFS目录：

hdfs dfs -ls /

----结束

全量更新备 Master 节点的原始客户端

步骤1 参见步骤1~步骤3登录备Master节点，执行如下命令切换到omm用户。

sudo su - omm

步骤2 在备master节点上执行如下命令，从主master节点拷贝下载的客户端包。

scp omm@*master1节点IP地址***:/tmp/MRS-client/MRS_Services_Client.tar /tmp/MRS-client/**

说明

- 该命令以master1节点为主master节点为例。
- 目的路径以备master节点的/tmp/MRS-client/目录为例，请根据实际路径修改。

步骤3 参见步骤4~步骤13，更新备Master节点的客户端。

----结束

6.3 各组件客户端使用实践

6.3.1 使用 ClickHouse 客户端

ClickHouse是面向联机分析处理的列式数据库，支持SQL查询，且查询性能好，特别是基于大宽表的聚合分析查询性能非常优异，比其他分析型数据库速度快一个数量级。

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。在使用客户端前，需要先下载并更新客户端配置文件，确认Manager的主管理节点后才能使用客户端。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限，具体请参见**ClickHouse用户及权限管理**章节，为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行本步骤。

1. 如果是MRS 3.1.0版本集群，则需要先执行：**export CLICKHOUSE_SECURITY_ENABLED=true**
2. **kinit** *组件业务用户*

   例如，**kinit** clickhouseuser。

步骤5 执行ClickHouse组件的客户端命令。

执行**clickhouse -h**，查看ClickHouse组件命令帮助。

回显信息如下：

```
Use one of the following commands:
clickhouse local [args]
clickhouse client [args]
clickhouse benchmark [args]
clickhouse server [args]
clickhouse performance-test [args]
clickhouse extract-from-config [args]
clickhouse compressor [args]
clickhouse format [args]
clickhouse copier [args]
clickhouse obfuscator [args]
...
```

MRS 3.1.0及之后版本，使用**clickhouse client**命令连接ClickHouse服务端：

- 例如，当前集群未启用Kerberos认证，使用ssl安全方式登录：

  clickhouse client --host *ClickHouse的实例IP* **--user** *用户名* **--password** *密码* **--port** 9440 **--secure**
- 例如，当前集群已启用Kerberos认证，使用ssl安全方式登录。

Kerberos集群场景下没有默认用户，必须在Manager上创建用户，详细参考**ClickHouse用户及权限管理**。

使用kinit认证成功后，客户端登录时可以不携带--user和--password参数，即使用kinit认证的用户登录。

clickhouse client --host *ClickHouse的实例IP* **--port** 9440 **--secure**

相关参数使用说明如下表：

表 6-3 clickhouse client 命令行参数说明

参数名	参数说明
--host	服务端的host名称，默认是localhost。您可以选择使用ClickHouse实例所在节点主机名或者IP地址。 **说明** ClickHouse的实例IP地址可登录集群FusionInsight Manager，然后选择“集群 > 服务 > ClickHouse > 实例”，获取ClickHouseServer实例对应的业务IP地址。
--port	连接的端口。 ● 如果使用ssl安全连接则默认端口为9440，并且需要携带参数--secure。具体的端口值可通过ClickHouseServer实例配置搜索“tcp_port_secure”参数获取。 ● 如果使用非ssl安全连接则默认端口为9000，不需要携带参数--secure。具体的端口值可通过ClickHouseServer实例配置搜索“tcp_port”参数获取。
--user	用户名。 可以在Manager上创建该用户名并绑定对应的角色权限，具体可以参考**ClickHouse用户及权限管理**。 ● 如果当前集群已启用Kerberos认证，使用kinit认证成功后，客户端登录时可以不携带--user和--password参数，即使用kinit认证的用户登录。Kerberos集群场景下没有默认用户，必须在Manager上创建该用户名。 ● 如果当前集群未启用Kerberos认证，客户端登录时可以指定Manager上创建的用户和密码。不携带用户和密码参数时则默认使用**default**用户登录。
--password	密码。 默认值：空字符串。该参数和--user参数配套使用，可以在Manager上创建用户名时设置该密码。
--query	使用非交互模式查询。
--database	默认当前操作的数据库。默认值：服务端默认的配置（默认是default）。
--multiline	如果指定，允许多行语句查询（Enter仅代表换行，不代表查询语句完结）。
--multiquery	如果指定，允许处理用;号分隔的多个查询，只在非交互模式下生效。
--format	使用指定的默认格式输出结果。

参数名	参数说明
--vertical	如果指定，默认情况下使用垂直格式输出结果。在这种格式中，每个值都在单独的行上打印，适用显示宽表的场景。
--time	如果指定，非交互模式下会打印查询执行的时间到stderr中。
--stacktrace	如果指定，如果出现异常，会打印堆栈跟踪信息。
--config-file	配置文件的名称。
--secure	如果指定，将通过ssl安全模式连接到服务器。
--history_file	存放命令历史的文件的路径。
--param_<name>	带有参数的查询，并将值从客户端传递给服务器。具体用法详见https://clickhouse.tech/docs/zh/interfaces/cli/#cli-queries-with-parameters。

----结束

6.3.2 使用 Flink 客户端

本节提供使用Flink运行wordcount作业的操作指导。

前提条件

- MRS集群中已安装Flink组件。
- 集群正常运行，已安装集群客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

使用 Flink 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行如下命令初始化环境变量。

source /opt/hadoopclient/bigdata_env

步骤4 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 登录Manager，下载认证凭据。

   登录集群的Manager界面，具体请参见访问MRS Manager（MRS 2.x及之前版本），选择“系统设置 > 用户管理”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/hadoopclient/Flink/flink/conf”目录下。如果是在

集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。

4. 配置安全认证，在“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/hadoopclient/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test

5. 参考“组件操作指南 > 使用Flink > 参考 > 签发证书样例”章节生成“generate_keystore.sh”脚本并放置在Flink的客户端bin目录下，执行如下命令进行安全加固，请参考“组件操作指南 > 使用Flink > 安全加固 > 认证和加密”，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中关于SSL的值，针对MRS2.x及之前版本，安全集群默认没有开启外部SSL，用户如果需要启用外部SSL，请参考“组件操作指南 > 使用Flink > 安全加固”进行配置后再次运行该脚本即可。

   说明

   - generate_keystore.sh脚本无需手动生成。
   - 执行认证和加密后会将生成的flink.keystore、flink.truststore、security.cookie自动填充到“flink-conf.yaml”对应配置项中。

6. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/hadoopclient/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/hadoopclient/Flink/flink/conf/”目录下新建目录，例如ssl。

        cd /opt/hadoopclient/Flink/flink/conf/

        mkdir ssl

     ii. 移动flink.keystore和flink.truststore文件到“/opt/hadoopclient/Flink/flink/conf/ssl/”中。

        mv flink.keystore ssl/

        mv flink.truststore ssl/

     iii. 修改flink-conf.yaml文件中如下两个参数为相对路径。

```
security.ssl.internal.keystore: ssl/flink.keystore
security.ssl.internal.truststore: ssl/flink.truststore
```

步骤5 运行wordcount作业。

须知

用户在Flink提交作业或者运行作业时，应具有如下权限：

- 如果启用Ranger鉴权，当前用户必须属于hadoop组或者已在Ranger中为该用户添加“/flink”的读写权限。
- 如果停用Ranger鉴权，当前用户必须属于hadoop组。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。

    yarn-session.sh -nm "*session-name*"

    flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - 执行如下命令在Yarn上提交单个作业。

    flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。

      yarn-session.sh -nm "*session-name*"

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - flink.keystore和flink.truststore文件路径为相对路径时：
    - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/hadoopclient/Flink/flink/conf/”，则在“opt/hadoopclient/Flink/flink/conf/”目录下执行命令。

      yarn-session.sh -t ssl/ -nm "*session-name*"

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster -yt ssl/ /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

步骤6 作业提交成功后，客户端界面显示如下。

图 6-4 在 Yarn 上提交作业成功

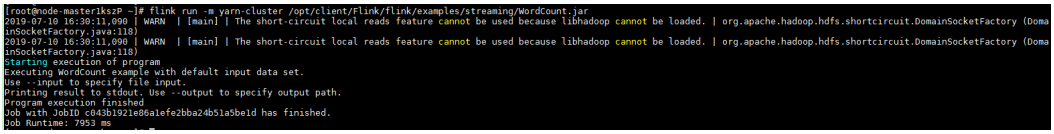
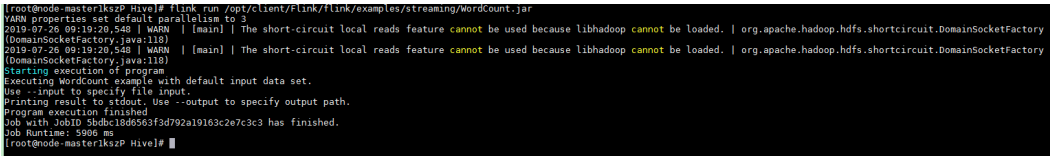


图 6-5 启动 session 成功

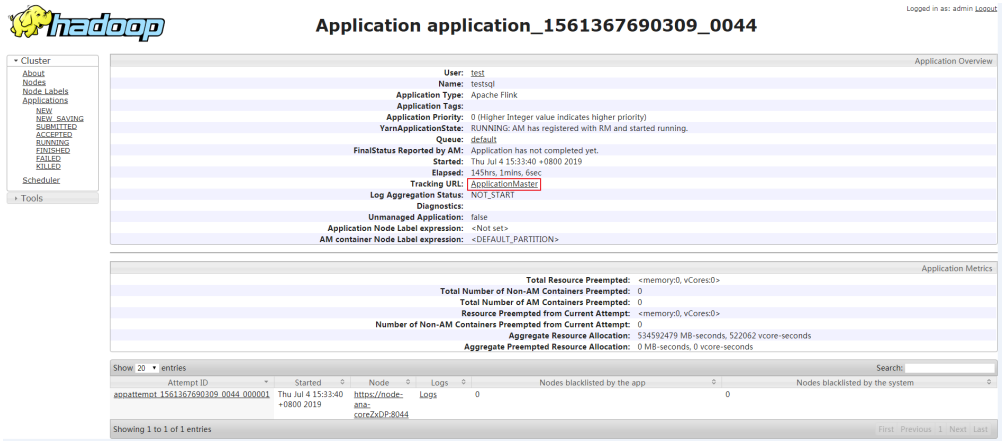
图 6-6 在 session 中提交作业成功



步骤7 使用运行用户进入Yarn服务的原生页面，具体操作参考“组件操作指南 > 使用Flink > 查看Flink作业”，找到对应作业的application，单击application名称，进入到作业详情页面。

- 若作业尚未结束，可单击“Tracking URL”链接进入到Flink的原生页面，查看作业的运行信息。
- 若作业已运行结束，对于在session中提交的作业，可以单击“Tracking URL”链接登录Flink原生页面查看作业信息。

图 6-7 application



----结束

使用 Flink 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行如下命令初始化环境变量。

source /opt/hadoopclient/bigdata_env

步骤4 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。

2. 登录Manager，下载认证凭据。

   登录集群的Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“系统 > 权限 > 用户”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。

3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/hadoopclient/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。

4. 将客户端安装节点的业务IP、Manager的浮动IP和Master节点IP添加到配置文件“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中的“jobmanager.web.access-control-allow-origin”和“jobmanager.web.allow-access-address”配置项中，IP地址之间使用英文逗号分隔。

   ```
   jobmanager.web.access-control-allow-origin: xx.xx.xxx.xxx,xx.xx.xxx.xxx,xx.xx.xxx.xxx
   jobmanager.web.allow-access-address: xx.xx.xxx.xxx,xx.xx.xxx.xxx,xx.xx.xxx.xxx
   ```

   说明

   - 客户端安装节点的业务IP获取方法：
     - 集群内节点：

       登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中当前的集群并单击集群名，进入集群信息页面。

       在“节点管理”中查看安装客户端所在的节点IP。

     - 集群外节点：安装客户端所在的弹性云服务器的IP。
   - Manager的浮动IP获取方法：
     - 登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中当前的集群并单击集群名，进入集群信息页面。

       在“节点管理”中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。

     - 远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示MRS Manager浮动IP地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到MRS Manager的浮动IP地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。

5. 配置安全认证，在“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/hadoopclient/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test

6. 参考“组件操作指南 > 使用Flink > 参考 > 签发证书样例”章节生成“generate_keystore.sh”脚本并放置在Flink的客户端bin目录下，执行如下命令进行安全加固，请参考“组件操作指南 > 使用Flink > 安全加固 > 认证和加密”，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中关于SSL的值。

   sh generate_keystore.sh <*password*>

说明

执行认证和加密后会在Flink客户端的“conf”目录下生成“flink.keystore”和“flink.truststore”文件，并且在客户端配置文件“flink-conf.yaml”中将以下配置项进行了默认赋值：

- 将配置项“security.ssl.keystore”设置为“flink.keystore”文件所在绝对路径。
- 将配置项“security.ssl.truststore”设置为“flink.truststore”文件所在的绝对路径。
- 将配置项“security.cookie”设置为“generate_keystore.sh”脚本自动生成的一串随机规则密码。
- 默认“flink-conf.yaml”中“security.ssl.encrypt.enabled: false”，“generate_keystore.sh”脚本将配置项“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值设置为调用“generate_keystore.sh”脚本时输入的密码。
- MRS 3.1.0及之后版本，如果需要使用密文时，设置“flink-conf.yaml”中“security.ssl.encrypt.enabled: true”，“generate_keystore.sh”脚本不会配置“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值，需要使用Manager明文加密API进行获取，执行**curl -k -i -u** *user name*:*password* **-X POST -HContent-type:application/json -d '{"plainText":"***password***"}' 'https://***x.x.x.x***:28443/web/api/v2/tools/encrypt'**

  其中*user name*:password分别为当前系统登录用户名和密码；**"plainText"**的password为调用“generate_keystore.sh”脚本时的密码；x.x.x.x为集群Manager的浮动IP。

7. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/hadoopclient/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/hadoopclient/Flink/flink/conf/”目录下新建目录，例如ssl。

     cd /opt/hadoopclient/Flink/flink/conf/

     mkdir ssl

     ii. 移动flink.keystore和flink.truststore文件到“/opt/hadoopclient/Flink/flink/conf/ssl/”中。

     mv flink.keystore ssl/

     mv flink.truststore ssl/

     iii. 修改flink-conf.yaml文件中如下两个参数为相对路径。

```
security.ssl.keystore: ssl/flink.keystore
security.ssl.truststore: ssl/flink.truststore
```

步骤5 运行wordcount作业。

须知

用户在Flink提交作业或者运行作业时，应具有如下权限：

- 如果启用Ranger鉴权，当前用户必须属于hadoop组或者已在Ranger中为该用户添加“/flink”的读写权限。
- 如果停用Ranger鉴权，当前用户必须属于hadoop组。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。

    yarn-session.sh -nm "*session-name***"**

    flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - 执行如下命令在Yarn上提交单个作业。

    flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。

      yarn-session.sh -nm "*session-name***"**

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - flink.keystore和flink.truststore文件路径为相对路径时：
    - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/hadoopclient/Flink/flink/conf/”，则在“opt/hadoopclient/Flink/flink/conf/”目录下执行命令。

      yarn-session.sh -t ssl/ -nm "*session-name***"**

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster -yt ssl/ /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

步骤6 作业提交成功后，客户端界面显示如下。

图 6-8 在 Yarn 上提交作业成功

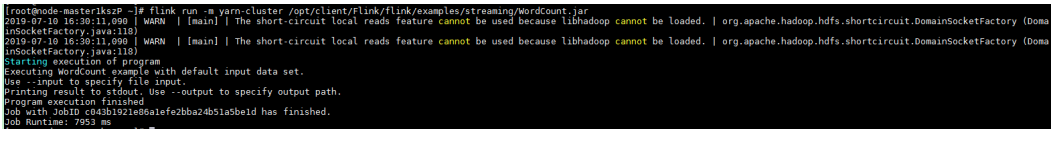


图 6-9 启动 session 成功

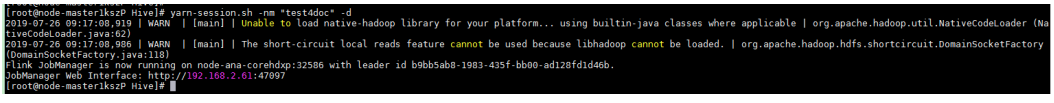
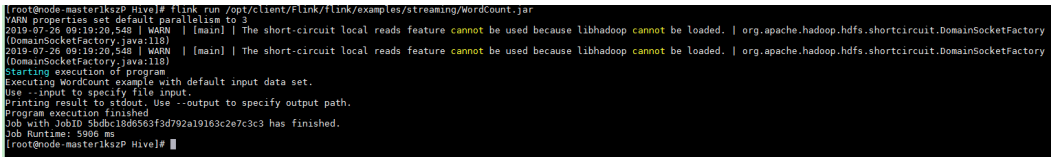


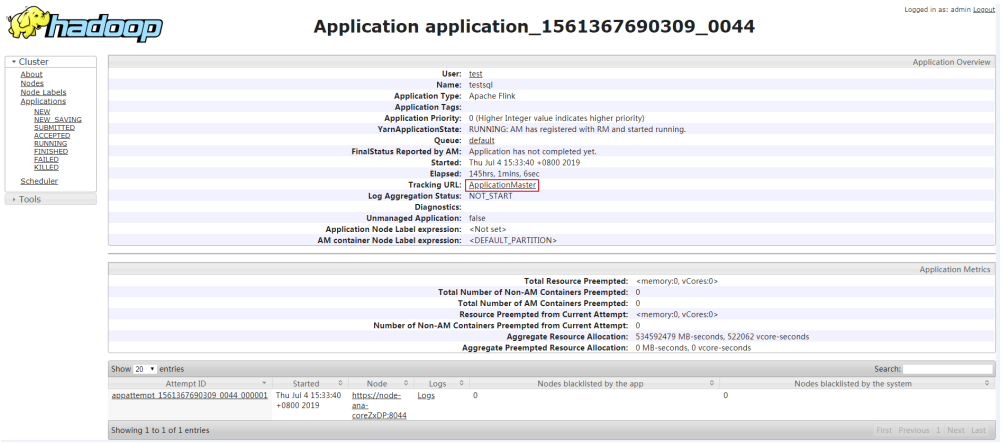
图 6-10 在 session 中提交作业成功



步骤7 使用运行用户进入Yarn服务的原生页面，具体操作参考“组件操作指南 > 使用Flink > 查看Flink作业”，找到对应作业的application，单击application名称，进入到作业详情页面

- 若作业尚未结束，可单击“Tracking URL”链接进入到Flink的原生页面，查看作业的运行信息。
- 若作业已运行结束，对于在session中提交的作业，可以单击“Tracking URL”链接登录Flink原生页面查看作业信息。

图 6-11 application



----结束

6.3.3 使用 Flume 客户端

操作场景

Flume支持将采集的日志信息导入到Kafka。

前提条件

- 已创建启用Kerberos认证的流集群。
- 已在日志生成节点安装Flume客户端，例如安装目录为“/opt/Flumeclient”，客户端安装请参见“组件操作指南 > 使用Flume > 安装Flume客户端”。以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 已配置网络，使日志生成节点与流集群互通。

使用 Flume 客户端（MRS 3.x 之前版本）

说明

普通集群不需要执行**步骤2-步骤6**。

步骤1 客户端安装。

步骤2 将Master1节点上的认证服务器配置文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/MRS_Current/1_X_KerberosClient/etc/kdc.conf”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤3 查看任一部署Flume角色节点的“业务IP”。

登录集群详情页面，选择“集群 > 组件管理 > Flume > 实例”，查看任一部署Flume角色节点的“业务IP”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤4 将此节点上的用户认证文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/MRS_XXX/install/FusionInsight-Flume-Flume组件版本号/flume/conf/flume.keytab”。

其中“XXX”为产品版本号，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤5 将此节点上的配置文件“jaas.conf”，复制到安装Flume客户端的节点，保存到Flume客户端中“conf”目录。

文件完整路径为“${BIGDATA_HOME}/MRS_Current/1_X_Flume/etc/jaas.conf”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤6 登录安装Flume客户端节点，切换到客户端安装目录，执行以下命令修改文件：

vi conf/jaas.conf

修改参数“keyTab”定义的用户认证文件完整路径即**步骤4**中保存用户认证文件的目录：“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”，然后保存并退出。

步骤7 执行以下命令，修改Flume客户端配置文件“flume-env.sh”：

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/flume-env.sh**

在“-XX:+UseCMSCompactAtFullCollection”后面，增加以下内容：

```
-Djava.security.krb5.conf=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/kdc.conf -Djava.security.auth.login.config=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/jaas.conf -Dzookeeper.request.timeout=120000
```

例如：**"-XX:+UseCMSCompactAtFullCollection -Djava.security.krb5.conf=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/kdc.conf -Djava.security.auth.login.config=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/jaas.conf -Dzookeeper.request.timeout=120000"**

请根据实际情况，修改“Flume客户端安装目录”，然后保存并退出。

步骤8 假设Flume客户端安装路径为“/opt/FlumeClient”，执行以下命令，重启Flume客户端：

cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/bin

./flume-manage.sh restart

步骤9 执行以下命令，修改Flume客户端配置文件“properties.properties”。

vi *Flume客户端安装目录*/fusioninsight-flume-*Flume组件版本号*/conf/properties.properties

将以下内容保存到文件中：

```
################################################################################################
client.sources = static_log_source
client.channels = static_log_channel
client.sinks = kafka_sink
################################################################################################
#LOG_TO_HDFS_ONLINE_1

client.sources.static_log_source.type = spooldir
client.sources.static_log_source.spoolDir = PATH
client.sources.static_log_source.fileSuffix = .COMPLETED
client.sources.static_log_source.ignorePattern = ^$
client.sources.static_log_source.trackerDir = PATH
client.sources.static_log_source.maxBlobLength = 16384
client.sources.static_log_source.batchSize = 51200
client.sources.static_log_source.inputCharset = UTF-8
client.sources.static_log_source.deserializer = LINE
client.sources.static_log_source.selector.type = replicating
client.sources.static_log_source.fileHeaderKey = file
client.sources.static_log_source.fileHeader = false
client.sources.static_log_source.basenameHeader = true
client.sources.static_log_source.basenameHeaderKey = basename
client.sources.static_log_source.deletePolicy = never

client.channels.static_log_channel.type = file
client.channels.static_log_channel.dataDirs = PATH
client.channels.static_log_channel.checkpointDir = PATH
client.channels.static_log_channel.maxFileSize = 2146435071
client.channels.static_log_channel.capacity = 1000000
client.channels.static_log_channel.transactionCapacity = 612000
client.channels.static_log_channel.minimumRequiredSpace = 524288000

client.sinks.kafka_sink.type = org.apache.flume.sink.kafka.KafkaSink
client.sinks.kafka_sink.kafka.topic = flume_test
client.sinks.kafka_sink.kafka.bootstrap.servers = XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口号
client.sinks.kafka_sink.flumeBatchSize = 1000
client.sinks.kafka_sink.kafka.producer.type = sync
client.sinks.kafka_sink.kafka.security.protocol  = SASL_PLAINTEXT
client.sinks.kafka_sink.kafka.kerberos.domain.name = hadoop.XXX.com
client.sinks.kafka_sink.requiredAcks = 0

client.sources.static_log_source.channels = static_log_channel
client.sinks.kafka_sink.channel = static_log_channel
```

请根据实际情况，修改以下参数，然后保存并退出。

- spoolDir
- trackerDir

- dataDirs
- checkpointDir
- topic

  如果kafka中该topic不存在，默认情况下会自动创建该topic。
- kafka.bootstrap.servers

  默认情况下，安全集群对应端口21007，普通集群对应端口9092。
- kafka.security.protocol

  安全集群请配置为SASL_PLAINTEXT，普通集群请配置为PLAINTEXT。
- “kafka.kerberos.domain.name”

  普通集群无需配置此参数。安全集群对应此参数的值为Kafka集群中“kerberos.domain.name”对应的值。

  具体可到Broker实例所在节点上查看“${BIGDATA_HOME}/MRS_Current/1_X_Broker/etc/server.properties”。

  其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤10 Flume客户端将自动加载“properties.properties”的内容。

当“spoolDir”生成新的日志文件，文件内容将发送到Kafka生产者，并支持Kafka消费者消费。

----结束

使用 Flume 客户端（MRS 3.x 及之后版本）

说明

普通集群不需要执行**步骤2**-**步骤6**。

步骤1 客户端安装。

步骤2 将Master1节点上的认证服务器配置文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_Current/1_X_KerberosClient/etc/kdc.conf”。其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤3 查看任一部署Flume角色节点的“业务IP”。

登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Flume > 实例”。查看任一部署Flume角色节点的“业务IP”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤4 将此节点上的用户认证文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_Porter_XXX/install/FusionInsight-Flume-Flume组件版本号/flume/conf/flume.keytab”。

其中“XXX”为产品版本号，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤5 将此节点上的配置文件“jaas.conf”，复制到安装Flume客户端的节点，保存到Flume客户端中“conf”目录。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_Current/1_X_Flume/etc/jaas.conf”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤6 登录安装Flume客户端节点，切换到客户端安装目录，执行以下命令修改文件：

vi conf/jaas.conf

修改参数“keyTab”定义的用户认证文件完整路径即**步骤4**中保存用户认证文件的目录：“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”，然后保存并退出。

步骤7 执行以下命令，修改Flume客户端配置文件“flume-env.sh”：

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/flume-env.sh**

在“-XX:+UseCMSCompactAtFullCollection”后面，增加以下内容：

```
-Djava.security.krb5.conf=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/kdc.conf -Djava.security.auth.login.config=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/jaas.conf -Dzookeeper.request.timeout=120000
```

例如：**"-XX:+UseCMSCompactAtFullCollection -Djava.security.krb5.conf=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/kdc.conf -Djava.security.auth.login.config=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/jaas.conf -Dzookeeper.request.timeout=120000"**

请根据实际情况，修改“Flume客户端安装目录”，然后保存并退出。

步骤8 假设Flume客户端安装路径为“/opt/FlumeClient”，执行以下命令，重启Flume客户端：

cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号***/bin**

./flume-manage.sh restart

步骤9 执行以下命令，修改Flume客户端配置文件“properties.properties”。

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/properties.properties**

将以下内容保存到文件中：

```
#################################################################################################
client.sources = static_log_source
client.channels = static_log_channel
client.sinks = kafka_sink
#################################################################################################
#LOG_TO_HDFS_ONLINE_1
```

```
client.sources.static_log_source.type = spooldir
client.sources.static_log_source.spoolDir = PATH
client.sources.static_log_source.fileSuffix = .COMPLETED
client.sources.static_log_source.ignorePattern = ^$
client.sources.static_log_source.trackerDir = PATH
client.sources.static_log_source.maxBlobLength = 16384
client.sources.static_log_source.batchSize = 51200
client.sources.static_log_source.inputCharset = UTF-8
client.sources.static_log_source.deserializer = LINE
client.sources.static_log_source.selector.type = replicating
client.sources.static_log_source.fileHeaderKey = file
client.sources.static_log_source.fileHeader = false
client.sources.static_log_source.basenameHeader = true
client.sources.static_log_source.basenameHeaderKey = basename
client.sources.static_log_source.deletePolicy = never

client.channels.static_log_channel.type = file
client.channels.static_log_channel.dataDirs = PATH
client.channels.static_log_channel.checkpointDir = PATH
client.channels.static_log_channel.maxFileSize = 2146435071
client.channels.static_log_channel.capacity = 1000000
client.channels.static_log_channel.transactionCapacity = 612000
client.channels.static_log_channel.minimumRequiredSpace = 524288000

client.sinks.kafka_sink.type = org.apache.flume.sink.kafka.KafkaSink
client.sinks.kafka_sink.kafka.topic = flume_test
client.sinks.kafka_sink.kafka.bootstrap.servers = XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口
号,XXX.XXX.XXX.XXX:kafka端口号
client.sinks.kafka_sink.flumeBatchSize = 1000
client.sinks.kafka_sink.kafka.producer.type = sync
client.sinks.kafka_sink.kafka.security.protocol  = SASL_PLAINTEXT
client.sinks.kafka_sink.kafka.kerberos.domain.name = hadoop.XXX.com
client.sinks.kafka_sink.requiredAcks = 0

client.sources.static_log_source.channels = static_log_channel
client.sinks.kafka_sink.channel = static_log_channel
```

请根据实际情况，修改以下参数，然后保存并退出。

- spoolDir
- trackerDir
- dataDirs
- checkpointDir
- topic

  如果kafka中该topic不存在，默认情况下会自动创建该topic。
- kafka.bootstrap.servers

  默认情况下，安全集群对应端口21007，普通集群对应端口9092。
- kafka.security.protocol

  安全集群请配置为SASL_PLAINTEXT，普通集群请配置为PLAINTEXT。
- “kafka.kerberos.domain.name”

  普通集群无需配置此参数。安全集群对应此参数的值为Kafka集群中“kerberos.domain.name”对应的值。

  具体可到Broker实例所在节点上查看“${BIGDATA_HOME}/FusionInsight_Current/1_X_Broker/etc/server.properties”。

  其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤10 Flume客户端将自动加载“properties.properties”的内容。

当“spoolDir”生成新的日志文件，文件内容将发送到Kafka生产者，并支持Kafka消费者消费。

----结束

6.3.4 使用 HBase 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用HBase客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。

  “机机”用户需要下载keytab文件，“人机”用户第一次登录时需修改密码。
- 非**root**用户使用HBase客户端，请确保该HBase客户端目录的属主为该用户，否则请参考如下命令修改属主。

  chown user:group -R *客户端安装目录***/HBase**

使用 Hbase 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令切换到客户端目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

步骤5 直接执行HBase组件的客户端命令。

hbase shell

----结束

使用 HBase 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令切换到客户端目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若安装了HBase多实例，在使用客户端连接具体HBase实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载HBase2实例变量：

source HBase2/component_env

步骤5 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

步骤6 直接执行HBase组件的客户端命令。

hbase shell

----结束

HBase 客户端常用命令

常用的HBase客户端命令如下表所示。更多命令可参考**http://hbase.apache.org/2.2/book.html**

表 6-4 HBase 客户端命令

命令	说明
create	创建一张表，例如**create 'test', 'f1', 'f2', 'f3'**。
disable	停止指定的表，例如**disable 'test'**。
enable	启动指定的表，例如**enable 'test'**。
alter	更改表结构。可以通过alter命令增加、修改、删除列族信息以及表相关的参数值，例如**alter 'test', {NAME => 'f3', METHOD => 'delete'}**。
describe	获取表的描述信息，例如**describe 'test'**。
drop	删除指定表。删除前表必须已经是停止状态，例如**drop 'test'**。
put	写入指定cell的value。Cell的定位由表、rowk、列组合起来唯一决定，例如**put 'test','r1','f1:c1','myvalue1'**。
get	获取行的值或者行的指定cell的值。例如**get 'test','r1'**。
scan	查询表数据。参数中指定表名和scanner，例如**scan 'test'**。

6.3.5 使用 HDFS 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用HDFS客户端。

前提条件

- 已安装客户端。

  例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）

使用 HDFS 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行HDFS Shell命令。例如：

hdfs dfs -ls /

----结束

HDFS 客户端常用命令

常用的HDFS客户端命令如下表所示。

更多命令可参考**https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/CommandsManual.html#User_Commands**

表 6-5 HDFS 客户端常用命令

命令	说明	样例
hdfs dfs -mkdir *文件夹名称*	创建文件夹	**hdfs dfs -mkdir /tmp/mydir**
hdfs dfs -ls *文件夹名称*	查看文件夹	**hdfs dfs -ls /tmp**
hdfs dfs -put *客户端节点上本地文件 HDFS指定路径*	上传本地文件到HDFS指定路径	**hdfs dfs -put /opt/test.txt /tmp** 上传客户端节点“/opt/test.txt”文件到HDFS的“/tmp”路径下
hdfs dfs -get *hdfs指定文件 客户端节点上指定路径*	下载HDFS文件到本地指定路径	**hdfs dfs -get /tmp/test.txt /opt/** 下载HDFS的“/tmp/test.txt”文件到客户端节点的“/opt”路径下

命令	说明	样例
hdfs dfs -rm -r -f *hdfs指定文件夹*	删除文件夹	**hdfs dfs -rm -r -f /tmp/mydir**
hdfs dfs -chmod *权限参数 文件目录*	为用户设置HDFS目录权限	**hdfs dfs -chmod 700 /tmp/test**

客户端常见使用问题

1. 当执行HDFS客户端命令时，客户端程序异常退出，报“java.lang.OutOfMemoryError”的错误。

   这个问题是由于HDFS客户端运行时的所需的内存超过了HDFS客户端设置的内存上限（默认为128MB）。可以通过修改“*<客户端安装路径>*/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改HDFS客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   CLIENT_GC_OPTS="-Xmx1G"
   ```

   在修改完后，使用如下命令刷新客户端配置，使之生效：

   source <*客户端安装路径*>/**bigdata_env**

2. 如何设置HDFS客户端运行时的日志级别?

   HDFS客户端运行时的日志是默认输出到Console控制台的，其级别默认是INFO级别。有的时候为了定位问题，需要开启DEBUG级别日志，可以通过导出一个环境变量来设置，命令如下：

   export HADOOP_ROOT_LOGGER=DEBUG,console

   在执行完上面命令后，再执行HDFS Shell命令时，即可打印出DEBUG级别日志。

   如果想恢复INFO级别日志，可执行如下命令：

   export HADOOP_ROOT_LOGGER=INFO,console

3. 如何彻底删除HDFS文件?

   由于HDFS的回收站机制，一般删除HDFS文件后，文件会移动到HDFS的回收站中。如果确认文件不再需要并且需要立马释放存储空间，可以继续清理对应的回收站目录（例如：hdfs://hacluster/user/*xxx*/.Trash/Current/*xxx*）。

6.3.6 使用 Hive 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Hive客户端。

前提条件

- 已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。

使用 Hive 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 根据集群认证模式，完成Hive客户端登录。

- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。

  kinit *组件业务用户*

  beeline

- 普通模式，则执行以下命令，登录Hive客户端，如果不指定组件业务用户，则会以当前操作系统用户登录。

  beeline -n *组件业务用户*

  说明

  进行beeline连接后，可以编写并提交HQL语句执行相关任务。如需执行Catalog客户端命令，需要先执行**!q**命令退出beeline环境。

步骤5 使用以下命令，执行HCatalog的客户端命令。

hcat -e *"cmd"*

其中*"cmd"*必须为Hive DDL语句，如**hcat -e "show tables"**。

说明

- 若要使用HCatalog客户端，必须从“组件管理”页面单击“下载客户端”，下载全部服务的客户端。Beeline客户端不受此限制。
- 由于权限模型不兼容，使用HCatalog客户端创建的表，在HiveServer客户端中不能访问，但可以使用WebHCat客户端访问。
- 在普通模式下使用HCatalog客户端，系统将以当前登录操作系统用户来执行DDL命令。
- 退出beeline客户端时请使用**!q**命令，不要使用“Ctrl + c”。否则会导致连接生成的临时文件无法删除，长期会累积产生大量的垃圾文件。
- 在使用beeline客户端时，如果需要在一行中输入多条语句，语句之间以“;”分隔，需要将“entireLineAsCommand”的值设置为“false”。

  设置方法：如果未启动beeline，则执行**beeline --entireLineAsCommand=false**命令；如果已启动beeline，则在beeline中执行**!set entireLineAsCommand false**命令。

  设置完成后，如果语句中含有不是表示语句结束的“;”，需要进行转义，例如**select concat_ws('\;', collect_set(col1)) from tbl**。

----结束

使用 Hive 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 MRS 3.X支持Hive多实例，若安装了Hive多实例，在使用客户端连接具体Hive实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载Hive2实例变量：

source Hive2/component_env

步骤5 根据集群认证模式，完成Hive客户端登录。

- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。

  kinit *组件业务用户*

  beeline

- 普通模式，则执行以下命令，登录Hive客户端，如果不指定组件业务用户，则会以当前操作系统用户登录。

  beeline -n *组件业务用户*

步骤6 使用以下命令，执行HCatalog的客户端命令。

hcat -e *"cmd"*

其中*"cmd"*必须为Hive DDL语句，如**hcat -e "show tables"**。

说明

- 若要使用HCatalog客户端，必须从服务页面选择“更多 > 下载客户端”，下载全部服务的客户端。Beeline客户端不受此限制。
- 由于权限模型不兼容，使用HCatalog客户端创建的表，在HiveServer客户端中不能访问，但可以使用WebHCat客户端访问。
- 在普通模式下使用HCatalog客户端，系统将以当前登录操作系统用户来执行DDL命令。
- 退出beeline客户端时请使用**!q**命令，不要使用“Ctrl + C”。否则会导致连接生成的临时文件无法删除，长期会累积产生大量的垃圾文件。
- 在使用beeline客户端时，如果需要在一行中输入多条语句，语句之间以“;”分隔，需要将“entireLineAsCommand”的值设置为“false”。

  设置方法：如果未启动beeline，则执行**beeline --entireLineAsCommand=false**命令；如果已启动beeline，则在beeline中执行**!set entireLineAsCommand false**命令。

  设置完成后，如果语句中含有不是表示语句结束的“;”，需要进行转义，例如**select concat_ws('\;', collect_set(col1)) from tbl**。

----结束

Hive 客户端常用命令

常用的Hive Beeline客户端命令如下表所示。

更多命令可参考**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients#HiveServer2Clients-BeelineCommands**。

表 6-6 Hive Beeline 客户端常用命令

命令	说明
set <key>=<value>	设置特定配置变量（键）的值。 **说明** 若变量名拼错，Beeline不会显示错误。
set	打印由用户或Hive覆盖的配置变量列表。
set -v	打印Hadoop和Hive的所有配置变量。
add FILE[S] <filepath> <filepath>*add JAR[S] <filepath> <filepath>*add ARCHIVE[S] <filepath> <filepath>*	将一个或多个文件、JAR文件或ARCHIVE文件添加至分布式缓存的资源列表中。
add FILE[S] <ivyurl> <ivyurl>* add JAR[S] <ivyurl> <ivyurl>* add ARCHIVE[S] <ivyurl> <ivyurl>*	使用“ivy://goup:module:version?query_string”格式的lvy URL，将一个或多个文件、JAR文件或ARCHIVE文件添加至分布式缓存的资源列表中。
list FILE[S]list JAR[S]list ARCHIVE[S]	列出已添加至分布式缓存中的资源。
list FILE[S] <filepath>*list JAR[S] <filepath>*list ARCHIVE[S] <filepath>*	检查给定的资源是否已添加至分布式缓存中。
delete FILE[S] <filepath>*delete JAR[S] <filepath>*delete ARCHIVE[S] <filepath>*	从分布式缓存中删除资源。
delete FILE[S] <ivyurl> <ivyurl>* delete JAR[S] <ivyurl> <ivyurl>* delete ARCHIVE[S] <ivyurl> <ivyurl>*	从分布式缓存中删除使用<ivyurl>添加的资源。
reload	使HiveServer2发现配置参数指定路径下JAR文件的变更“hive.reloadable.aux.jars.path”（无需重启HiveServer2）。更改操作包括添加、删除或更新JAR文件。
dfs <dfs command>	执行dfs命令。
<query string>	执行Hive查询，并将结果打印到标准输出。

6.3.7 使用 Impala 客户端

Impala是用于处理存储在Hadoop集群中的大量数据的MPP（大规模并行处理）SQL查询引擎。 它是一个用C++和Java编写的开源软件。 与其他Hadoop的SQL引擎相比，它拥有高性能和低延迟的特点。

背景信息

假定用户开发一个应用程序，用于管理企业中的使用A业务的用户信息，使用Impala客户端实现A业务操作流程如下：

普通表的操作：

- 创建用户信息表user_info。
- 在用户信息中新增用户的学历、职称信息。
- 根据用户编号查询用户姓名和地址。
- A业务结束后，删除用户信息表。

表 6-7 用户信息

编号	姓名	性别	年龄	地址
12005000201	A	男	19	A城市
12005000202	B	女	23	B城市
12005000203	C	男	26	C城市
12005000204	D	男	18	D城市
12005000205	E	女	21	E城市
12005000206	F	男	32	F城市
12005000207	G	女	29	G城市
12005000208	H	女	30	H城市
12005000209	I	男	26	I城市
12005000210	J	女	25	J城市

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 运行Impala客户端命令，实现A业务。

直接执行Impala组件的客户端命令：

impala-shell

说明

默认情况下，**impala-shell**尝试连接到localhost的21000端口上的Impala守护程序。如需连接到其他主机，请使用**-i <host:port>**选项，例如：impala-shell -i xxx.xxx.xxx.xxx:21000。要自动连接到特定的Impala数据库，请使用**-d <database>**选项。例如，如果您的所有Kudu表都位于数据库“impala_kudu”中，则**-d impala_kudu**可以使用此数据库。要退出Impala Shell，请使用**quit**命令。

内部表的操作：

1. 根据表6-7创建用户信息表user_info并添加相关数据。
   ```
   create table user_info(id string,name string,gender string,age int,addr string);
   insert into table user_info(id,name,gender,age,addr) values("12005000201","A","男",19,"A城市");
   ```
   ……（其他语句相同）
2. 在用户信息表user_info中新增用户的学历、职称信息。
   以增加编号为12005000201的用户的学历、职称信息为例，其他用户类似。
   ```
   alter table user_info add columns(education string,technical string);
   ```
3. 根据用户编号查询用户姓名和地址。
   以查询编号为12005000201的用户姓名和地址为例，其他用户类似。
   ```
   select name,addr from user_info where id='12005000201';
   ```
4. 删除用户信息表。
   ```
   drop table user_info;
   ```

外部分区表的操作：

创建外部分区表并导入数据

1. 创建外部表数据存储路径。
   - 安全模式（集群开启了Kerberos认证）：
     cd /opt/hadoopclient
     source bigdata_env
     kinit *hive*

     说明

     用户hive需要具有Hive管理员权限。

     impala-shell
     hdfs dfs -mkdir /hive
     hdfs dfs -mkdir /hive/user_info
   - 普通模式（集群关闭了Kerberos认证）：
     su - omm
     cd /opt/hadoopclient
     source bigdata_env

impala-shell

hdfs dfs -mkdir /hive

hdfs dfs -mkdir /hive/user_info

2. 建表。

```
create external table user_info(id string,name string,gender string,age int,addr string) partitioned by(year string) row format delimited fields terminated by ' ' lines terminated by '\n' stored as textfile location '/hive/user_info';
```

说明

fields terminated指明分隔的字符,如按空格分隔，' '。

lines terminated 指明分行的字符，如按换行分隔，'\n'。

/hive/user_info为数据文件的路径。

3. 导入数据。
   a. 使用insert语句插入数据。

```
insert into user_info partition(year="2018") values ("12005000201","A","男",19,"A城市");
```

   b. 使用load data命令导入文件数据。
      i. 根据表6-7数据创建文件。如，文件名为txt.log，以空格拆分字段，以换行符作为行分隔符。
      ii. 上传文件至hdfs。

hdfs dfs -put txt.log /tmp

      iii. 加载数据到表中。

load data inpath '/tmp/txt.log' into table user_info partition (year='2018');

4. 查询导入数据。

```
select * from user_info;
```

5. 删除用户信息表。

```
drop table user_info;
```

----结束

6.3.8 使用 Kafka 客户端

操作场景

用户可以在集群客户端完成Topic的创建、查询、删除等基本操作。

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

使用 Kafka 客户端（MRS 3.x 之前版本）

步骤1 进入ZooKeeper实例页面：

单击集群名称，登录集群详情页面，选择“组件管理 > ZooKeeper > 实例”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient/Kafka/kafka/bin”。

cd /opt/hadoopclient/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/hadoopclient/bigdata_env

步骤6 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Kafka用户*

步骤7 创建一个Topic：

sh kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份个数* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤8 执行以下命令，查询集群中的Topic信息：

sh kafka-topics.sh --list --zookeeper *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤9 删除**步骤7**中创建的Topic：

sh kafka-topics.sh --delete --topic *主题名称* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

输入 "y"，回车。

----结束

使用 Kafka 客户端（MRS 3.x 及之后版本）

步骤1 进入ZooKeeper实例页面：

登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 实例”。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient/Kafka/kafka/bin”。

cd /opt/hadoopclient/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/hadoopclient/bigdata_env

步骤6 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Kafka用户*

步骤7 登录FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 配置 > 全部配置”，搜索参数“clientPort”,记录“clientPort”的参数值。

步骤8 创建一个Topic：

sh kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份个数* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤9 执行以下命令，查询集群中的Topic信息：

sh kafka-topics.sh --list --zookeeper *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤10 删除步骤8中创建的Topic：

sh kafka-topics.sh --delete --topic *主题名称* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

----结束

6.3.9 使用 Kudu 客户端

Kudu是专为Apache Hadoop平台开发的列式存储管理器。Kudu具有Hadoop生态系统应用程序的共同技术特性：可水平扩展，并支持高可用性操作。

前提条件

已安装集群客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 运行Kudu命令行工具。

直接执行Kudu组件的命令行工具，查看帮助。

kudu -h

回显信息如下：

```
Usage: kudu <command> [<args>]

<command> can be one of the following:
        cluster   Operate on a Kudu cluster
       diagnose   Diagnostic tools for Kudu servers and clusters
             fs   Operate on a local Kudu filesystem
           hms    Operate on remote Hive Metastores
  local_replica   Operate on local tablet replicas via the local filesystem
```

```
      master   Operate on a Kudu Master
         pbc   Operate on PBC (protobuf container) files
        perf   Measure the performance of a Kudu cluster
remote_replica   Operate on remote tablet replicas on a Kudu Tablet Server
       table   Operate on Kudu tables
      tablet   Operate on remote Kudu tablets
        test   Various test actions
     tserver   Operate on a Kudu Tablet Server
         wal   Operate on WAL (write-ahead log) files
```

说明

kudu命令行工具不提供DDL、DML等操作，但提供针对cluster、master、tserver、fs、table等的细化查询功能。

常用操作：

- 查看当前集群下有哪些表。

  kudu table list *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051*

- 查询Kudu服务KuduMaster实例的配置信息。

  kudu master get_flags *KuduMaster实例IP:7051*

- 查询表的schema。

  kudu table describe *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051 tablename*

- 删除表。

  kudu table delete *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051 tablename*

  说明

  KuduMaster实例IP获取方式：在集群详情页面，选择“组件管理 > Kudu > 实例”，获取角色KuduMaster的IP地址。

----结束

6.3.10 使用 Oozie 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Oozie客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。

使用 Oozie 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录，该操作的客户端目录只是举例，请根据实际安装目录修改。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 判断集群认证模式。

- 安全模式，执行以下命令进行用户认证。*exampleUser*为提交任务的用户名。

  kinit *exampleUser*

- 普通模式，执行**步骤5**。

步骤5 配置Hue。

1. spark2x环境配置（如果不涉及spark2x任务，可以跳过此步骤）：

   hdfs dfs -put /opt/client/Spark2x/spark/jars/*.jar /user/oozie/share/lib/spark2x/

   当HDFS目录“/user/oozie/share”中的Jar包发生变化时，需要重启Oozie服务。

2. 上传Oozie配置文件以及Jar包至HDFS：

   hdfs dfs -mkdir /user/*exampleUser*

   hdfs dfs -put -f /opt/client/Oozie/oozie-client-*/examples /user/*exampleUser*/

   说明

   - *exampleUser*为提交任务的用户名。
   - 在提交任务的用户和非job.properties文件均无变更的前提下，客户端安装目录/Oozie/oozie-client-*/examples目录一经上传HDFS，后续可重复使用，无需多次提交。
   - 解决Spark和Yarn关于jetty的jar冲突。

     hdfs dfs -rm -f /user/oozie/share/lib/spark/jetty-all-9.2.22.v20170606.jar
   - 普通模式下，上传过程如果遇到“Permission denied”的问题，可执行以下命令进行处理。

     su - omm

     source /opt/client/bigdata_env

     hdfs dfs -chmod -R 777 /user/oozie

     exit

----结束

6.3.11 使用 Storm 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Storm客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/hadoopclient”。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）

操作步骤

步骤1 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若安装了Storm多实例，在使用Storm命令提交拓扑时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，Storm-2实例：

source Storm-2/component_env

步骤5 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤6 执行命令进行客户端操作。

例如执行以下命令：

- **cql**
- **storm**

说明

同一个storm客户端不能同时连接安全和非安全的ZooKeeper。

----结束

6.3.12 使用 Yarn 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Yarn客户端。

前提条件

- 已安装客户端。

  例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。普通模式不需要下载keytab文件及修改密码操作。

使用 Yarn 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行Yarn命令。例如：

yarn application -list

----结束

客户端常见使用问题

1. 当执行Yarn客户端命令时，客户端程序异常退出，报“java.lang.OutOfMemoryError”的错误。

   这个问题是由于Yarn客户端运行时的所需的内存超过了Yarn客户端设置的内存上限（默认为128MB）。对于MRS 3.x后续版本集群，可以通过修改“<客户端安装路径>/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改Yarn客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   export CLIENT_GC_OPTS="-Xmx1G"
   ```

   对于MRS 3.x之前版本集群，可以通过修改“<客户端安装路径>/HDFS/component_env”中的“GC_OPTS_YARN”来修改Yarn客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   export GC_OPTS_YARN="-Xmx1G"
   ```

   在修改完后，使用如下命令刷新客户端配置，使之生效：

   source <*客户端安装路径*>/**bigdata_env**

2. 如何设置Yarn客户端运行时的日志级别?

   Yarn客户端运行时的日志是默认输出到Console控制台的，其级别默认是INFO级别。有的时候为了定位问题，需要开启DEBUG级别日志，可以通过导出一个环境变量来设置，命令如下：

   export YARN_ROOT_LOGGER=DEBUG,console

   在执行完上面命令后，再执行Yarn Shell命令时，即可打印出DEBUG级别日志。

   如果想恢复INFO级别日志，可执行如下命令：

   export YARN_ROOT_LOGGER=INFO,console

7 配置存算分离

7.1 存算分离简介

MRS支持在大数据存储容量大、计算资源需要弹性扩展的场景下，用户将数据存储在OBS服务中，使用MRS集群仅做数据计算处理的存算分离模式。

说明

大数据存算分离场景，请务必使用OBS并行文件系统，使用普通对象桶会对集群性能产生较大影响。

存算分离功能使用流程：

1. 配置存算分离集群。

   请选择如下其中一种配置即可（推荐使用委托方式）。

   - 通过为MRS集群绑定ECS委托方式访问OBS，避免了AK/SK直接暴露在配置文件中的风险，具体请参考**配置存算分离集群（委托方式）**。
   - 在MRS集群中配置AK/SK，AK/SK会明文暴露在配置文件中，请谨慎使用，具体请参考**配置存算分离集群（AKSK方式）**。

2. 使用存算分离集群。

   各个组件使用存算分离的具体操作请参考如下内容。

   - **Flink对接OBS文件系统**
   - **Flume对接OBS文件系统**
   - **HDFS客户端对接OBS文件系统**
   - **Hive对接OBS文件系统**
   - **MapReduce对接OBS文件系统**
   - **Spark2x对接OBS文件系统**
   - **Sqoop对接外部存储系统**

7.2 配置存算分离集群（委托方式）

MRS支持用户将数据存储在OBS服务中，使用MRS集群仅做数据计算处理的存算模式。MRS通过IAM服务的“委托”机制进行简单配置， 实现使用ECS自动获取的临时AK/SK访问OBS。避免了AK/SK直接暴露在配置文件中的风险。

通过绑定委托，ECS或BMS云服务将有权限来管理您的部分资源，请根据实际业务场景需求确认是否需要配置委托。

MRS提供如下访问OBS的配置方式，请选择其中一种配置即可（推荐使用委托方式）：

- 通过为MRS集群绑定ECS委托方式访问OBS，避免了AK/SK直接暴露在配置文件中的风险，具体请参考本章节。
- 在MRS集群中配置AK/SK，AK/SK会明文暴露在配置文件中，请谨慎使用，具体请参考**配置存算分离集群（AKSK方式）**。

集群的Hadoop、Hive、Spark、Presto、Flink组件支持该功能。

步骤一：创建具有访问 OBS 权限的 ECS 委托

说明

- MRS在IAM的委托列表中预置了**MRS_ECS_DEFAULT_AGENCY**委托，可在集群创建过程中可以选择该委托，该委托拥有对象存储服务的OBS OperateAccess权限和在集群所在区域拥有CES FullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。同时请勿在IAM修改**MRS_ECS_DEFAULT_AGENCY**委托。
- 如需使用预置的委托，请跳过创建委托步骤。如需使用自定义委托，请参考如下步骤进行创建委托（创建或修改委托需要用户具有Security Administrator权限）。

1. 登录管理控制台。
2. 在服务列表中选择“管理与监管 > 统一身份认证服务”。
3. 选择“委托 > 创建委托”。
4. 设置“委托名称”。例如：mrs_ecs_obs。
5. “委托类型”选择“云服务”，在“云服务”中选择“弹性云服务器ECS 裸金属服务器BMS”，授权ECS或BMS调用OBS服务。
6. “持续时间”选择“永久”并单击“下一步”。
7. 在弹出授权页面的搜索框内，搜索“OBS OperateAccess”策略，勾选“OBS OperateAccess”策略。
8. 单击“下一步”，选择权限范围方案，默认选择“所有资源”，单击“展开其他方案”，选择“全局服务资源”。
9. 在弹出的提示框中单击“知道了”，开始授权。界面提示“授权成功。”，单击“完成”，委托成功创建。

步骤二：创建存算分离集群

配置存算分离支持在新建集群中配置委托实现，也可以通过为已有集群绑定委托实现。本示例以开启Kerberos认证的集群为例介绍。

新创建存算分离集群：

1. 登录MRS服务控制台。
2. 单击“创建集群”，进入“创建集群”页面。
3. 在集群页面，选择“自定义创建”页签。
4. 在“自定义创建”页签，填写“软件配置”参数。
   - 区域：请根据需要选择。
   - 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。

- 集群版本：请选择集群版本。
- 集群类型：选择“分析集群”或“混合集群”并勾选所有组件。
- 元数据：选择“本地元数据”。

5. 单击“下一步”，并配置硬件相关参数。
   - 可用区：默认即可。
   - 虚拟私有云：默认即可。
   - 子网：默认即可。
   - 安全组：默认即可。
   - 弹性公网IP：默认即可。
   - 企业项目：默认即可。
   - 集群节点：请根据自身需求选择节点规格和数量。
6. 单击“下一步”，并配置相关参数。
   - Kerberos认证：默认开启，请根据自身需要选择。
   - 用户名：默认为“admin”，用于登录集群管理页面。
   - 密码：设置admin用户密码。
   - 确认密码：再次输入设置的admin用户密码。
   - 登录方式：选择登录ECS节点的登录方式，本例选择密码方式。
   - 用户名：默认为“root”，用于远程登录ECS机器。
   - 密码：设置root用户密码。
   - 确认密码：再次输入设置的root用户密码。
7. 本例以配置委托为例介绍，其他参数暂不配置，如需配置请参考**高级配置（可选）**。

   委托：选择**步骤一：创建具有访问OBS权限的ECS委托**所创建的委托或MRS在IAM服务中预置的委托MRS_ECS_DEFAULT_AGENCY。
8. 通信安全授权请勾选“确认授权”，详细信息请参见**授权安全通信**。
9. 单击“立即”，等待集群创建成功。

   当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

为已有集群配置存算分离功能：

1. 登录MRS控制台，在导航栏选择“集群列表 > 现有集群”。
2. 单击集群名称，进入集群详情页面。
3. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。
4. 在集群详情页的“概览”页签，单击委托右侧的“管理委托”选择需要绑定的委托并单击“确定”进行绑定，或单击“新建委托”进入IAM控制台进行创建后再在此处进行绑定。

步骤三：创建 OBS 文件系统用于存放数据

说明

大数据存算分离场景，请务必使用OBS并行文件系统，使用普通对象桶会对集群性能产生较大影响。

1. 登录OBS控制台。
2. 单击“并行文件系统 > 创建并行文件系统”。
3. 填写文件系统名称，例如“mrs-word001”。
   其他参数请根据需要填写。
4. 单击“立即创建”。
5. 在OBS控制台并行文件系统列表中，单击文件系统名称进入详情页面。
6. 在左侧导航栏选择“文件”，新建program、input文件夹。
   - program：请上传程序包到该文件夹。
   - input：请上传输入数据到该文件夹。

步骤四：访问 OBS 文件系统

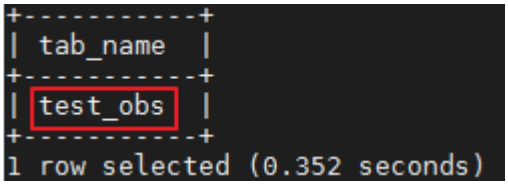
1. 用root用户登录集群Master节点，具体请参见**登录集群节点**。
2. 配置环境变量。
   MRS 3.x之前版本请执行：**source /opt/client/bigdata_env**
   MRS 3.x及之后版本请执行：**source /opt/Bigdata/client/bigdata_env**
3. 验证Hadoop访问OBS。
   a. 查看文件系统mrs-word001下面的文件列表。
      hadoop fs -ls obs://mrs-word001/
   b. 返回文件列表即表示访问OBS成功。

      图 7-1 Hadoop 验证返回文件列表

      ```
      Found 2 items
      drwxrwxrwx   - root root          0 2019-12-21 11:04 obs://mrs-word001/input
      drwxrwxrwx   - root root          0 2019-12-21 11:04 obs://mrs-word001/program
      ```

4. 验证Hive访问OBS。
   a. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建Hive表的权限，具体请参见**创建角色**配置拥有对应权限的角色，参考**创建用户**创建用户并为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行此命令。
      kinit MRS集群用户
      例如，kinit hiveuser
   b. 执行Hive组件的客户端命令。
      beeline
   c. 在beeline中直接使用obs的目录进行访问。例如，执行如下命令创建Hive表并指定数据存储在mrs-word001文件系统的test_obs目录中。
      create table test_obs(a int, b string) row format delimited fields terminated by "," stored as textfile location "obs://mrs-word001/test_obs";
   d. 执行如下命令查询所有表，返回结果中存在表test_obs，即表示访问OBS成功。
      show tables;

图 7-2 Hive 验证返回已创建的表名



e. 使用“Ctrl + C”退出hive beeline。

5. 验证Spark访问OBS。

a. 执行Spark组件的客户端命令。

spark-beeline

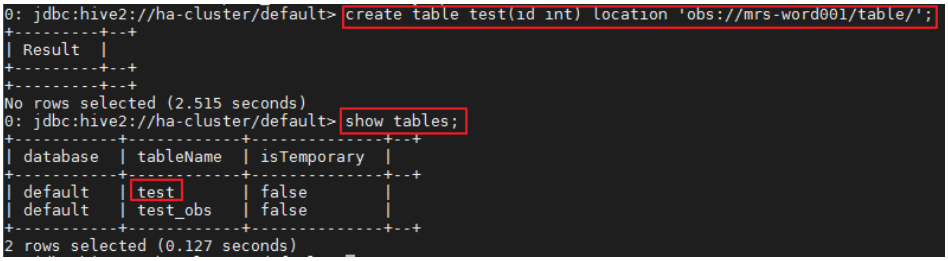
b. 在spark-beeline中访问OBS，例如在obs://mrs-word001/table/目录中创建表test。

create table test(id int) location 'obs://mrs-word001/table/';

c. 执行如下命令查询所有表，返回结果中存在表test，即表示访问OBS成功。

show tables;

图 7-3 Spark 验证返回已创建的表名



d. 使用“Ctrl + C”退出退出spark beeline。

6. 验证Presto访问OBS。

- 未开启Kerberos认证的普通集群

i. 执行如下命令连接客户端。

presto_cli.sh

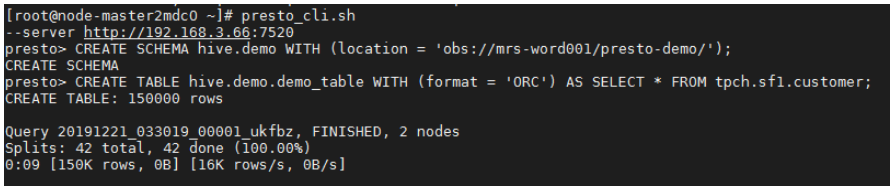
ii. 在Presto客户端中执行语句创建schema，指定location为OBS路径，例如：

CREATE SCHEMA hive.demo01 WITH (location = 'obs://mrs-word001/presto-demo002/');

iii. 在该schema中建表，该表的数据即存储在OBS文件系统内，例如：

CREATE TABLE hive.demo.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

图 7-4 普通集群 Presto 验证返回结果



iv. 执行**exit**退出客户端。

- 开启Kerberos认证的安全集群

i. 登录MRS Manager创建一个拥有“Hive Admin Privilege”权限的角色，例如prestorole，创建角色请参考**创建角色**。

ii. 创建一个属于“Presto”和“Hive”组的用户，同时为该用户绑定**6.i**中创建的角色，例如presto001，创建用户请参考**创建用户**。

iii. 认证当前用户。

kinit presto001

iv. 下载用户凭证。

1) 针对MRS 3.x之前版本集群，在MRS Manager页面，选择“系统设置 > 用户管理 ”，单击新增用户所在行的“更多 > 下载认证凭据”。

图 7-5 下载 Presto 用户认证凭据

2) 针对MRS 3.x及之后版本，在FusionInsight Manager页面，选择“系统 > 权限 > 用户”，单击新增用户所在行的“更多 > 下载认证凭据”。

v. 解压下载的用户凭证文件，得到“krb5.conf”和“user.keytab”两个文件并放入客户端目录，例如“/opt/Bigdata/client/Presto/”。

vi. 执行如下命令获取用户principal。

klist -kt /opt/Bigdata/client/Presto/user.keytab

vii. 启用Kerberos认证的集群，执行以下命令连接本集群的Presto Server。

presto_cli.sh --krb5-config-path {krb5.conf文件路径} --krb5-principal {用户principal} --krb5-keytab-path {user.keytab文件路径} --user {presto用户名}

- krb5.conf文件路径：请替换为**6.v**中设置的文件存放路径，例如“/opt/Bigdata/client/Presto/krb5.conf”
- user.keytab文件路径：请替换为**6.v**中设置的文件存放路径，例如“/opt/Bigdata/client/Presto/user.keytab”
- 用户principal：请替换为**6.vi**中返回的结果
- presto用户名：请替换为**6.ii**中创建的用户名，例如“presto001”

例如：presto_cli.sh --krb5-config-path /opt/Bigdata/client/Presto/krb5.conf --krb5-principal prest001@xxx_xxx_xxx_xxx.COM --krb5-keytab-path /opt/Bigdata/client/Presto/user.keytab --user presto001

viii. 在Presto客户端中执行语句创建schema，指定location为OBS路径，例如：

CREATE SCHEMA hive.demo01 WITH (location = 'obs://mrs-word001/presto-demo002/');

ix. 在该schema中建表，该表的数据即存储在OBS文件系统内，例如：

CREATE TABLE hive.demo01.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

图 7-6 安全集群 Presto 验证返回结果

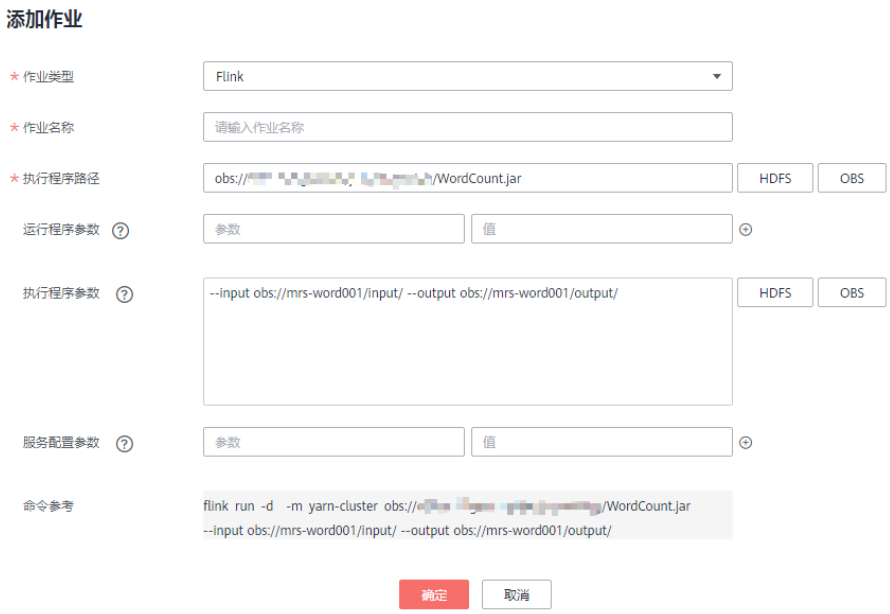
x. 执行**exit**退出客户端。

7. 验证Flink访问OBS。

a. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。

b. 用户同步完成后，在集群详情页选择“作业管理 > 添加”提交Flink作业，在“执行程序参数”中按照“--input <作业输入路径> --output <作业输出路径>”格式填写，其中作业输入路径选择OBS，输出路径请手动输入一个不存在的目录，例如obs://mrs-word001/output/，如图7-7所示。

图 7-7 添加 Flink 作业



c. 在OBS控制台，进入提交作业时选择的输出路径，即可查看到输出目录已经自动创建并存放着作业执行结果，表示访问OBS成功。

图 7-8 Flink 作业执行结果

相关参考

如需对访问OBS的权限进行控制，请参考**配置MRS多用户访问OBS细粒度权限**。

7.3 配置存算分离集群（AKSK 方式）

MRS支持使用**obs://**的方式对接OBS服务，当前主要支持的组件为Hadoop、Hive、Spark、Presto、Flink。其中HBase组件使用**obs://**的方式对接OBS服务暂不支持。

MRS提供如下访问OBS的配置方式，请选择其中一种配置即可（推荐使用委托方式）：

- 通过为MRS集群绑定ECS委托方式访问OBS，避免了AK/SK直接暴露在配置文件中的风险，具体请参考**配置存算分离集群（委托方式）**。
- 在MRS集群中配置AK/SK，AK/SK会明文暴露在配置文件中，请谨慎使用，具体请参考本章节。

说明

- 为了提高数据写入性能，可以修改对应服务的配置参数fs.obs.buffer.dir的值为数据盘目录。
- 大数据存算分离场景，请务必使用OBS并行文件系统，使用普通对象桶会对集群性能产生较大影响。

Hadoop 访问 OBS

- 在MRS客户端的HDFS目录($client_home/ HDFS/hadoop/etc/hadoop)中修改core-site.xml文件，增加如下内容。

```
<property>
    <name>fs.obs.access.key</name>
    <value>ak</value>
</property>
<property>
    <name>fs.obs.secret.key</name>
    <value>sk</value>
</property>
<property>
    <name>fs.obs.endpoint</name>
    <value>obs endpoint</value>
</property>
```

须知

在文件中设置AK/SK会明文暴露在配置文件中，请谨慎使用。

添加配置后无需手动添加AK/SK、endpoint就可以直接访问OBS上的数据。例如执行如下命令查看文件系统obs-test下面的文件夹test_obs_orc的文件列表。

hadoop fs -ls "obs://obs-test/test_obs_orc"

- 每次在命令行中手动添加AK/SK、endpoint访问OBS上的数据。

  hadoop fs -Dfs.obs.endpoint=xxx -Dfs.obs.access.key=xx -Dfs.obs.secret.key=xx -ls "obs://obs-test/ test_obs_orc"

Hive 访问 OBS

步骤1 登录服务配置页面。

- 针对MRS 3.x之前版本，登录集群详情页面，选择“组件管理 > Hive > 服务配置”。
- 针对MRS 3.x及之后版本，登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Hive > 配置”。

步骤2 将“基础配置”切换为“全部配置”。

步骤3 搜索“fs.obs.access.key”和“fs.obs.secret.key”参数，并分别配置为OBS的AK和SK。

若当前集群中搜索不到如上两个参数，请在左侧导航选择“Hive > 自定义”，在自定义参数“core.site.customized.configs”中增加如上两个参数。

步骤4 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Hive服务。

步骤5 在beeline中直接使用obs的目录进行访问。例如，执行如下命令创建Hive表并指定数据存储在test-bucket文件系统的test_obs目录中。

create table test_obs(a int, b string) row format delimited fields terminated by "," stored as textfile location "obs://test-bucket/test_obs";

----结束

Spark 访问 OBS

说明

由于SparkSQL依赖Hive，所以在Spark上配置OBS时，需要同时修改**Hive访问OBS**的OBS配置。

- spark-beeline和spark-sql

  可以通过在shell中增加如下OBS的属性实现访问OBS。

  ```
  set fs.obs.endpoint=xxx
  set fs.obs.access.key=xxx
  set fs.obs.secret.key=xxx
  ```

- spark-beeline

  spark-beeline也可以通过在Manager中配置服务参数实现访问OBS。操作如下：

  a. 登录服务配置页面。

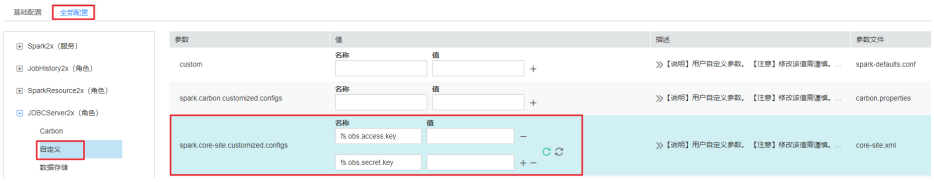
  - 针对MRS 3.x之前版本，登录集群详情页面，选择“组件管理 > Spark > 服务配置”。
  - 针对MRS 3.x及之后版本，登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Spark2x > 配置”。

b. 将“基础配置”切换为“全部配置”。

c. 选择“JDBCServer > OBS”配置fs.obs.access.key、fs.obs.secret.key参数。

若当前集群中没有如上两个参数，请在左侧导航选择“JDBCServer > 自定义”，在自定义参数“spark.core-site.customized.configs”中增加如上两个参数。

图 7-9 自定义添加 OBS 参数



d. 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Spark服务。

e. 在spark-beeline中访问OBS，例如访问obs://obs-demo-input/table/目录：

create table test(id int) location 'obs://obs-demo-input/table/';

- spark-sql和spark-submit

  spark-sql也可以通过修改core-site.xml配置文件实现访问OBS。

  使用spark-sql和使用spark-submit提交任务访问OBS时，配置文件修改方法一致。

  修改MRS客户端中Spark配置文件夹（$client_home/Spark/spark/conf）中的core-site.xml，增加如下内容：

```
<property>
    <name>fs.obs.access.key</name>
    <value>ak</value>
</property>
<property>
    <name>fs.obs.secret.key</name>
    <value>sk</value>
</property>
<property>
    <name>fs.obs.endpoint</name>
    <value>obs endpoint</value>
</property>
```

Presto 访问 OBS

步骤1 登录集群详情页面，选择“组件管理 > Presto > 服务配置”。

步骤2 将“基础配置”切换为“全部配置”。

步骤3 搜索并配置如下参数。

- fs.obs.access.key配置为用户AK
- fs.obs.secret.key配置为用户SK

若当前集群中搜索不到如上两个参数，请在左侧导航选择“Presto > Hive”，在自定义参数“core.site.customized.configs”中增加如上两个参数。

步骤4 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Presto服务。

步骤5 选择“组件管理 > Hive > 服务配置”。

步骤6 将“基础配置”切换为“全部配置”。

步骤7 搜索并配置如下参数。

- fs.obs.access.key配置为用户AK
- fs.obs.secret.key配置为用户SK

步骤8 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Hive服务。

步骤9 在Presto客户端中执行语句创建schema，指定location为OBS路径，例如：

CREATE SCHEMA hive.demo WITH (location = 'obs://obs-demo/presto-demo/');

步骤10 在该schema中建表，该表的数据即存储在OBS文件系统内，例如：

CREATE TABLE hive.demo.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

----结束

Flink 访问 OBS

在MRS客户端的Flink配置文件“客户端安装路径/Flink/flink/conf/flink-conf.yaml”中，增加如下内容。

```
fs.obs.access.key：ak
fs.obs.secret.key: sk
fs.obs.endpoint: obs endpoint
```

须知

在文件中设置AK/SK会明文暴露在配置文件中，请谨慎使用。

添加配置后无需手动添加AK/SK、endpoint就可以直接访问OBS上的数据。

7.4 使用存算分离集群

7.4.1 Flink 对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 使用安装客户端的用户登录Flink客户端安装节点。

步骤2 执行如下命令初始化环境变量。

source ${client_home}/bigdata_env

步骤3 需要配置好Flink客户端。具体配置参考**安装客户端（3.x及之后版本）**。

步骤4 如果是安全集群，使用以下命令进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

步骤5 在Flink命令行显式添加要访问的OBS文件系统。

./bin/flink run --class com.xxx.bigdata.flink.examples.FlinkProcessingTimeAPIMain ./config/FlinkCheckpointJavaExample.jar --chkPath obs://*OBS并行文件系统名称*

----结束

说明

由于Flink作业是On Yarn运行，在配置Flink对接OBS文件系统之前需要确保Yarn对接OBS文件系统功能是正常的。

7.4.2 Flume 对接 OBS 文件系统

本章节适用于MRS 3.x及之后的版本。

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 配置委托。

1. 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”。
2. 单击集群名称，进入集群详情页面。
3. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。
4. 单击委托右侧的“管理委托”，选择需要绑定的委托并单击“确定”进行绑定。

步骤2 创建OBS文件系统用于存放数据。

1. 登录OBS控制台。
2. 单击“并行文件系统”进入并行文件系统页面，单击“创建并行文件系统”。
3. 填写文件系统名称，例如“esdk-c-test-pfs1”，其他参数请根据需要填写。单击“立即创建”等待创建完成。
4. 在OBS控制台并行文件系统列表中，单击已新建的文件系统名称进入详情页面。
5. 在左侧导航栏选择“文件 > 新建文件夹”新建“testFlumeOutput”文件夹。

步骤3 准备properties.properties文件并将上传至“/opt/flumeInput”目录。

1. 在本地准备“properties.properties”文件，文件内容如下：

```
# source
server.sources = r1
# channels
server.channels = c1
# sink
server.sinks = obs_sink
# ----- define net source -----
server.sources.r1.type = seq
server.sources.r1.spooldir = /opt/flumeInput
# ---- define OBS sink ----
server.sinks.obs_sink.type = hdfs
server.sinks.obs_sink.hdfs.path = obs://esdk-c-test-pfs1/testFlumeOutput
server.sinks.obs_sink.hdfs.filePrefix = %[localhost]
server.sinks.obs_sink.hdfs.useLocalTimeStamp = true
# set file size to trigger roll
server.sinks.obs_sink.hdfs.rollSize = 0
server.sinks.obs_sink.hdfs.rollCount = 0
server.sinks.obs_sink.hdfs.rollInterval = 5
#server.sinks.obs_sink.hdfs.threadsPoolSize = 30
server.sinks.obs_sink.hdfs.fileType = DataStream
server.sinks.obs_sink.hdfs.writeFormat = Text
server.sinks.obs_sink.hdfs.fileCloseByEndEvent = false
```

```
# define channel
server.channels.c1.type = memory
server.channels.c1.capacity = 1000
# transaction size
server.channels.c1.transactionCapacity = 1000
server.channels.c1.byteCapacity = 800000
server.channels.c1.byteCapacityBufferPercentage = 20
server.channels.c1.keep-alive = 60
server.sources.r1.channels = c1
server.sinks.obs_sink.channel = c1
```

说明

参数“server.sinks.obs_sink.hdfs.path”中的值为**步骤2**中新建的OBS文件系统。

2. 使用**root**用户登录安装Flume客户端的节点。
3. 新建“/opt/flumeInput”目录，并在该目录下新建一个内容自定义的txt文件。
4. 登录FusionInsight Manager。
5. 选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置”，在参数“flume.config.file”的“值”中单击“上传文件”，上传**步骤3.1**准备的“properties.properties”文件，单击“保存”。

步骤4 在OBS系统中查看结果。

1. 登录OBS控制台。
2. 单击“并行文件系统”，进入**步骤2**中创建的并行文件系统中的文件夹查看结果。

----结束

7.4.3 HDFS 客户端对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 以客户端安装用户登录安装了HDFS客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd ${client_home}

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 在hdfs命令行显式添加要访问的OBS文件系统。

例如：

- 使用以下命令访问OBS文件系统。
  hdfs dfs -ls obs://*OBS并行文件系统名称|路径*
- 使用以下命令上传客户端节点“/opt/test.txt”文件到HDFS的“/tmp”路径下。
  hdfs dfs -put /opt/test.txt /tmp

----结束

说明

OBS文件系统打印大量日志可能导致读写性能受影响，可通过调整OBS客户端日志级别优化，日志调整方式如下：

cd ${client_home}/HDFS/hadoop/etc/hadoop

vi log4j.properties

在文件中添加OBS日志级别配置

log4j.logger.org.apache.hadoop.fs.obs=WARN

log4j.logger.com.obs=WARN

```
[root@node-master1AuKK hadoop]# tail -4 log4j.properties
log4j.logger.org.apache.commons.beanutils=WARN

log4j.logger.org.apache.hadoop.fs.obs=WARN
log4j.logger.com.obs=WARN
[root@node-master1AuKK hadoop]#
```

7.4.4 Hive 对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

建表时指定 Location 为 OBS 路径

步骤1 使用安装客户端用户登录客户端安装节点。

步骤2 执行如下命令初始化环境变量。

source ${client_home}/bigdata_env

步骤3 如果是安全集群，执行以下命令进行用户认证（该用户需要具有Hive操作的权限），如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Hive组件操作用户*

步骤4 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 配置 > 全部配置”。

在左侧的导航列表中选择“Hive > 自定义”。在自定义配置项中，给参数“hdfs.site.customized.configs”添加配置项“dfs.namenode.acls.enabled”，设置值为“false”。

步骤5 单击“保存”，保存配置。单击“概览”，选择“更多 > 重启服务”，输入当前用户密码，单击“确定”，并勾选“同时重启上层服务。”，单击“确定”，重启Hive服务。

步骤6 进入beeline客户端，在创建表时指定Location为OBS文件系统路径。

beeline

例如，创建一个表“test”，该表的Location为“obs://OBS并行文件系统名称/user/hive/warehouse/”：

create table test(name string) location "obs://*OBS并行文件系统名称/user/hive/warehouse/***";**

说明

需要添加组件操作用户到Ranger策略中的URL策略，URL填写对象在obs上的完整路径。权限选择Read, Write 权限，其他权限不涉及URL策略。

----结束

指定创建的 Hive 表默认 Location 为 OBS 路径

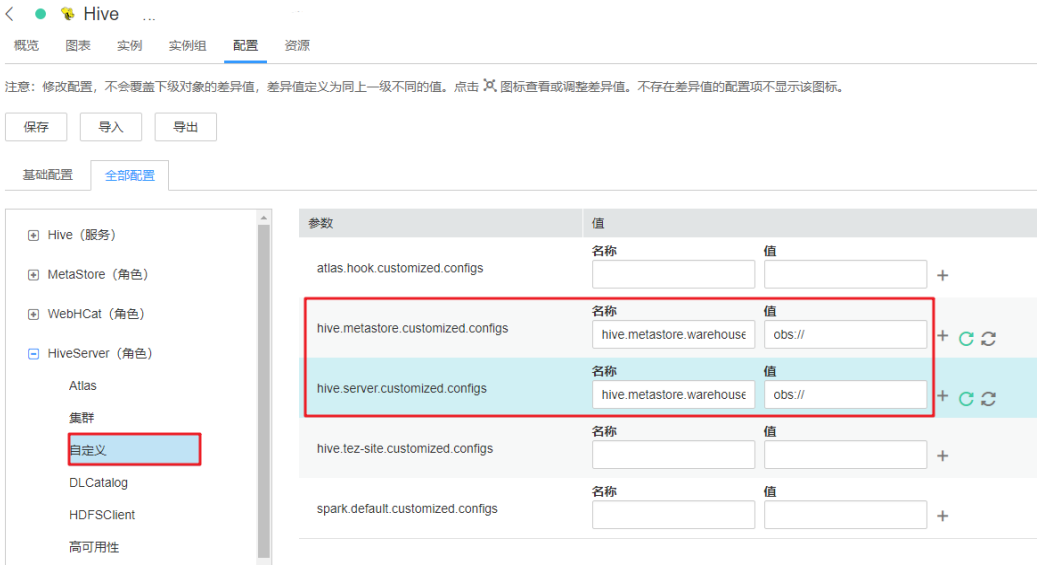
步骤1 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 配置 > 全部配置”。

步骤2 在左侧的导航列表中选择“MetaStore > 自定义”。在自定义配置项中，给参数“hive.metastore.customized.configs”添加配置项“hive.metastore.warehouse.dir”，设置值为OBS路径。

图 7-10 hive.metastore.warehouse.dir 配置

步骤3 在左侧的导航列表中选择“HiveServer > 自定义”。在自定义配置项中，给参数“hive.metastore.customized.configs”和“hive.server.customized.configs”添加配置项“hive.metastore.warehouse.dir”，设置值为OBS路径。

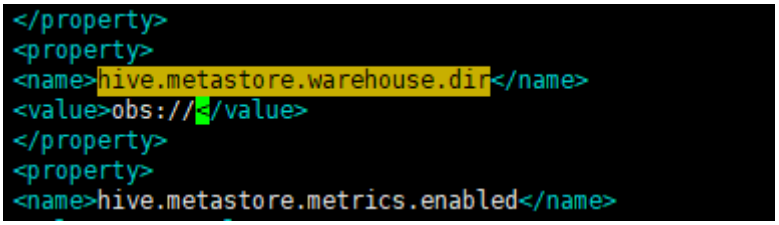
图 7-11 hive.metastore.warehouse.dir 配置



步骤4 保存并重启Hive服务。

步骤5 更新客户端配置文件。

1. 执行以下命令修改客户端Hive配置文件目录下的“hivemetastore-site.xml”。
   vim /opt/Bigdata/client/Hive/config/hivemetastore-site.xml
2. 将“hive.metastore.warehouse.dir”的值修改为对应的OBS路径。



步骤6 进入beeline客户端，创建表并确认Location为OBS路径。

beeline

create table test(name string);

desc formatted test;

说明

如果当前数据库Location已指向HDFS，那么在当前数据库下建表（不指定Location）默认也指向当前HDFS。如需修改默认建表策略可以修改数据库的Location重新指向OBS。操作如下：

1. 执行以下命令查看数据库Location。

   show create database *obs_test*;

   ```
   +----------------------------------------------------+
   |                   createdb_stmt                    |
   +----------------------------------------------------+
   | CREATE DATABASE `obs_test`                         |
   | LOCATION                                           |
   |   'hdfs://hacluster/user/hive/warehouse/obs_test.db'  |
   +----------------------------------------------------+
   3 rows selected (0.038 seconds)
   ```

2. 执行以下命令修改数据库Location。

   alter database *obs_test* **set location** *'obs://test1231/'*

   执行命令**show create database** *obs_test*，查看数据库Location已经指向OBS。

   ```
   INFO  : Concurrency mode is disabled, not creating
   +----------------------------+
   |        createdb_stmt       |
   +----------------------------+
   | CREATE DATABASE `obs_test` |
   | LOCATION                   |
   |   'obs://test1231/'        |
   +----------------------------+
   3 rows selected (0.063 seconds)
   ```

3. 执行以下命令修改表的Location。

   alter table *user_info* **set location** *'obs://test1231/'*

   如果表已有业务数据，需要同步迁移原数据文件至修改后的Location地址。

----结束

7.4.5 MapReduce 对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 登录MRS管理控制台，单击集群名称进入集群详情页面。

步骤2 选择“组件管理 > Mapreduce”，进入Mapreduce服务“全部配置”页面，在左侧的导航列表中选择“Mapreduce > 自定义”。在自定义配置项中，给参数文件“core-site.xml”添加配置项“mapreduce.jobhistory.always-scan-user-dir”，设置值为“true”。

步骤3 保存配置，并重启Mapreduce服务。

----结束

7.4.6 Spark2x 对接 OBS 文件系统

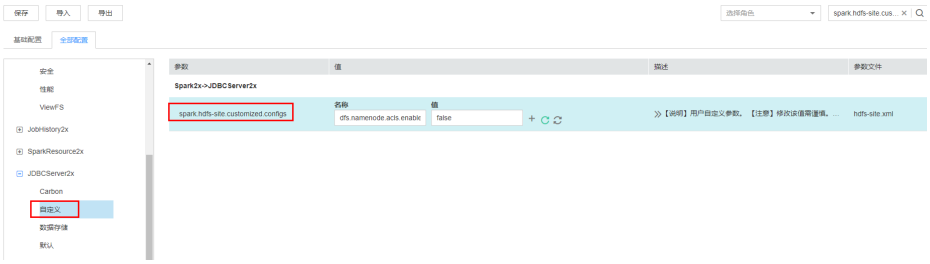
MRS集群支持Spark2x在集群安装完成后对接OBS文件系统。

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

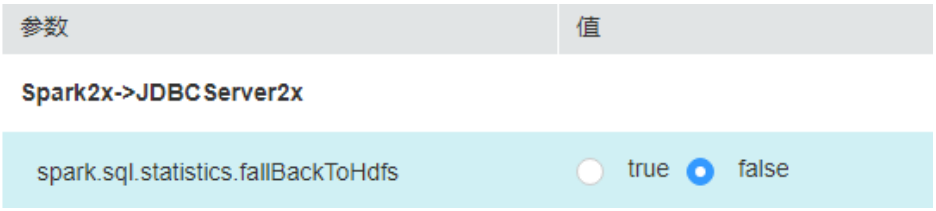
集群安装后使用 spark beeline

步骤1 登录FusionInsight Manager，选择“集群 > 服务 > Spark2x > 配置 > 全部配置”。

在左侧的导航列表中选择“JDBCServer2x > 自定义”。在参数“spark.hdfs-site.customized.configs”中添加配置项“dfs.namenode.acls.enabled”，值为“false”。



步骤2 在搜索框中搜索参数“spark.sql.statistics.fallBackToHdfs”，修改该参数值为“false”。



步骤3 保存配置并重启JDBCServer2x实例。

步骤4 使用安装客户端用户登录客户端安装节点。

步骤5 配置环境变量。

source ${client_home}/bigdata_env

步骤6 如果是安全集群，使用以下命令用户进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

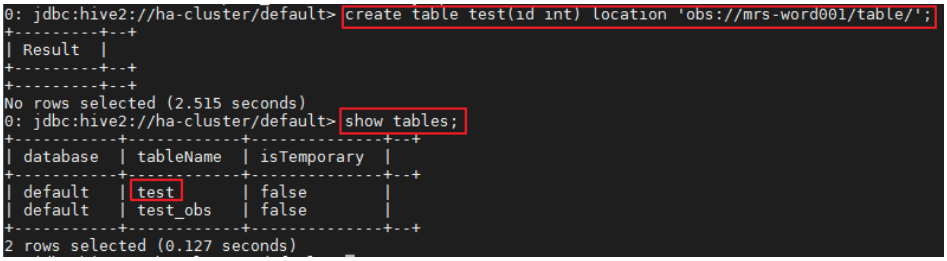
步骤7 在spark-beeline中访问OBS，例如在“obs://mrs-word001/table/”目录中创建表“test”。

create table test(id int) location '*obs://mrs-word001/table/*';

步骤8 执行如下命令查询所有表，返回结果中存在表test，即表示访问OBS成功。

show tables;

图 7-12 Spark2x 验证返回已创建的表名



步骤9 使用“Ctrl + C”退出spark beeline。

----结束

集群安装后使用 spark sql

步骤1 使用安装客户端用户登录客户端安装节点。

步骤2 配置环境变量。

source ${client_home}/bigdata_env

步骤3 修改配置文件：

vim ${client_home}/Spark2x/spark/conf/hdfs-site.xml

```
<property>
<name>dfs.namenode.acls.enabled</name>
<value>false</value>
</property>
```

步骤4 如果是安全集群，使用以下命令用户进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

步骤5 在spark-sql中访问OBS，例如在“obs://mrs-word001/table/”目录中创建表“test”。

步骤6 进入spark bin目录：**cd ${client_home}/Spark2x/spark/bin**，执行**./spark-sql**登录spark-sql命令行。

步骤7 在spark-sql命令行执行以下命令：

create table test(id int) location '*obs://mrs-word001/table/***';**

步骤8 执行语句**show tables;**查看表是否存在。

步骤9 执行**exit;**退出spark-sql命令行。

说明

OBS文件系统打印大量日志可能导致读写性能受影响，可通过调整OBS客户端日志级别优化，日志调整方式如下：

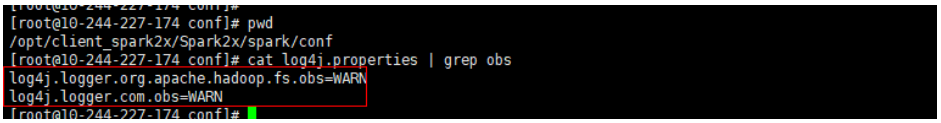
cd ${client_home}/Spark2x/spark/conf

vi log4j.properties

在文件中添加OBS日志级别配置

log4j.logger.org.apache.hadoop.fs.obs=WARN

log4j.logger.com.obs=WARN



----结束

7.4.7 Sqoop 对接外部存储系统

sqoop export (HDFS 到 MySQL)

步骤1 登录客户端所在节点。

步骤2 执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤3 使用sqoop命令操作sqoop客户端。

sqoop export --connect jdbc:mysql://10.100.231.134:3306/test --username root --password xxxxxx --table component13 -export-dir hdfs://hacluster/user/hive/warehouse/component_test3 --fields-terminated-by ',' -m 1

表 7-1 参数说明

参数	说明
-direct	快速模式，利用了数据库的导入工具，如MySQL的mysqlimport，可以比jdbc连接的方式更为高效的将数据导入到关系数据库中。
-export-dir <dir>	存放数据的HDFS的源目录。
-m或-num-mappers <n>	启动n个map来并行导入数据，默认是4个，该值请勿高于集群的最大Map数。
-table <table-name>	要导入的目的关系数据库表。
-update-key <col-name>	后面接条件列名，通过该参数可以将关系数据库中已经存在的数据进行更新操作，类似于关系数据库中的update操作。

参数	说明
-update-mode <mode>	更新模式，有两个值updateonly和默认的allowinsert，该参数只能在关系数据表里不存在要导入的记录时才能使用，比如要导入的hdfs中有一条id=1的记录，如果在表里已经有一条记录id=2，那么更新会失败。
-input-null-string <null-string>	可选参数，如果没有指定，则字符串null将被使用。
-input-null-non-string <null-string>	可选参数，如果没有指定，则字符串null将被使用。
-staging-table <staging-table-name>	创建一个与导入目标表同样数据结构的表，将所有数据先存放在该表中，然后由该表通过一次事务将结果写入到目标表中。 该参数是用来保证在数据导入关系数据库表的过程中的事务安全性，因为在导入的过程中可能会有多个事务，那么一个事务失败会影响到其它事务，比如导入的数据会出现错误或出现重复的记录等等情况，那么通过该参数可以避免这种情况。
-clear-staging-table	如果该staging-table非空，则通过该参数可以在运行导入前清除staging-table里的数据。

----结束

sqoop import（MySQL 到 Hive 表）

步骤1 登录客户端所在节点。

步骤2 执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤3 使用sqoop命令操作sqoop客户端。

sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password xxxxxx --table component --hive-import --hive-table component_test2 --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile

表 7-2 参数说明

参数	说明
-append	将数据追加到hdfs中已经存在的dataset中。使用该参数，sqoop将把数据先导入到一个临时目录中，然后重新给文件命名到一个正式的目录中，以避免和该目录中已存在的文件重名。
-as-avrodatafile	将数据导入到一个Avro数据文件中。
-as-sequencefile	将数据导入到一个sequence文件中。

参数	说明
-as-textfile	将数据导入到一个普通文本文件中，生成该文本文件后，可以在hive中通过sql语句查询出结果。
-boundary-query <statement>	边界查询，在导入前先通过SQL查询得到一个结果集，然后导入的数据就是该结果集内的数据，格式如：**-boundary-query 'select id,creationdate from person where id = 3'**，表示导入的数据为id=3的记录，或者**select min(<split-by>), max(<split-by>) from <table name>**。 注意：查询的字段中不能有数据类型为字符串的字段，否则会报错：java.sql.SQLException: Invalid value for getLong()。
-columns<col,col,col…>	指定要导入的字段值，格式如：-columns id,username
-direct	快速模式，利用了数据库的导入工具，如MySQL的mysqlimport，可以比jdbc连接的方式更为高效的将数据导入到关系数据库中。
-direct-split-size	在使用上面direct直接导入的基础上，对导入的流按字节数分块，特别是使用直连模式从PostgreSQL导入数据时，可以将一个到达设定大小的文件分为几个独立的文件。
-inline-lob-limit	设定大对象数据类型的最大值。
-m或-num-mappers	启动n个map来并行导入数据，默认是4个，该值请勿高于集群的最大Map数。
-query，-e<statement>	从查询结果中导入数据，该参数使用时必须指定-target-dir、-hive-table，在查询语句中一定要有where条件且在where条件中需要包含$CONDITIONS。 示例：-query 'select * from person where $CONDITIONS ' -target-dir /user/hive/warehouse/person -hive-table person
-split-by<column-name>	表的列名，用来切分工作单元，一般后面跟主键ID。
-table <table-name>	关系数据库表名，数据从该表中获取。
-target-dir <dir>	指定hdfs路径。
-warehouse-dir <dir>	与**-target-dir**不能同时使用，指定数据导入的存放目录，适用于导入hdfs，不适合导入hive目录。
-where	从关系数据库导入数据时的查询条件，示例：-where 'id = 2'
-z,-compress	压缩参数，默认数据不压缩，通过该参数可以使用gzip压缩算法对数据进行压缩，适用于SequenceFile，text文本文件，和Avro文件。
–compression-codec	Hadoop压缩编码，默认为gzip。

参数	说明
–null-string <null-string>	替换null字符串，如果没有指定，则字符串null将被使用。
–null-non-string<null-string>	替换非String的null字符串，如果没有指定，则字符串null将被使用。
-check-column (col)	增量导入参数，用来作为判断的列名，如id。
-incremental (mode) append 或lastmodified	增量导入参数。 append：追加，比如对大于last-value指定的值之后的记录进行追加导入。 lastmodified：最后的修改时间，追加last-value指定的日期之后的记录。
-last-value (value)	增量导入参数，指定自从上次导入后列的最大值（大于该指定的值），也可以自己设定某一值。

----结束

Sqoop 使用样例

- sqoop import（MySQL到HDFS）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* --query 'SELECT * FROM component where $CONDITIONS and component_id ="MRS 1.0_002"' --target-dir /tmp/component_test --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile

- sqoop export （obs到MySQL）

  sqoop export --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* --table component14 -export-dir obs://obs-file-bucket/xx/part-m-00000 --fields-terminated-by ',' -m 1

- sqoop import（MySQL到obs）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* --table component --target-dir obs://obs-file-bucket/xx --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile

- sqoop import（MySQL到Hive外obs表）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* --table component --hive-import --hive-table component_test01 --fields-terminated-by "," -m 1 --as-textfile

8 访问 MRS 集群上托管的开源组件 Web 页面

8.1 开源组件 Web 站点

场景介绍

MRS集群默认在集群的Master节点或Core节点创建并托管了不同组件的Web站点，用户可以通过这些Web站点查看组件相关信息。

访问开源组件Web站点步骤：

1. 配置访问方式。

   MRS提供如下访问开源组件Web站点的方式：

   - **通过弹性公网IP访问**：推荐使用该方式，为集群绑定弹性公网IP，简便易操作。
   - **通过Windows弹性云服务器访问**：需要创建单独的ECS并进行相关配置。
   - **创建连接MRS集群的SSH隧道并配置浏览器**：当用户和MRS集群处于不同的网络中时可以使用该方式访问。

2. 访问站点。请参考**表8-1**的地址进行访问。

Web 站点一览

说明

对于开启Kerberos认证的集群，admin用户不具备各组件的管理权限，如需正常访问各组件的Web UI界面，请提前添加具有对应组件管理权限的用户。

表 8-1 开源组件 Web 站点地址

集群类型	站点类型	站点地址
全部类型	MRS Manager	• 适用于所有版本集群 https://Manager浮动IP地址:28443/web **说明** 1. 确保本地机器与MRS集群网络互通。 2. 远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示MRS Manager浮动IP地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到MRS Manager的浮动IP地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。 • MRS 3.x之前版本集群 https://<弹性公网IP>:9022/mrsmanager?locale=zh-cn 具体请参见**访问MRS Manager（MRS 2.x及之前版本）**。 • MRS 3.x及以后版本请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。
分析集群	HDFS NameNode	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > HDFS > NameNode WebUI > NameNode (主)” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > HDFS > NameNode WebUI > NameNode (*主机名称*，主)”
	HBase HMaster	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > HBase > HMaster WebUI > HMaster (主)” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > HBase > HMaster WebUI > HMaster (*主机名称*，主)”

集群类型	站点类型	站点地址
	MapReduce JobHistoryServer	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Mapreduce > JobHistoryServer WebUI > JobHistoryServer” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Mapreduce > JobHistoryServer WebUI > JobHistoryServer (*主机名称*，主)”
	YARN ResourceManager	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Yarn > ResourceManager WebUI > ResourceManager (主)” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Yarn > ResourceManager WebUI > ResourceManager (*主机名称*，主)”
	Spark JobHistory	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Spark > Spark WebUI > JobHistory” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Spark2x > Spark2x WebUI > JobHistory2x (*主机名称*)”
	Hue	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Hue > Hue WebUI > Hue (主)” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Hue > Hue WebUI > Hue (*主机名称*，主)” Loader页面是基于开放源代码Sqoop WebUI的图形化数据迁移管理工具，由Hue WebUI承载。
	Tez	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Tez > Tez WebUI > TezUI” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Tez > Tez WebUI > TezUI (*主机名称*)”

集群类型	站点类型	站点地址
	Presto	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Presto > Presto WebUI > Coordinator (主)” ● 在Manager页面选择“集群 > 服务 > Presto > Coordinator WebUI > Coordinator(Coordinator)”
	Ranger	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Ranger > Ranger WebUI > RangerAdmin (主)” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Ranger > Ranger WebUI > RangerAdmin”
流处理集群	Storm	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Storm > Storm WebUI > UI” ● 在Manager页面选择“集群 > 服务 > Storm > Storm WebUI > UI (*主机名称*)”

8.2 开源组件端口列表

HBase 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
hbase.master.port	16000	HMaster RPC端口。该端口用于HBase客户端连接到HMaster。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口	端口说明
hbase.master.info.port	16010	HMaster HTTPS端口。该端口用于远程Web客户端连接到HMaster UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.regionserver.port	16020	RS (RegoinServer) RPC端口。该端口用于HBase客户端连接到RegionServer。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.regionserver.info.port	16030	Region server HTTPS端口。该端口用于远程Web客户端连接到RegionServer UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.thrift.info.port	9095	Thrift Server的Thrift Server侦听端口。 该端口用于： 客户端链接时使用该端口侦听。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.regionserver.thrift.port	9090	RegionServer的Thrift Server侦听端口 。 该端口用于： 客户端链接RegionServer时使用该端口侦听。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.rest.info.port	8085	RegionServer RESTServer原生web界面的端口
-	21309	RegionServer RESTServer的REST端口

HDFS 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
dfs.namenode.rpc.port	• 9820（MRS 3.x之前版本） • 8020（MRS 3.x及之后版本）	NameNode RPC 端口。 该端口用于： 1. HDFS客户端与Namenode间的通信。 2. Datanode与NameNode之间的连接。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
dfs.namenode.http.port	9870	HDFS HTTP端口(NameNode)。 该端口用于： 1. 点对点的NameNode检查点操作。 2. 远程Web客户端连接NameNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
dfs.namenode.https.port	9871	HDFS HTTPS端口(NameNode)。 该端口用于： 1. 点对点的NameNode检查点操作。 2. 远程Web客户端连接NameNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
dfs.datanode.ipc.port	9867	Datanode IPC 服务器端口。 该端口用于： 客户端连接DataNode用来执行RPC操作。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是

配置参数	默认端口	端口说明
dfs.datanode.port	9866	Datanode数据传输端口。 该端口用于： 1. HDFS客户端从DataNode传输数据或传输数据到DataNode。 2. 点对点的Datanode传输数据。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.http.port	9864	Datanode HTTP端口。 该端口用于： 安全模式下，远程Web客户端连接DataNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.https.port	9865	Datanode HTTPS端口。 该端口用于： 安全模式下，远程Web客户端连接DataNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.JournalNode.rpc.port	8485	JournalNode RPC端口。 该端口用于： 客户端通信用于访问多种信息。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口	端口说明
dfs.journalnode.http.port	8480	JournalNode HTTP端口。 该端口用于： 安全模式下，远程Web客户端链接JournalNode。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.journalnode.https.port	8481	JournalNode HTTPS端口。 该端口用于： 安全模式下，远程Web客户端链接JournalNode。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
httpfs.http.port	14000	HttpFS HTTP服务器侦听的端口。 该端口用于： 远程REST接口连接HttpFS。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Hive 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
templeton.port	9111	WebHCat 提供REST 服务的端口。 该端口用于： WebHCat客户端与WebHCat服务端之间的通信。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口	端口说明
hive.server2.thrift.port	10000	HiveServer 提供Thrift 服务的端口。 该端口用于： HiveServer客户端与HiveServer之间的通信。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hive.metastore.port	9083	MetaStore 提供Thrift 服务的端口。 该端口用于： MetaStore客户端与MetaStore之间的通信，即HiveServer与MetaStore之间通信。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hive.server2.webui.port	10002	Hive的WEB UI端口。 该端口用Web请求与Hive UI服务器进行HTTPS/HTTP通信。

Hue 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
HTTP_PORT	8888	hue提供Https服务端口。 该端口用于：https方式提供web服务，支持修改。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Kafka 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
port	9092	Broker提供数据接收、获取服务
ssl.port	9093	Broker提供数据接收、获取服务的SSL端口

配置参数	默认端口	端口说明
sasl.port	21007	Broker提供SASL安全认证端口，提供安全Kafka服务
sasl-ssl.port	21009	Broker提供SASL安全认证和SSL通信的端口,提供安全认证及通信加密服务

Loader 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
LOADER_HTTPS_PORT	21351	该端口用于提供Loader作业配置、运行的REST接口 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Manager 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
-	8080	WebService提供的供用户访问端口 该端口用于使用http协议访问Web UI ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
-	28443	WebService提供的供用户访问端口 该端口用于使用https协议访问Web UI ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

MapReduce 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
mapreduce.jobhistory.webapp.port	19888	Job history服务器Web http端口。 该端口用于：查看Job History服务器的Web页面。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
mapreduce.jobhistory.port	10020	Job history服务器端口。 该端口用于： 1. 用于MapReduce客户端恢复任务的数据。 2. 用于Job客户端获取任务报告。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
mapreduce.jobhistory.webapp.https.port	19890	Job history服务器Web https端口。 该端口用于查看Job History服务器的Web页面。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Spark 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
hive.server2.thrift.port	22550	JDBC thrift端口。 该端口用于： Spark2.1.0 CLI/JDBC与Spark2.1.0 CLI/JDBC服务器进行socket通信。 **说明** 如果hive.server2.thrift.port被占用，将抛端口被占用异常。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口	端口说明
spark.ui.port	4040	JDBC的Web UI端口 该端口用于：Web请求与JDBC Server Web UI服务器进行HTTPS/HTTP通信。 **说明** 系统会根据端口的设置取值，并验证其有效性；如果无效，端口+1，直到取到有效值为止（上限16次，重试次数可以通过配置spark.port.maxRetries改变）。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
spark.history.ui.port	18080	JobHistory Web UI端口 该端口用于：Web请求与Spark2.1.0 History Server间的HTTPS/HTTP通信 **说明** 系统会根据端口的设置取值，并验证其有效性；如果无效，端口+1，直到取到有效值为止（上限16次，重试次数可以通过配置spark.port.maxRetries改变）。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是

Storm 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
nimbus.thrift.port	6627	nimbus提供thrift服务
supervisor.slots.ports	6700,6701,6702,6703	接收由其它服务器转发过来的请求
logviewer.https.port	29248	logviewer提供Https服务
ui.https.port	29243	Storm ui提供Https服务(ui.https.port)

YARN 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
yarn.resourcemanager.webapp.port	8088	ResourceManager服务的web http 端口。
yarn.resourcemanager.webapp.https.port	8090	ResourceManager服务的web https 端口。 该端口用于：安全模式下，接入Resource Manager Web应用。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
yarn.nodemanager.webapp.port	8042	NodeManager Web http端口
yarn.nodemanager.webapp.https.port	8044	NodeManager Web https端口。 该端口用于： 安全模式下，接入NodeManager web应用。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是

ZooKeeper 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
clientPort	2181	Zookeeper客户端端口。 该端口用于： ZooKeeper客户端连接ZooKeeper服务器。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Kerberos 常用端口

表中涉及端口的协议类型均为：UDP。

配置参数	默认端口	端口说明
kdc_ports	21732	KerberOS服务端端口 该端口用于： 组件向kerberos服务认证。配置集群互信可能会用到； **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Opentsdb 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
tsd.network.port	4242	Opentsdb的WEB UI端口。 该端口用于：Web请求与Opentsdb UI服务器进行HTTPS/HTTP通信。

Tez 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
tez.ui.port	28888	Tez的WEB UI端口。

KafkaManager 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
kafka_manager_port	9099	KafkaManager的WEB UI端口。

Presto 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
http-server.http.port	7520	presto coordinator对外提供服务的http端口。
http-server.https.port	7521	presto coordinator对外提供服务的https端口。
http-server.http.port	7530	presto worker对外提供服务的http端口。
http-server.https.port	7531	presto worker对外提供服务的https端口。

Flink 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
jobmanager.web.port	32261-32325	Flink的WEB UI端口。 用于Client Web请求与Flink server进行HTTP/HTTPS通信。

ClickHouse 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
tcp_port	9000	业务客户端TCP接入端口。
http_port	8123	业务客户端HTTP接入端口。
https_port	8443	业务客户端HTTPS接入端口。
tcp_port_secure	9440	业务客户端TCP With SSL接入端口。默认仅在安全模式下开放。

Impala 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
--beeswax_port	21000	提供给impala-shell通信的端口。
--hs2_port	21050	提供给Impala应用通信的端口。
--hs2_http_port	28000	Impala对外提供HiveServer2协议的端口。

8.3 通过专线访问

MRS为您提供云专线（Direct Connect）方式访问MRS集群。云专线用于搭建用户本地数据中心与线上云VPC之间高速、低时延、稳定安全的专属连接通道，充分利用线上云服务优势的同时，继续使用现有的IT设施，实现灵活一体，可伸缩的混合云计算环境。

前提条件

云专线服务可用，并已打通本地数据中心到线上VPC的连接通道。

通过专线访问 MRS 集群

步骤1 登录MRS管理控制台。

步骤2 单击集群名称进入集群详情页。

步骤3 在集群详情页面的“概览”页签，单击“集群管理页面”右侧的“前往 Manager”。

步骤4 “访问方式”选择“专线访问”，并勾选“我确认已打通本地与浮动IP的网络，可使用专线直接访问MRS Manager。”。

浮动IP为MRS为您访问MRS Manager页面自动分配的IP地址，使用专线访问MRS Manager之前您确保云专线服务已打通本地数据中心到线上VPC的连接通道。

步骤5 单击“确定”，进入MRS Manager登录页面，用户名使用“admin”，密码为创建集群时设置的admin密码。

----结束

切换 MRS Manager 访问方式

为了便于用户操作，浏览器缓存会记录用户所选择的访问Manager的方式，如需切换访问Manager方式，参考如下步骤操作。

步骤1 登录MRS管理控制台。

步骤2 单击集群名称进入集群详情页。

步骤3 在集群详情页面的“概览”页签，单击“集群管理页面”右侧的 按钮。

步骤4 在弹出页面重新选择“访问方式”即可。

- 若由“EIP访问”切换为“专线访问”，请在专线网路互通的前提下，在弹出页面的“访问方式”选择“专线访问”并勾选“我确认已打通本地与浮动IP的网络，可使用专线直接访问MRS Manager。”后单击“确定”。
- 若由“专线访问”切换为“EIP访问”，在弹出页面的“访问方式”选择“EIP访问”并参考**通过弹性公网IP访问Manager**配置EIP。若集群已配置过公网IP，直接单击“确定”以EIP方式访问Manager。

----结束

8.4 通过弹性公网 IP 访问

为了方便用户访问开源组件的Web站点，MRS集群支持通过为集群绑定弹性公网IP的方式，访问MRS集群上托管的开源组件。该方式更加简便易操作，推荐使用该方式访问开源组件的Web站点。

为集群绑定弹性公网 IP 并添加安全组规则

1. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，待IAM用户同步成功后，在集群详情页会出现“组件管理”页签。

2. 单击“集群管理页面”右侧的“前往 Manager”。
3. 弹出访问MRS Manager页面，绑定弹性公网IP并添加安全组规则。仅首次访问该集群的组件开源站点时，需要如下配置。
   a. 绑定弹性公网IP，在弹性公网IP下拉框中选择可用的弹性公网IP。若没有可用的弹性公网IP，请单击“管理弹性公网IP”弹性公网IP后在该页面引用。若创建集群时已绑定弹性公网IP，请跳过该步骤。
   b. 选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
   c. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

      说明

      - 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
      - 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问MRS组件。

   d. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”
   e. 单击“确定”，进入登录页面，用户名使用“admin”，密码为创建集群时设置的admin密码。
4. 登录Manager页面，选择“集群 > 服务 > HDFS > NameNode WebUI > NameNode (*主机名称*，主)”，访问开源组件Web站点。此处仅以HDFS NameNode为例介绍，其他组件访问地址请参考**开源组件Web站点**页面提供的站点地址。

8.5 通过 Windows 弹性云服务器访问

MRS支持通过Windows弹性云服务器访问开源组件Web站点。该方式操作较为复杂，推荐不支持EIP功能的MRS集群使用。

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在 “现有集群” 列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“集群控制台地址”、“安全组”。

说明

集群控制台地址获取方式：远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示集群控制台地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到集群控制台地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。

步骤3 在ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个标准镜像“Windows Server 2012 R2 Standard 64bit(40GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和MRS集群的“安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为MRS集群的安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组中添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的帐号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

例如Windows 2012操作系统可以使用Internet Explorer 11。

Manager访问地址形式为**https://*集群控制台地址*:28443/web**。访问时需要输入MRS集群的用户名和密码，例如“admin”用户。

说明

- 集群控制台地址：远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示集群控制台地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到集群控制台地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。
- 如果使用其他MRS集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

步骤7 请参考**开源组件Web站点**页面提供的站点地址访问开源组件Web站点。

----结束

相关任务

配置集群节点名称与IP地址映射

步骤1 登录Manager，单击“主机管理”。

记录集群中所有节点的“主机名称”和“管理IP”。

步骤2 在工作环境使用“记事本”打开“hosts”文件，将节点名称与IP地址的对应关系填写到文件中。

每个对应关系填写一行，填写效果例如：

```
192.168.4.127 node-core-Jh3ER
192.168.4.225 node-master2-PaWVE
192.168.4.19 node-core-mtZ81
192.168.4.33 node-master1-zbYN8
192.168.4.233 node-core-7KoGY
```

保存修改。

----结束

8.6 创建连接 MRS 集群的 SSH 隧道并配置浏览器

操作场景

用户和MRS集群处于不同的网络中，需要创建一个SSH隧道连接，使用户访问站点的数据请求，可以发送到MRS集群并动态转发到对应的站点。

MAC系统暂不支持该功能访问MRS，请参考**通过弹性公网IP访问**内容访问MRS。

前提条件

- 准备一个SSH客户端用于创建SSH隧道，例如使用开源SSH客户端Git。请下载并安装。
- 已创建好集群，并准备pem格式的密钥文件或创建集群时的密码。
- 用户本地环境可以访问互联网。

操作步骤

步骤1 登录MRS管理控制台，选择“集群列表 > 现有集群”。

步骤2 单击指定名称的MRS集群。

记录集群的“安全组”。

步骤3 为集群Master节点的安全组添加一条需要访问MRS集群的IP地址的入规则，允许指定来源的数据访问端口“22”。

具体请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 查询集群的主管理节点，具体请参考**如何确认Manager的主备管理节点**。

步骤5 为集群的主管理节点绑定一个弹性IP地址。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤6 在本地启动Git Bash，执行以下命令登录集群的主管理节点：**ssh root@*弹性IP地址***或者**ssh -i *密钥文件路径* root@*弹性IP地址***

步骤7 执行以下命令查看数据转发配置：

cat /etc/sysctl.conf | grep net.ipv4.ip_forward

- 系统查询到“net.ipv4.ip_forward=1”表示已配置转发，则请执行**步骤9**。
- 系统查询到“net.ipv4.ip_forward=0”表示未配置转发，则请执行**步骤8**。
- 系统查询不到“net.ipv4.ip_forward”参数表示该参数未配置，则请执行以下命令后再执行**步骤9**。

  echo "net.ipv4.ip_forward = 1" >> /etc/sysctl.conf

步骤8 修改节点转发配置：

1. 执行以下命令切换**root**用户：

   sudo su - root

2. 执行以下命令，修改转发配置：

   echo 1 > /proc/sys/net/ipv4/ip_forward

   sed -i "s/net.ipv4.ip_forward=0/net.ipv4.ip_forward = 1/g" /etc/sysctl.conf

   sysctl -w net.ipv4.ip_forward=1

3. 执行以下命令，修改sshd配置文件：

   vi /etc/ssh/sshd_config

   按I进入编辑模式，查找“AllowTcpForwarding”和“GatewayPorts”，并删除注释符号，修改内容如下，然后保存并退出：

   ```
   AllowTcpForwarding yes
   GatewayPorts yes
   ```

4. 执行以下命令，重启sshd服务：

   service sshd restart

步骤9 执行以下命令查看浮动IP地址：

ifconfig

系统显示的“eth0:FI_HUE”表示为Hue的浮动IP地址，“eth0:wsom”表示Manager浮动IP地址，请记录“inet”的实际参数值。

然后退出登录：**exit**

步骤10 在本地机器执行以下命令创建支持动态端口转发的SSH隧道：

使用命令**ssh -i *密钥文件路径* -v -ND *本地端口地址* root@*弹性IP地址***或者***ssh -v -ND 本地端口地址 root@弹性IP地址***，然后输入创建集群时的密码。

其中，“本地端口地址”需要指定一个用户本地环境未被使用的端口，建议选择8157。

创建后的SSH隧道，通过“-D”启用动态端口转发功能。默认情况下，动态端口转发功能将启动一个SOCKS代理进程并侦听用户本地端口，端口的数据将由SSH隧道转发到集群的主管理节点。

步骤11 执行如下命令配置浏览器代理。

1. 进入本地Google Chrome浏览器客户端安装目录。
2. 按住“shift+鼠标右键”，选择“在此处打开命令窗口”，打开CMD窗口后输入如下命令：

   chrome --proxy-server="socks5://localhost:8157" --host-resolver-rules="MAP * 0.0.0.0 , EXCLUDE localhost" --user-data-dir=c:/tmppath --proxy-bypass-list="*google*com,*gstatic.com,*gvt*.com,*:80"

   说明

   - 8157为**步骤10**中配置的本地代理端口。
   - 若本地操作系统为Windows 10，请打开Windows操作系统“开始”菜单，输入**cmd**命令，打开一个命令行窗口执行**步骤11.2**中的命令。若该方式不能成功，请打开Windows操作系统“开始”菜单后，在搜索框中输入并执行**步骤11.2**中的命令。

步骤12 在新弹出的浏览器地址栏，输入Manager的访问地址。

Manager访问地址形式为**https://*Manager浮动IP地址*:28443/web**。

访问启用Kerberos认证的集群时，需要输入MRS集群的用户名和密码，例如“admin”用户。未启用Kerberos认证的集群则不需要。

第一次访问时，请根据浏览器提示，添加站点信任以继续打开页面。

步骤13 准备站点的访问地址。

1. 参考**Web站点一览**，获取Web站点的地址格式及对应的角色实例。
2. 单击“服务管理”。
3. 单击指定的服务名称，例如HDFS。
4. 单击“实例”，查看NameNode的主角色实例“NameNode(主)”的“业务IP”。

步骤14 在浏览器输入访问Web站点真实地址并访问。

步骤15 退出访问Web站点时，请终止并关闭SSH隧道。

----结束

9 访问集群 Manager

9.1 访问 FusionInsight Manager（MRS 3.x 及之后版本）

操作场景

MRS 3.x及之后版本的集群使用FusionInsight Manager对集群进行监控、配置和管理。用户在集群安装后可使用帐号登录FusionInsight Manager。

说明

如果不能正常登录组件的WebUI页面，请参考**通过ECS访问FusionInsight Manager**方式访问FusionInsight Manager。

通过弹性 IP 访问 FusionInsight Manager

步骤1 登录MRS管理控制台页面。

步骤2 单击“集群列表 > 现有集群”，在集群列表中单击指定的集群名称，进入集群信息页面。

步骤3 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中配置弹性IP信息。

1. 若创建MRS集群时暂未绑定弹性公网IP，在“弹性公网IP”下拉框中选择可用的弹性公网IP。若用户创建集群时已经绑定弹性公网IP，直接执行**步骤3.2**

   说明

   如果没有弹性公网IP，可先单击“管理弹性公网IP”弹性公网IP后，然后在弹性公网IP下拉框中选择的弹性公网IP。

2. 在“安全组”中选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
3. 添加安全组规则，默认填充的是用户访问弹性IP地址的规则，如需开放多个IP段为可信范围用于访问Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。
4. 勾选确认信息后，单击“确定”。

步骤4 单击“确定”，进入Manager登录页面。

步骤5 输入默认用户名“admin”及创建集群时设置的密码，单击“登录”进入Manager页面。

步骤6 在MRS管理控制台，在“现有集群” 列表，单击指定的集群名称，进入集群信息页面。

说明

如需给其他用户开通访问Manager的权限，请执行**步骤6**～**步骤9**，添加对应用户访问公网的IP地址为可信范围。

步骤7 单击弹性公网IP后边的“添加安全组规则”。

步骤8 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的*公网IP*/*端口号*是可信任的公网访问IP范围，我了解使用0.0.0.0/0会带来安全风险”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤6**-**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤9 单击“确定”完成安全组规则添加。

----结束

通过 ECS 访问 FusionInsight Manager

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在“现有集群”列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“集群管理页面”、“安全组”。

步骤3 在管理控制台首页服务列表中选择“弹性云服务器”，进入ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个标准镜像“Windows Server 2012 R2 Standard 64bit(40GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和Master节点的“默认安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为Master节点的默认安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的帐号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

例如Windows 2012操作系统可以使用Internet Explorer 11。

Manager访问地址为“集群管理页面”地址。访问时需要输入集群的用户名和密码，例如“admin”用户。

说明

- 如果使用其他集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。请咨询管理员。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

步骤7 注销用户退出Manager时移动鼠标到右上角  ，然后单击“注销”。

----结束

9.2 访问 MRS Manager（MRS 2.x 及之前版本）

操作场景

MRS使用Manager对集群进行监控、配置和管理，用户可以在MRS控制台页面打开Manager管理页面，使用创建集群时设置的admin帐号和密码登录Manager。

通过弹性公网 IP 访问 Manager

步骤1 登录MRS管理控制台页面。

步骤2 单击“集群列表 > 现有集群”，在集群列表中单击指定的集群名称，进入集群信息页面。

步骤3 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中“访问方式”选择“EIP访问”。专线访问请参考**通过专线访问**。

1. 若用户创建集群时暂未绑定弹性公网IP，在弹性公网IP下拉框中选择可用的弹性公网IP。若用户创建集群时已经绑定弹性公网IP，直接执行**步骤3.2**。

   说明

   如果没有弹性公网IP，可先单击“管理弹性公网IP”弹性公网IP后，然后在弹性公网IP下拉框中选择的弹性公网IP。
2. 选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
3. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则，如需开放多个IP段为可信范围用于访问MRS Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问MRS Manager服务。

4. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

步骤4 单击“确定”，进入MRS Manager登录页面。

步骤5 输入默认用户名“admin”及创建集群时设置的密码，单击“登录”进入MRS Manager页面。

步骤6 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

说明

如需给其他用户开通访问MRS Manager的权限，请执行**步骤6**-**步骤9**，添加对应用户访问公网的IP地址为可信范围。

步骤7 单击弹性公网IP后边的“添加安全组规则”。

步骤8 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的授权对象是可信任的公网访问IP范围，禁止使用0.0.0.0/0,否则会有安全风险。”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤6**-**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤9 单击“确定”完成安全组规则添加。

----结束

通过 ECS 访问 Manager

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在 “现有集群” 列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“安全组”。

步骤3 在ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个标准镜像“Windows Server 2012 R2 Standard 64bit(40GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和MRS集群的“默认安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为MRS集群的默认安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组中添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的帐号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

例如Windows 2012操作系统可以使用Internet Explorer 11。

Manager访问地址形式为**https://*集群控制台地址*:28443/web**。访问时需要输入MRS集群的用户名和密码，例如“admin”用户。

说明

- 集群控制台地址：远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示集群控制台地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到集群控制台地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。
- 如果使用其他MRS集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

步骤7 注销用户退出Manager时移动鼠标到右上角  ，然后单击“注销”。

----结束

为集群更换弹性公网 IP

步骤1 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤2 查看“弹性公网IP”。

步骤3 登录“虚拟私有云 VPC”管理控制台。

步骤4 选择“弹性公网IP和带宽 > 弹性公网IP”。

步骤5 查找MRS集群所绑定的弹性公网IP，并在“操作”列单击“解绑”解绑MRS集群绑定的弹性公网IP。

步骤6 登录MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

此时，集群详情页面“弹性公网IP”显示“暂未绑定”。

步骤7 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中“访问方式”选择“EIP访问”。

步骤8 在弹性公网IP下拉框中选择新的弹性公网IP并配置他参数，具体请参考**通过弹性公网IP访问Manager**。

----结束

为其他用户开通访问 MRS Manager 的权限

步骤1 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤2 单击弹性公网IP后边的“添加安全组规则”。

步骤3 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的授权对象是可信任的公网访问IP范围，禁止使用0.0.0.0/0,否则会有安全风险。”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤1**-**步骤4**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤4 单击“确定”完成安全组规则添加。

----结束

10 FusionInsight Manager 操作指导（适用于 3.x）

10.1 从这里开始

10.1.1 FusionInsight Manager 入门指导

概述

MRS为用户提供海量数据的管理及分析功能，快速从结构化和非结构化的海量数据中挖掘您所需要的价值数据。开源组件结构复杂，安装、配置、管理过程费时费力，使用FusionInsight Manager将为您提供企业级的集群的统一管理平台：

- 提供集群状态的监控功能，您能快速掌握服务及主机的运行状态。
- 提供图形化的指标监控及定制，您能及时的获取系统的关键信息。
- 提供服务属性的配置功能，满足您实际业务的性能需求。
- 提供集群、服务、角色实例的操作功能，满足您一键启停等操作需求。
- 提供权限管理及审计功能，您能设置访问控制及管理操作日志。

浏览器支持能力

- Google Chrome
  推荐使用Google Chrome 93～95版本。
- Edge
  支持Windows 10系统自带的Edge浏览器。

说明

推荐使用Windows平台的浏览器访问FusionInsight Manager。

系统界面简介

FusionInsight Manager提供统一的集群管理平台，帮助您快捷、直观的完成集群的运行维护。

界面最上方为操作栏，中部为显示区，最下方为任务栏。

操作栏各操作入口的详细功能如**表10-1**所示。

表 10-1 界面操作入口功能描述

入口	功能描述
主页	提供柱状图、折线图、表格等多种图表方式展示集群的主要监控指标、主机的状态统计。您可以定制关键监控信息面板，并拖动到任意位置。系统概览支持数据自动刷新，请参见**主页**。
集群	提供各集群内服务监控、服务操作向导以及服务配置，帮助您对服务进行统一管理。请参见**集群**。
主机	提供主机监控、主机操作向导，帮助您对主机进行统一管理。请参见**主机**。
运维	提供告警查询、告警处理指导功能。帮助您及时发现产品故障及潜在隐患，并进行定位排除，以保证系统正常运行。请参见**运维**。
审计	提供审计日志查询及导出功能。帮助您查阅所有用户活动及操作。请参见**审计**。
租户资源	提供统一租户管理平台。请参见**租户资源**。
系统	提供对FusionInsight Manager的系统管理设置，例如用户权限设置。请参见**系统设置**。

10.1.2 查询 FusionInsight Manager 版本号

管理员通过查看FusionInsight Manager版本号，可以进行下一步的系统升级及日常维护操作。

- 界面方式

  登录FusionInsight Manager，在主页界面，单击右上角的  ，在下拉框中单击“关于”，在弹框中查看FusionInsight Manager版本号。

- 命令方式

  a. 以**root**用户登录FusionInsight Manager主管理节点。

  b. 执行如下命令，查看FusionInsight_Manager的版本号及平台信息。

  su - omm

  cd ${BIGDATA_HOME}/om-server/om/sbin/pack

  ./queryManager.sh

  显示如下：

  ```
  Version          Package                          Cputype
  ***              FusionInsight_Manager_***        x86_64
  ```

  说明

  此处版本号***以实际查询的版本号为准。

10.1.3 登录管理系统

操作场景

该任务指导用户在Manager安装后使用帐号登录FusionInsight Manager。

操作步骤

步骤1 获取FusionInsight Manager的网络地址。

步骤2 打开页面后，输入系统用户和密码。

步骤3 新用户登录需要修改密码。

用户密码策略：

- 密码字符长度必须为8～64个字符。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符`~!@#$%^&*()-_=+|[{}];',<.>/\?中的4种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与当前密码相同。

步骤4 将光标移动到FusionInsight Manager右上角的，在弹出窗口中单击“注销”，单击“确定”后可退出当前登录用户。

----结束

10.1.4 登录管理节点

操作场景

部分运维操作的脚本与命令需要或只支持在主管理节点上运行。管理员可以根据以下指导确认并登录主或备管理节点。

在 Manager 查看主备管理节点并登录

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > OMS”。

在“基本信息”区域，“当前主用”表示主管理节点的主机名，“当前备用”表示备管理节点的主机名。

单击主机名可进入对应的主机详情页面。记录主机的IP地址信息。

步骤3 以**root**用户登录主或备管理节点。

----结束

执行脚本确定主备管理节点并登录

步骤1 以**root**用户登录任意部署Manager的节点。

步骤2 执行以下命令确认主备管理节点。

su - omm

sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh

界面打印信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“Master1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“Master2”为备管理节点）。

```
HAMode
double
NodeName            HostName        HAVersion         StartTime             HAActive
HAAllResOK          HARunPhase
192-168-0-30        Master1         V100R001C01       xxxx-09-01 07:12:05     active
normal              Actived
192-168-0-24        Master2         V100R001C01       xxxx-09-01 07:14:02     standby
normal              Deactived
```

步骤3 执行如下命令获取主备管理节点IP地址。

cat /etc/hosts

获取的主备管理节点IP地址示例如下：

```
127.0.0.1       localhost
192.168.0.30    Master1
192.168.0.24    Master2
```

步骤4 以**root**用户登录主或备管理节点。

----结束

10.2 主页

10.2.1 主页概述

登录FusionInsight Manager以后，Manager界面将默认显示“主页”标签中的内容，“综述”页面提供各集群服务状态预览区及监控状态报表，“告警分析”页面展示TOP告警统计及分析。

- 主页右侧可查看集群的不同级别告警个数、运行任务个数、当前用户和帮助信息等内容。
  - 单击  可查看“任务管理中心”中近100次操作任务的任务名、集群、状态、进度、开始时间和结束时间。

    说明

    对于启动、停止、重启以及滚动重启操作，在任务执行过程中，单击任务列表中的对应任务名称，单击“中止”按钮，根据界面提示输入管理员密码后，用户可中止该任务。中止后，任务将不再继续执行。

  - 单击  可获得帮助信息，如表10-2所示。

    表 10-2 帮助信息一览表

项目	描述
“关于”	提供FusionInsight Manager版本号信息。

- 主页底部任务栏显示FusionInsight Manager的语言选项和当前集群时间及时区信息，可切换系统语言。

服务状态预览区

主页界面的左侧展示各集群主机个数及已安装服务个数，可通过单击 ，展开对应集群的全部服务信息，查看当前集群已安装各服务的状态和告警情况。

通过单击 ，对当前集群进行基本的运维管理操作，详情请参考**表10-3**。

每个服务名称左侧的表示当前该服务运行状态良好，表示当前服务启动失败，表示当前服务未启动。

同时服务名称右侧可查看当前该服务是否产生了告警，如果存在告警，则以图标区分告警的级别并显示告警数。

对于支持多服务特性的组件，若在同一集群中安装了多个服务，服务的右侧会显示安装的个数。

如果服务右侧显示则表示该服务配置已过期。

监控状态报表

主页界面的右侧为图表区，包含关键监控状态的报表，例如集群中所有主机的状态、主机CPU使用率、主机内存使用率等。用户可以自定义在图表区展示的监控报表，管理监控指标请参考**管理监控指标数据报表**。

监控图表的数据来源可在图表的左下方查看，每个监控报表可以放大查看具体数值，也可以关闭不再显示。

告警分析

“告警分析”页面展示“Top20告警统计”表和“Top3告警分析”图。单击“Top20告警统计”中的告警名称，可以在告警分析中只展示该告警信息。该功能支持告警统计，可以展示TOP告警以及发生的时间规律，可以有针对性地解决告警，提升系统稳定性。

10.2.2 管理监控指标数据报表

操作场景

FusionInsight Manager支持用户自定义在主页进行展示的监控项，也可以导出监控数据。

说明

历史报表根据所自定义的时间长度不同，图表横轴中每个时间间隔也会不同，具体监控数据的规则如下：

- 0～25小时：每个间隔5分钟，要求集群至少安装10分钟以上，最多保留15天监控数据。
- 25小时～150小时：每个间隔30分钟，要求集群至少安装30分钟以上，最多保留3个月监控数据。
- 150小时～300小时：每个间隔1小时，要求集群至少安装1小时以上，最多保留3个月监控数据。
- 300小时～300天：每个间隔1天，要求集群至少安装1天以上，最多保留6个月监控数据。
- 300天以上：每个间隔7天，要求集群安装7天以上，最多保留一年的监控数据。
- 如果FusionInsight Manager存储所用的GaussDB所在分区的磁盘使用率超过80%时，会清理实时监控数据和周期为5分钟的监控数据。
- 若为“租户资源”下的“存储资源(HDFS)”表，0小时～300小时：每个间隔1小时，要求集群至少安装1小时以上，最多保留3个月监控数据。

自定义监控指标报表

步骤1 登录FusionInsight Manager。

步骤2 单击“主页”。

步骤3 在图表区的右上角，单击 ▼ ，在弹出菜单中选择“定制”。

说明

监控时段以5分钟为单位，显示最近1小时的监控数据；从进入“实时监控”页面后，在监控图右侧以5分钟为单位显示实时监控数据。

步骤4 在窗口左侧分类中，选择一项监控资源主体。

步骤5 在右侧监控列表勾选一个或多个监控指标。

步骤6 单击“确定”。

----结束

导出全部监控数据

步骤1 登录FusionInsight Manager。

步骤2 单击“主页”。

步骤3 在所需要操作的集群的图表区的右上角，选择一个时间范围获取监控数据，例如“1周”。

默认为实时数据，无法导出。单击可以自定义监控数据时间范围。

步骤4 在图表区的右上角，单击 ▼ ，在弹出菜单中选择“导出”。

----结束

导出指定监控项数据

步骤1 登录FusionInsight Manager。

步骤2 单击“主页”。

步骤3 在所需要操作的集群的图表区任意一个监控报表窗格的右上角，单击。

步骤4 选择一个时间范围获取监控数据，例如“1周”。

默认为实时数据，无法导出。单击可以自定义监控数据时间范围。

步骤5 单击“导出”。

----结束

10.3 集群

10.3.1 管理集群

10.3.1.1 集群管理概述

总览

登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 概览”可以查看当前集群的主要状态信息。

在“概览”页面上可对当前集群进行基本管理操作，如启动、停止、滚动重启、同步配置等，具体如**表10-3**所示。

表 10-3 维护管理功能

操作入口	说明
“启动”	将集群中所有服务启动。
“停止”	将集群中所有服务停止。
“更多 > 重启”	将集群中所有服务重启。
“更多 > 滚动重启”	为集群中所有服务提供不中断业务的重启操作，具体可参考**滚动重启集群**。
“更多 > 同步配置”	为集群中所有服务启用新的配置参数。
“更多 > 重启配置过期的实例”	为集群中所有服务重启配置过期的实例，具体可参考**管理配置过期**。
“更多 > 健康检查”	为OMS、集群所有服务和所有节点进行健康检查，健康检查可以包含三方面检查项：各检查对象的运行状态、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“运行状态”。 健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。如果发现问题，可以单击“查看帮助”。
“ 更多 > 下载客户端”	为用户下载默认的客户端，具体可参考**下载客户端**。

操作入口	说明
“ 更多 > 导出安装模板”	将集群所有安装配置批量导出，例如集群认证模式、节点信息、服务配置等，可用于相同环境下集群重新安装的场景。
“更多 > 导出配置”	将集群所有服务的配置批量导出。
“更多 > 进入维护模式/退出维护模式”	配置集群进入/退出维护模式。
“更多 > 维护模式视图”	查看集群进入维护状态的服务或主机。

10.3.1.2 滚动重启集群

操作场景

滚动重启指当集群中服务角色升级更新或修改配置后，在尽可能不中断业务的前提下的重启操作。

如果需要批量为集群中所有服务进行重启且不中断业务，可执行集群滚动重启操作。

说明

- 部分服务不支持滚动重启，在执行滚动重启集群的过程中，不支持滚动重启的服务将进行普通重启，业务可能会中断。请根据界面提示是否可以执行操作。
- 如果修改了端口类等需要尽快生效的配置（例如服务端的端口），则不建议通过滚动重启的方式使之生效，建议采用普通重启。

对系统的影响

与普通重启相比，滚动重启不会导致服务业务中断，但是滚动重启将比普通重启要花费更长的时间，且对应服务的吞吐量、性能等可能会受到影响。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 滚动重启”。

步骤3 输入当前登录的用户密码确认身份，单击“确定”。

步骤4 根据实际情况调整相关参数，如表10-4所示。

表 10-4 滚动重启参数

参数名称	描述
“只重启集群内配置过期的实例”	是否只重启集群内修改过配置的实例。

参数名称	描述
“启用机架策略”	是否启用机架并发滚动重启策略，只对满足机架策略滚动重启的角色（角色支持机架感知功能，且角色下的实例归属于2个或2个以上的机架）生效。 **说明** 该参数仅在滚动重启HDFS、Yarn时可设置。
“数据节点滚动重启并发数”	采用分批并发滚动重启策略的数据节点实例每一个批次重启的实例数，默认为1。 **说明** ● 该参数仅对同时满足“采用并发滚动策略”和“实例为数据节点”两个条件时才有效。 ● 当启用机架策略时，该参数将失效，集群以机架策略默认配置的最大实例数（默认值为20）作为一个机架内分批并发重启的最大实例数。 ● 该参数仅在滚动重启HDFS、HBase、Yarn、Kafka、Storm、Flume时可设置。 ● HBase的RegionServer滚动重启的并发数不支持手动配置，会根据RegionServer的节点数自行调整，调整规则为：30节点以内，每个批次1个节点；300节点以内，每个批次2个节点；300节点以上(含300节点)，每个批次1%(向下取整)个节点。
“批次时间间隔”	滚动重启实例批次之间的间隔时间，默认为0。
“退服超时时间”	角色实例在滚动重启过程中的退服等待时间，默认为1800s。 部分角色（例如HiveServer、JDBCServer）在滚动重启前会暂时停止提供服务，该状态下的实例不可再接入新的客户端连接，而已经存在的连接需要等待一段时间才能完成，配置合适的超时时间参数能尽可能地保证业务不中断。 **说明** 该参数仅在滚动重启Hive、Spark2x时可设置。
“批次容错阈值”	滚动重启实例批次执行失败容错次数，默认为0，即表示任意一个批次的实例重启失败后，滚动重启任务终止。

说明

“数据节点滚动重启并发数”、“批次时间间隔”、“批次容错阈值”等高级参数需要根据实际情况合理设置，否则可能导致服务业务中断或者严重影响性能，请谨慎调整。

例如：

- “数据节点滚动重启并发数”过大，同时重启多个实例导致服务业务中断或者由于剩余工作实例较少严重影响性能。
- “批次容错阈值”过大，某一批次实例失败后继续重启下一批次实例，导致服务业务中断。

步骤5 单击“确定”，等待滚动重启完成。

----结束

10.3.1.3 管理配置过期

操作场景

某个新的配置需要同时下发到集群所有服务，或修改某项配置后导致多个不同服务的“配置状态”为“配置过期”或“失败”时，表示这些服务的配置参数值未同步且未生效，管理员可以对集群执行同步配置功能，并在同步配置后重启相关服务实例，使所有服务启用新的配置参数。

若集群中服务配置均已同步但未生效，需重启配置过期的实例。

对系统的影响

- 集群执行同步配置后，需要重启配置过期的服务。重启时对应的服务不可用。
- 重启配置过期的实例时，该实例不可用。

操作步骤

同步配置

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 概览”。

步骤3 选择“更多 > 同步配置”。

步骤4 在弹出窗口中单击“确定”，开始为当前集群同步配置。

----结束

重启配置过期的实例

步骤1 选择“更多 > 重启配置过期的实例”。

步骤2 在弹出窗口中输入当前登录的用户密码确认身份，然后单击“确定”。

步骤3 在确认重启实例的对话框中单击“确定”。

支持单击“查看实例”打开所有配置已过期的实例列表，确认可以执行重启任务。

----结束

10.3.1.4 下载客户端

操作场景

MRS集群提供了默认的客户端，用户可以通过客户端执行管理操作、运行业务或进行二次开发。使用客户端前需要下载客户端软件包。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 下载客户端”。

界面显示“下载集群客户端”对话框。

步骤3 在“选择客户端类型”选择一个类型。

- “完整客户端”表示下载包中包含了脚本、编译文件和配置文件。
- “仅配置文件”表示下载包仅包含客户端配置文件。

  一般适用于应用开发任务。例如完整客户端已下载并安装后，管理员通过Manager界面修改了服务配置，开发人员需要更新客户端配置文件的场景。

  说明

  平台类型包括x86_64和aarch64两种，可分别在x86和TaiShan节点上安装使用。默认情况下，下载的客户端平台类型和服务端保持一致。

步骤4 是否在集群的节点中生成客户端软件包文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件。

  文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client/”。支持修改为其他目录且**omm**用户拥有目录的读、写与执行权限。如果路径中已存在客户端文件，会覆盖路径下已有的客户端文件。

  等待文件生成后，使用**omm**用户或客户端安装用户将获取的下载包复制到其他目录，例如“/opt/Bigdata/client”。
- 否，单击“确定”，下载客户端文件至本地。

  开始下载客户端软件包，并等待下载完成。

客户端下载成功后，参考**安装客户端**进行客户端的安装。

----结束

10.3.1.5 修改集群属性

操作场景

FusionInsight Manager支持用户在集群安装完成后查看基本属性。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 集群属性”。

默认可查看集群名称、集群描述、产品类型、集群ID、认证模式、创建时间和已安装部件信息。

步骤3 修改“集群名称”。

1. 单击，填入新的名称。

   支持的命名规则：集群名称只能包含汉字、字母、数字、下划线（_）、中划线（-）和空格，仅以汉字、字母、数字、下划线（_）或中划线（-）开头，只能在中间包含空格，并且最小长度为2个字符，最大长度不能超过199个字符。
2. 单击“确定”使新的集群名称生效。

步骤4 修改“集群描述”。

1. 单击，填入新的描述信息。

   只能包含汉字、英文字母、数字、中英文逗号、中英文句号、下划线（_）、空格和换行符，并且不能超过199个字符。

2. 单击“确定”使新的描述生效。

----结束

10.3.1.6 管理集群配置

操作场景

FusionInsight Manager支持一键查看集群内各服务配置参数的变动情况，方便用户快速排查定位问题，提升配置管理效率。

管理员可通过配置界面快速查看集群内各服务所有非初始默认值、同一角色实例之间非统一值、集群配置修改的历史记录、集群内当前配置状态为过期的参数。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 配置”。

步骤3 根据操作场景，选择对应操作页面：

- 查看所有非默认值：
  a. 单击“所有非默认值”，界面将显示当前集群内各服务、角色或实例的配置参数中，与初始默认值不一致的参数项。

     单击参数值后面的图标可快速恢复配置项的参数值至系统默认值，单击图标可查看该配置项的历史修改记录。

     配置参数较多时，可通过界面右上角的服务过滤框进行筛选，或者在搜索框中直接搜索关键字。
  b. 如需修改配置项参数值，根据参数描述修改配置后，单击“保存”，在弹出的窗口中单击“确定”。
- 查看所有非统一值：
  a. 单击“所有非统一值”，界面将显示当前集群内角色级别、服务级别、实例组级别或实例级别的存在差异化配置的配置项。

     单击参数值后面的图标，在弹出的窗口中可查看具体的差异项。

  b. 如需修改配置项参数值，可单击图标取消下层的配置差异化或手动调整，然后单击“确定”，再单击“保存”，在弹出的窗口中单击“确定”。
- 查看过期配置：
  a. 单击“过期配置”，界面将显示当前集群内配置过期的配置项。
  b. 可通过界面上方的服务过滤框进行筛选，查看不同服务的过期配置，或者在搜索框中直接搜索关键字。
  c. 处于过期状态的配置项并未完全生效，在不影响业务情况下，请及时重启配置过期的服务或实例。
- 查看历史配置记录：
  a. 单击“历史配置”，界面将显示当前集群的历史配置变更记录，用户可查看具体的参数值变动详情，包括所属服务、修改前与修改后的参数值、参数文件等内容。

b. 如需还原某次配置变更，可单击记录所在行“操作”列的“还原配置”按钮，在弹出的窗口中单击“确定”。

说明

部分配置项在修改参数值后需重启对应服务才会生效，在保存配置后请及时重启配置过期的服务或实例。

----结束

10.3.1.7 静态服务池

10.3.1.7.1 静态服务资源

简介

集群分配给各个服务的资源是静态服务资源，这些服务包括Flume、HBase、HDFS和Yarn。每个服务的计算资源总量固定，不与其他服务共享，是静态的。租户通过独占或共享一个服务来获取这个服务运行时需要的资源。

静态服务池

静态服务池用来指定服务资源的配置。

在服务级别上，静态服务池对各服务可使用的资源进行统一管理：

- 限制服务使用的资源总量，支持配置Flume、HBase、HDFS和Yarn在部署节点可使用的CPU、I/O和内存总量。
- 实现服务级别的资源隔离，可将集群中的服务与其他服务隔离，使一个服务上的负载对其他服务产生的影响有限。

调度机制

静态服务资源支持基于时间的动态调度机制，可以在不同时间段为服务配置不同的资源量，优化客户业务运行环境，提高集群的效率。

在一个复杂的集群环境中，多种服务共享使用集群资源，但是各服务的资源使用周期可能会有比较大的区别。

例如以下业务场景，对于一个银行客户：

- 在白天HBase查询服务的业务多。
- 在晚上查询服务的业务少而Hive分析服务业务多。

如果只给每个服务设置固定的资源可能会导致：

- 白天查询服务的资源不够用，分析服务的资源空闲。
- 晚上分析服务的资源不够用，查询服务的资源空闲。

集群资源利用率不高，而且服务能力也打了折扣。因此：

- 白天多配置HBase服务资源。
- 晚上多配置Hive服务资源。

这种基于时间的动态调度机制可以更高效的利用资源、运行任务。

10.3.1.7.2 配置集群静态资源

操作场景

当需要控制集群服务可以使用节点资源的情况，或者控制集群服务在不同时间段节点可用配额的CPU与I/O资源时，管理员可以在FusionInsight Manager调整资源基数，并自定义资源配置组。

对系统的影响

- 配置静态服务池后，受影响的服务的“配置状态”将显示为“配置过期”，需要重启服务，重启期间服务不可用。
- 配置静态服务池后，各服务及角色实例使用的最大资源将不能超过限制。

操作步骤

修改资源调整基数

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 静态服务池”。

步骤2 单击右上角“配置”，进入静态资源池配置向导。

步骤3 在“系统资源调整基数”分别修改参数“CPU（%）”和“Memory（%）”。

修改“系统资源调整基数”将限制服务能够使用节点的最大物理CPU和内存资源百分比。如果多个服务部署在同一节点，则所有服务使用的最大物理资源百分比不能超过此参数值。

步骤4 单击“下一步”。

需要重新修改参数，可单击“上一步”返回。

修改资源池默认“default”配置组

步骤5 单击“default”，在“权重配置”表格中各服务对应的“CPU LIMIT(%)”、“CPU SHARE(%)”、“I/O(%)”和“Memory(%)”填写各服务的资源使用百分比数量。

说明

- 所有服务使用的“CPU LIMIT(%)”和 “CPU SHARE(%)”资源配置总和可以大于100%。
- 所有服务使用的“I/O(%)”资源配置总和可以大于100%，不能为0。
- 所有服务使用的“Memory(%)”资源配置总和可以小于或等于100%，也可以大于100%。
- “Memory(%)”不支持动态生效，仅在“default”配置组中可以修改。
- “CPU LIMIT(%)”用于配置服务可使用的CPU核数与节点可分配的CPU核数占比。
- “CPU SHARE(%)”用于配置服务在与其他服务使用同一个CPU核的时间占比，即多个服务在使用同一个CPU核发生争抢时的时间占比。

步骤6 单击“根据权重配置生成详细配置”，FusionInsight Manager将根据集群硬件资源与分配情况，生成资源池实际参数配置值。

步骤7 单击“确定”。

在弹出窗口单击“确定”，确认保存配置。

添加自定义资源配置组

步骤8 是否需要在不同时间段自动调整资源配置?

- 是，执行**步骤9**。
- 否，只需要使用“default”在所有时间段生效，任务结束。

步骤9 单击“配置”，修改“系统资源调整基数”，然后单击“下一步”。

步骤10 单击“添加”增加新的资源配置组。

步骤11 在“第一步：调度时间”，单击“配置”显示时间策略配置页面。

根据业务需要修改以下参数，并单击“确定”保存：

- “重复”：勾选时表示此资源配置组按调度周期重复运行。不勾选时请设置一个资源配置组应用的日期与时间。
- “重复策略”：支持“每天”、“每周”和“每月”。仅在“重复”模式中生效。
- “在”：表示资源配置组应用的开始与结束时间。请设置一个唯一的时间区间，如果与已有配置组的时间区间有重叠，则无法保存。

说明

- “default”配置组会在所有未定义的时间段内生效。
- 新增加的配置组属于动态生效的配置项集合，在配置组应用的时间区间内可直接生效。
- 新增加的配置组可以被删除。最多增加4个动态生效的配置组。
- 选择任一种“重复策略”，如果结束时间小于开始时间，默认标识为第二天的结束时间。例如“22:00”到“6:00”表示调度时间为当天22点到第二天6点。
- 若多个配置组的“重复策略”类型不相同，则时间区间可以重叠，且生效的策略优先级从低到高的顺序为“每天”、“每周”、“每月”。例如，有“每月”与“每天”的调度配置组，时间区间分别为4:00到7:00，6:00到8:00，此时以每月的配置组为准。
- 若多个配置组的“重复策略”类型相同，当日期不相同时，则时间区间可以重叠。例如，有两个“每周”的调度配置组，可以分别指定时间区间为周一和周三的4:00到7:00。

步骤12 在“第二步：权重配置”修改各服务资源配置。

步骤13 单击“根据权重配置生成详细配置”，FusionInsight Manager将根据集群硬件资源与分配情况，生成资源池实际参数配置值。

步骤14 单击“确定”。

在弹出窗口单击“确定”，确认保存配置。

----结束

10.3.1.7.3 查看集群静态资源

操作场景

大数据管理平台支持通过静态服务资源池对没有运行在Yarn上的服务资源进行管理和隔离。系统支持基于时间的静态服务资源池自动调整策略，使集群在不同的时间段自动调整参数值，从而更有效地利用资源。

系统管理员可以在FusionInsight Manager查看静态服务池各个服务使用资源的监控指标结果，包含监控指标如下：

- 服务总体CPU使用率

- 服务总体磁盘IO读速率
- 服务总体磁盘IO写速率
- 服务总体内存使用大小

说明

启用多实例功能后，支持管理HBase所有服务实例使用的CPU、I/O和内存总量。

操作步骤

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 静态服务池”。

步骤2 在“配置组列表”，单击一个配置组，例如“default”。

步骤3 查看系统资源调整基数。

- “系统资源调整基数”表示集群中每个节点可以被集群服务使用的最大资源。如果节点只有一个服务，则表示此服务独占节点可用资源。如果节点有多个服务，则表示所有服务共同使用节点可用资源。
- “CPU”表示节点中服务可使用的最大CPU。
- “Memory”表示节点中服务可使用的最大内存。

步骤4 在图表区域，查看集群服务资源使用状态指标数据图表。

说明

- 可通过“为图标添加服务”，将特定服务的静态服务资源数据添至图表，最多可选择12个服务。
- 管理单个图表的操作，可参见**管理监控指标数据报表**。

----结束

10.3.1.8 客户端管理

10.3.1.8.1 管理客户端

操作场景

FusionInsight Manager支持统一管理集群的客户端安装信息，用户下载并安装客户端后，界面可自动记录已安装（注册）客户端的信息，方便查询管理。同时系统支持手动添加、修改未自动注册的客户端信息（如历史版本已安装的客户端）。

操作步骤

查看客户端信息

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 >*待操作集群的名称* > 客户端管理”，即可查看当前集群已安装的客户端信息。

用户可查看客户端所在节点的IP地址、安装路径、组件列表、注册时间及安装用户等信息。

在当前最新版本集群下载并安装客户端时，客户端信息会自动注册。

添加客户端信息

步骤3 如需手动添加已安装好的客户端信息，单击“添加”，根据界面提示手动添加客户端的IP地址、安装路径、用户、平台信息、注册信息等内容。

步骤4 配置好客户端信息，单击“确定”，添加成功。

修改客户端信息

步骤5 手动注册的客户端信息可以手动修改。

在“客户端管理”界面选择待修改的客户端，单击“修改”。修改信息后，单击“确定”完成修改。

删除客户端信息

步骤6 在“客户端管理”界面选择待删除的客户端，单击“删除”，在弹出的窗口中单击“确定”，即可删除客户端信息。

如需删除多个客户端信息，勾选待删除的客户端，单击“批量删除”，在弹出的窗口中单击“确定”，即可删除客户端信息。

导出客户端信息

步骤7 在“客户端管理”界面选择待操作的客户端，单击“导出全部”可导出所有已注册的客户端信息到本地。

说明

客户端管理界面上组件列表栏只展示有真实客户端的组件，因此部分没有客户端的组件和客户端特殊的组件不会显示在组件列表栏。

不显示的组件有：

LdapServer、KrbServer、DBService、Hue、Mapreduce、Flume

----结束

10.3.1.8.2 批量升级客户端

操作场景

在FusionInsight Manager界面上下载的客户端包中包含客户端批量升级工具，当集群升级或扩容后需要对多个客户端进行升级时，可以使用该工具对客户端进行批量一键升级。同时客户端批量升级工具提供了轻量级的批量刷新客户端所在节点“/etc/hosts”文件的功能。

操作步骤

客户端升级前准备

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 >*待操作集群的名称* > 更多 > 下载客户端”，下载完整客户端到服务端指定目录。

具体操作看参考**下载客户端**。

解压新下载的客户端，在解压后的目录找到batch_upgrade目录，例如“/tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig/batch_upgrade”。

步骤3 选择“集群 >*待操作集群的名称* > 客户端管理”，进入客户端管理界面，单击“导出全部”，将所选的客户端信息导出到本地。

步骤4 解压导出的客户端信息，将client-info.cfg文件上传到客户端解压目录的batch_upgrade目录下。

步骤5 参见**参考信息**，补全“client-info.cfg”中缺失的密码。

批量升级客户端

步骤6 执行**sh client_batch_upgrade.sh -u -f /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar -g /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig/batch_upgrade/client-info.cfg**，进行升级。

须知

由于配置了密码信息，执行完升级后建议尽快删除**client-info.cfg**文件。

步骤7 升级执行完成后确认结果。确保客户端升级无误后执行**sh client_batch_upgrade.sh -c**，确认升级结果。

步骤8 如果客户端升级后存在问题，可以执行**sh client_batch_upgrade.sh -s**，回滚客户端。

说明

- 客户端批量升级工具本身是将原客户端move至备份目录，然后再使用**-f**参数指定的客户端包再次安装客户端。因此若原客户端中有定制的内容，请在执行**-c**命令之前，将定制的内容从备份目录手动保存或者移至升级后的客户端目录。客户端备份路径为：*{原客户端路径}*_backup。
- 参数**-u**是**-c**和**-s**的前提，必须在**-u**命令执行了升级之后，才能选择是要执行**-c**进行提交还是**-s**进行回滚。
- 升级命令（**-u**）可以多次执行，每次执行只升级前面升级失败的客户端，跳过升级成功的客户端。
- 客户端批量升级工具也支持升级之前的旧客户端。
- 执行非root用户安装的客户端升级时，请确保相应用户在目标节点客户端所在目录及父目录的读写权限，否则会升级失败。
- **-f**参数输入的客户端包必须为全量客户端，不支持单组件或部分组件客户端包作为输入。

----结束

参考信息

批量升级客户端前，需手动配置远程登录客户端节点的用户密码信息：

执行**vi client-info.cfg**命令，添加用户密码信息。

例如：

```
clientIp,clientPath,user,password
10.10.10.100,/home/omm/client /home/omm/client2,omm,密码
```

配置文件各字段含义如下：

- clientIp：表示客户端所在节点IP地址。

- clientPath：客户端安装路径，可以包含多个路径，以空格分隔多个路径。注意路径不要以“/”结尾。
- user：节点用户名。
- password：节点用户密码信息。

  说明

  - password为密码。
  - 如果执行失败，请在执行目录的**work_space/log_*XXX***下查看**node.log**日志。

10.3.1.8.3 批量刷新 hosts 文件

操作场景

在FusionInsight Manager界面上下载的客户端包中包含客户端批量升级工具，该工具在提供批量升级客户端功能的同时，也提供了轻量级的批量刷新客户端所在节点“/etc/hosts”文件的功能。

前提条件

更新前准备请参考**批量升级客户端**章节“客户端升级前准备”步骤。

批量更新 hosts 文件

步骤1 检查需要更新“/etc/hosts”文件的节点的配置用户是否为“root”。

- 是，执行**步骤2**。
- 否，更改配置用户为“root”，再执行**步骤2**。

步骤2 执行**sh client_batch_upgrade.sh -r -f /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar -g /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig/batch_upgrade/client-info.cfg**，批量刷新客户端所在节点的“/etc/hosts”文件。

说明

- 执行批量刷新“/etc/hosts”文件时，输入的客户端包可以是完整客户端，也可以是仅包含配置文件的客户端软件包，推荐使用仅包含配置文件的客户端软件包。
- 需要更新“/etc/hosts”文件的主机所配置的用户必须为**root**用户，否则会刷新失败。

----结束

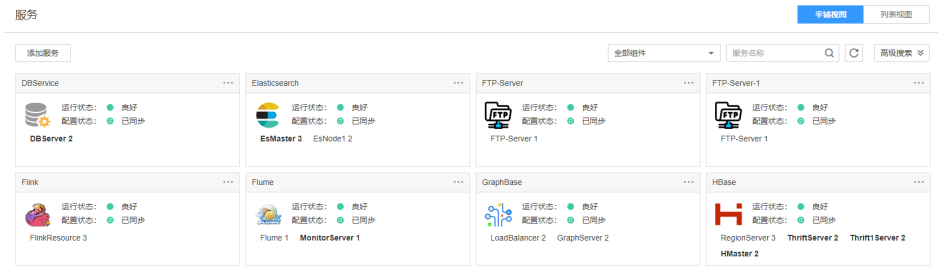
10.3.2 管理服务

10.3.2.1 服务管理概述

总览

登录FusionInsight Manager以后，选择“集群 > *待操作集群的名称* > 服务”后，打开服务管理页面，包含功能区和服务列表。

图 10-1 服务管理页面



功能区

服务管理页面的功能区支持选择视图类型，以及通过服务类型筛选和搜索服务。通过高级搜索可以根据“运行状态”和“配置状态”选择所需要的服务。

服务列表

服务管理页面的服务列表包含了集群中所有已安装的服务。如果选择“平铺视图”，则显示为窗格样式；如果选择“列表视图”，则显示为表格样式。

说明

本章节默认以“平铺视图”进行介绍。

服务列表可显示每个服务的运行状态、配置状态、角色的类型以及对应的实例个数。同时可以执行部分服务维护任务，例如启动、停止、重启服务等。

表 10-5 服务运行状态

状态	说明
良好	表示服务当前运行正常。
故障	表示服务当前无法正常工作。
亚健康	表示服务部分增强功能无法正常工作。
未启动	表示服务已停止。
未知	表示服务的初始状态信息无法检测。
正在启动	表示服务正在执行启动过程。
正在停止	表示服务正在执行停止过程。
启动失败	表示服务启动操作失败。
停止失败	表示服务停止操作失败。

说明

- 服务的运行状态为“故障”，会触发告警，请根据告警信息处理。
- HBase、Hive、Spark和Loader可显示“亚健康”（Subhealthy）状态。
  - Yarn已安装且不正常时，HBase处于“亚健康”状态。如启用多实例功能，则已安装的所有HBase服务实例处于“亚健康”状态。
  - HBase已安装且状态不正常时，Hive、Spark和Loader处于“亚健康”状态。
  - 启用多实例功能后，任意一个HBase服务实例已安装且不正常时，Loader处于“亚健康”状态。
  - 启用多实例功能后，某一个HBase服务实例已安装且不正常时，对应的Hive和Spark服务实例处于“亚健康”状态，即HBase2已安装且不正常时，Hive2和Spark2为“亚健康”状态。

表 10-6 服务配置状态

状态	说明
已同步	表示服务所有参数配置已在集群内全部生效。
配置过期	表示修改服务参数后，最新的配置未同步且未生效，需要同步配置且重启相应服务。可单击配置状态后的 图标查看过期的配置项。
失败	表示同步参数配置过程中出现通信或读写异常等操作。尝试使用“同步配置”恢复。
正在同步	表示正在同步服务参数配置。
未知	表示服务配置的初始状态信息无法检测。

服务列表中单击服务对应菜单，可对服务进行简单的维护管理操作，具体如表10-7所示。

表 10-7 基本维护管理功能

操作入口	说明
“启动服务”	启动集群中指定服务。
“停止服务”	将集群中指定服务停止。
“重启服务”	将集群中指定服务重启。 **说明** 某个服务可能被其他服务依赖，重启该服务则导致其他服务不可用，需要勾选“同时重启上层服务”。请根据对话框的服务列表确认是否可以执行操作，集群中由于依赖关系服务的重启为串行进行。单个服务的重启时长如表10-8所示。
“滚动重启服务”	为集群中指定服务提供不中断业务的重启操作，具体参数配置可参考表10-4。

操作入口	说明
“同步配置”	• 为集群中指定服务启用新的配置参数。 • 为集群中“配置状态”为“配置过期”的服务，下发新的配置参数。 **说明** 部分服务同步配置后需重启服务使配置生效。

表 10-8 重启时长

服务名称	重启时长	启动时长	附加说明
ClickHouse	4min	ClickHouseServer：2min ClickHouseBalancer：2min	-
HDFS	10min+x	NameNode：4min+x DataNode：2min JournalNode：2min Zkfc：2min	x为NameNode元数据加载时长，每千万文件大约耗时2分钟，例如5000万文件x为10min。由于受DataNode数据块上报影响启动时间有一定浮动。
Yarn	5min+x	ResourceManager：3min+x NodeManager：2min	x为ResourceManager保留任务数恢复时长，每1万保留任务大约需要1分钟
MapReduce	2min+x	JobHistoryServer：2min+x	x为历史任务扫描时长，每10万任务大约2.5min
Zookeeper	2min+x	quorumpeer：2min+x	x为加载znode节点时长，每100万znode大约1min
Hive	3.5min	HiveServer：3min MetaStore：1min30s WebHcat：1min Hive整体服务：3min	-

服务名称	重启时长	启动时长	附加说明
Spark2x	5min	JobHistory2x：5min SparkResource2x：5min JDBCServer2x：5min	-
Flink	4min	FlinkResource：1min FlinkServer：3min	-
Kafka	2min+x	Broker：1min+x	x为数据恢复时长，单实例20000 partition启动所需时长大约2mins。
Storm	6min	Nimbus：3mins UI：1min Supervisor：1min Logviewer：1min	-
Flume	3min	Flume：2 min MonitorServer：1min	-

10.3.2.2 其他服务管理操作

10.3.2.2.1 服务详情概述

总览

登录FusionInsight Manager以后，选择“集群 > *待操作集群的名称* > 服务”，在服务列表单击指定的服务名称打开服务详情页面，包含“概览”、“实例”、“实例组”和“配置”等页面，以及功能区。部分服务还支持显示自定义的管理工具页面，具体支持列表如表10-9所示。

表 10-9 自定义管理工具名称一览表

工具名称	对应服务	说明
Flume配置工具	Flume	用于为Flume的服务端和客户端配置采集参数。

工具名称	对应服务	说明
Flume客户端管理工具	Flume	查看Flume客户端监控信息。
Kafka Topic监控工具	Kafka	用于为Kafka的Topic提供监控与管理。

其中“概览”为默认页，包含基本信息、角色列表、依赖关系表和监控图表等，右上角可对服务进行管理，基本管理如启动、停止、滚动重启、同步配置请参考**表10-7**，其他服务管理操作如**表10-10**所示：

表 10-10 服务管理操作

操作入口	说明
“更多 > 健康检查”	为当前服务进行健康检查，健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“运行状态”。 健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。如果发现问题，可以单击“查看帮助”。
“更多 > 下载客户端”	为用户下载默认的仅包含具体服务的客户端，通过客户端执行管理操作、运行业务或进行二次开发，具体可参考**下载客户端**。
“更多 > 修改服务名称”	修改当前服务名称。
“更多 > 执行*角色名称*倒换”	具体请参考**执行角色实例主备倒换**。
“更多 > 进入维护模式/退出维护模式”	配置服务进入/退出维护模式。
“配置 > 导入/导出”	在迁移服务到新集群场景或者重新部署相同服务的场景下，为具体服务的所有配置数据做导入或者导出操作，实现配置结果的快速复制。

基本信息

“概览”的基本信息包含该服务的基本状态数据，即运行状态、配置状态、版本，还包含各个服务自身关键信息。如果服务支持开源WebUI，则在基本信息区域可通过WebUI的链接访问开源WebUI。

说明

当前版本“admin”用户没有权限访问服务的开源WebUI完整功能。请另外创建组件业务管理员并访问WebUI地址。

角色列表

“概览”页面的角色列表包含了该服务中所有的角色。角色列表可显示每个角色的运行状态和角色的实例个数。

依赖关系表

“概览”页面的依赖关系表支持展示该服务依赖的服务，以及依赖此服务的其他服务。

告警和事件的历史记录

告警和事件的历史记录区显示了当前服务上报的关键告警与事件记录，系统最大可显示20条历史记录。

图表

“概览”页面的右侧展示图表区，包含该服务的各个关键监控指标报表。用户可以自定义在图表区展示的监控报表、可以打开监控指标的解释说明或导出监控数据。对于定制类别为资源贡献类的图表，支持放大后切换趋势图和分布图。

说明

集群中部分服务提供服务级别的资源监控项，具体请参考**资源监控**。

10.3.2.2.2 执行角色实例主备倒换

操作场景

部分服务的角色以主备高可用的模式进行部署，在需要对主实例进行维护不能提供服务，或者其他维护需要时，可以手动触发实例主备倒换。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在服务详情页面单击“更多”，选择“执行*角色实例*倒换”。

步骤5 输入当前登录的用户密码确认身份，单击“确定”。

步骤6 在弹出界面单击“确定”，执行角色实例主备倒换。

说明

- Manager部件包支持的主备倒换角色实例的服务有：DBService。
- HD部件包支持的主备倒换角色实例的服务有：HDFS、Yarn、Storm、HBase、Mapreduce。
- HDFS的角色NameNode在进行主备倒换时，需要指定NameService。
- Porter部件包支持的主备倒换角色实例的服务有：Loader。
- 其他角色实例则不支持此功能。

----结束

10.3.2.2.3 资源监控

集群中部分服务提供服务级别的资源监控项，默认显示12小时的监控数据。用户可单击 自定义时间区间，缺省时间区间包括：12小时、1天、1周、1月。单击 可导出相应报表信息，无数据的监控项无法导出报表。支持资源监控的服务及监控项如**表10-11**所示。

登录FusionInsight Manager以后，选择“集群 > *待操作集群的名称* > 服务”后，选择待操作的服务，单击“资源”，进入资源监控页面。

表 10-11 服务资源监控

服务	监控指标	说明
HDFS	资源使用（按租户）	• 按租户统计HDFS的资源使用情况。 • 可选择按“容量”或“文件对象数”观察。
	资源使用（按用户）	• 按用户统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件对象数”观察。
	资源使用（按目录）	• 按目录统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件对象数”观察。 • 单击 配置空间监控，可以指定HDFS文件系统目录进行监控。
	资源使用（按副本）	• 按副本数统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件数”观察。
	资源使用（按文件大小）	• 按文件大小统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件数”观察。
	回收站（按用户）	• 按用户统计HDFS回收站的使用情况。 • 可选择按“回收站容量”或“文件对象数”观察。
	操作数	• 统计HDFS中操作数。
	自动balance	• 统计HDFS自动balancer的执行速度以及本次balancer当前迁移的总容量大小。
	NameNode RPC连接数（按用户）	• 按用户统计连接到NameNode的Client RPC请求中，各个用户的连接数。

服务	监控指标	说明
	慢DataNode节点	集群中数据传输或处理慢的DataNode节点。
	慢磁盘	集群中DataNode节点上数据处理慢的磁盘。
HBase	表级别操作请求次数	所有RegionServer上的所有表中put、delete、get、scan、increment、append操作请求次数。
	RegionServer级别操作请求次数	RegionServer中put、delete、get、scan、increment、append操作请求次数以及所有操作请求次数。
	服务级别操作请求次数	RegionServer上所有Region中put、delete、get、scan、increment、append操作请求次数。
	RegionServer级别HFile数	所有RegionServer中HFile数。
Hive	HiveServer2-Background-Pool线程数（按IP）	周期内统计并显示Top用户的HiveServer2-Background-Pool线程数。
	HiveServer2-Handler-Pool线程数（按IP）	周期内统计并显示Top用户的HiveServer2-Handler-Pool数监控。
	MetaStore使用数（按IP）	Hive周期内统计并显示Top用户的MetaStore使用数。
	Hive的Job数	Hive周期内统计并显示用户相关的Job数目。
	Split阶段访问的文件数	统计Hive周期内Split阶段访问底层文件存储系统（默认：HDFS）的文件数。
	Hive基本操作时间	Hive周期内统计底层创建目录（mkdirTime）、创建文件（touchTime）、写文件（writeFileTime）、重命名文件（renameTime）、移动文件（moveTime）、删除文件（deleteFileTime）、删除目录（deleteCatalogTime）所用的时间。
	表分区个数	Hive所有表分区个数监控，返回值的格式为：数据库#表名，表分区个数。
	HQL的Map数	Hive周期内执行的HQL与执行过程中调用的Map数统计，展示的信息包括：用户、HQL语句、Map数目。
	HQL访问次数	周期内HQL访问次数统计信息。
Kafka	Kafka磁盘使用率分布	Kafka集群的磁盘使用率分布统计。

服务	监控指标	说明
Spark2x	HQL访问次数	周期内HQL访问次数统计信息，展示信息包括用户名，HQL语句，执行该语句的次数。
Yarn	资源使用（按任务）	• 任务使用的CPU核数和内存。 • 可选择“按内存”或“按CPU”观察。
	资源使用（按租户）	• 租户所使用的CPU核数和内存。 • 可选择“按内存”或“按CPU”观察。
	资源使用比例（按租户）	• 租户所使用的CPU核数和内存的比例。 • 可选择“按内存”或“按CPU”观察。
	任务耗时排序	对Yarn任务耗时进行排序显示。
	ResourceManager RPC连接数（按用户）	统计连接到RM的Client RPC请求中，各个用户的连接数。
	操作数	统计Yarn每种操作类型对应的操作数及占比。
	队列中任务资源使用排序	• 在界面上选择某个队列（租户）后，显示在该队列中正在运行任务的消耗资源排序。 • 可选择“按内存”或“按CPU”观察。
	队列中用户资源使用排序	• 在界面上选择某个队列（租户）后，显示在该队列中正在运行任务的用户消耗的资源排序。 • 可选择“按内存”或“按CPU”观察。
ZooKeeper	资源使用（按二级Znode）	• ZooKeeper服务二级znode资源状况。 • 可选择“按Znode数量”或“按容量”观察
	连接数（按客户端IP）	ZooKeeper客户端连接资源状况。

10.3.2.2.4 采集堆栈信息

操作场景

为了满足实际业务的需求，管理员可以在FusionInsight Manager中采集指定角色或实例的堆栈信息，保存到本地目录，并支持下载。采集内容包括：

1. jstack栈信息。
2. jmap -histo堆统计信息。
3. jmap -dump堆信息快照。
4. 对于jstack和jmap-histo信息，支持连续采集以便对比。

操作步骤

采集堆栈

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > *待收集服务的名称*”。

步骤3 选择“更多 > 采集堆栈”。

说明

- 采集多个实例的堆栈信息：进入实例列表，勾选要采集的实例名称，选择“更多 > 采集堆栈”。
- 采集单个实例的堆栈信息：单击要采集的实例，选择“更多 > 采集堆栈”。

步骤4 根据界面提示，在弹框中选择需要采集的角色，采集内容，配置高级选项（若无特殊需求，保持默认配置即可），单击“确定”。

步骤5 采集成功后，单击“下载”。

下载堆栈信息

步骤6 选择“集群 > *待操作集群的名称* > 服务 > *待操作服务的名称*”。选择右上角“更多 > 下载堆栈信息”。

步骤7 选择需要下载的角色和内容，单击“下载”，可直接下载相关堆栈信息到本地。

清理堆栈信息

步骤8 选择“集群 > *待操作集群的名称* > 服务 > *待操作服务的名称*”。

步骤9 选择右上角“更多 > 清理堆栈信息”。

步骤10 选择需要清理的角色和内容，并配置“文件目录”。单击“确定”执行清理操作。

----结束

10.3.2.2.5 切换 Ranger 鉴权

操作场景

新安装的安全模式集群默认即安装了Ranger服务并启用了Ranger鉴权，用户可以通过组件的权限插件对组件资源的访问设置细粒度的安全访问策略。若不需使用Ranger进行鉴权，管理员可在服务页面手动停用Ranger鉴权，停用Ranger鉴权后，访问组件资源时系统将继续基于FusionInsight Manager的角色模型进行权限控制。

从历史版本升级的集群，用户访问组件资源时默认不使用Ranger鉴权，管理员可在安装了Ranger服务后手动启用Ranger鉴权。

说明

- 安全模式集群中，支持使用Ranger鉴权的组件包括：HDFS、Yarn、Kafka、Hive、HBase、Storm、Spark2x、Impala。
- 非安全模式集群中，Ranger可以支持基于OS用户进行组件资源的权限控制，支持启用Ranger鉴权的组件包括：HBase、HDFS、Hive、Spark2x、Yarn。
- 启用Ranger鉴权后，该组件所有鉴权将由Ranger统一管理，原鉴权插件设置的权限将会失效（HDFS与Yarn的组件ACL规则仍将生效），请谨慎操作，建议提前在Ranger上做好权限部署。
- 停用Ranger鉴权后，该组件所有鉴权将由组件自身权限插件管理，Ranger上设置的权限将会失效，请谨慎操作，建议提前在Manager上做好权限部署。

启用 Ranger 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在服务详情页面单击“更多”，选择“启用Ranger鉴权”。

步骤5 输入当前登录的用户密码确认身份，单击“确定”。

步骤6 在服务列表，重启配置过期的服务。

----结束

停用 Ranger 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在服务详情页面单击“更多”，选择“停用Ranger鉴权”。

步骤5 输入当前登录的用户密码确认身份，单击“确定”，在弹出框中单击“确定”。

步骤6 在服务列表，重启配置过期的服务。

----结束

10.3.2.3 服务配置

10.3.2.3.1 修改服务配置参数

操作场景

为了满足实际业务的需求，管理员可以在FusionInsight Manager中快速查看及修改服务默认的配置。请务必参照配置描述中的建议进行参数配置。

说明

集群中只剩下一个DBService角色实例时，不支持修改DBService服务的参数。

对系统的影响

- 配置服务属性后，需要重启此服务，重启期间该服务不可用。如果不重启，则服务“配置状态”为“配置过期”。
- 修改服务配置参数并重启生效后，需要重新下载并安装客户端，或者下载配置文件刷新客户端。例如HBase、HDFS、Hive、Spark、Yarn、Mapreduce。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 单击“配置”。

默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。

例如下图所示，第一个“LdapServer”表示服务名称，配置项针对整个服务；第二个“SlapdServer”表示角色名称，配置项针对角色的全部实例。

图 10-2 配置参数导航树

步骤5 在导航树选择指定的参数分类，并在右侧修改参数值。

说明

对于端口类参数值请从右侧描述中的取值范围中选取，请确保同一个服务中所有参数项配置的值均在取值范围内且唯一，否则会导致服务启动失败。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

步骤6 单击“保存”，并在确认对话框中单击“确定”。

等待界面提示“操作成功”，单击“完成”，配置已修改。

说明

- 更新Yarn服务队列的配置且不重启服务时，选择“更多 > 刷新队列”更新队列使配置生效。
- 配置Flume参数“flume.config.file”时，支持“上传文件”和“下载文件”功能。上传配置文件后旧文件将被覆盖，再下载文件只能获取新文件。如果未保存配置并重启服务，那么新文件设置未生效，请及时保存配置。
- 修改服务配置参数后如需重启服务使配置生效，可在服务页面单击右上角“更多 > 重启服务”。

----结束

10.3.2.3.2 修改服务自定义配置参数

操作场景

MRS集群各个组件支持开源的所有参数，其中部分关键使用场景的参数支持在FusionInsight Manager界面进行修改，且部分组件的客户端可能不包含开源特性的所有参数。如果需要修改其他Manager未直接支持的组件参数，管理员可以在Manager通过自定义配置项功能为组件添加新参数。添加的新参数最终将保存在组件的配置文件中并在重启后生效。

对系统的影响

- 配置服务属性后，需要重启此服务，重启期间该服务不可用。如果不重启，则服务“配置状态”为“配置过期”。
- 修改服务配置参数并重启生效后，需要重新下载并安装客户端，或者下载配置文件刷新客户端。

前提条件

管理员已充分了解需要新添加的参数意义、生效的配置文件以及对组件的影响。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 选择“配置 > 全部配置”。

步骤5 在左侧导航栏定位到某个一级节点，并选择“自定义”，Manager将显示当前组件的自定义参数。

“参数文件”显示保存管理员新添加的自定义参数的配置文件。每个配置文件中可能支持相同名称的开源参数，设置不同参数值后生效结果由组件加载配置文件的顺序决定。自定义参数支持服务级别与角色级别，请根据业务实际需要选择。不支持单个角色实例添加自定义参数。

步骤6 在对应参数项所在行“名称”列输入组件支持的参数名，在“值”列输入此参数的参数值。

支持单击“+”或“-”增加或删除一条自定义参数。

步骤7 单击“保存”，在弹出的“保存配置”窗口中确认修改参数，单击“确定”。界面提示“操作成功。”，单击“完成”，配置保存成功。

保存完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

任务示例（配置 Hive 自定义参数）

Hive依赖于HDFS，默认情况下hive访问HDFS使用的是HDFS的客户端，生效的配置参数统一由HDFS控制。例如HDFS参数“ipc.client.rpc.timeout”影响所有客户端连接HDFS服务端的RPC超时时间，如果管理员需要单独修改Hive连接HDFS的超时时间，

可以使用自定义配置项功能进行设置。在Hive的“core-site.xml”文件增加此参数可被Hive服务识别并代替HDFS的设置。

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”。

步骤2 选择“Hive > 配置 > 全部配置”。

步骤3 在左侧导航栏选择Hive服务级别“自定义”，Manager将显示Hive支持的服务级别自定义参数。

步骤4 在“core-site.xml”对应参数“core.site.customized.configs”的“名称”输入“ipc.client.rpc.timeout”，“值”输入新的参数值，例如“150000”。单位为毫秒。

步骤5 单击“保存”，在弹出的“保存配置”窗口中确认修改参数并单击“确定”。界面提示“操作成功。”，单击“完成”，配置保存成功。

保存完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

10.3.3 管理实例

10.3.3.1 实例管理概述

总览

登录FusionInsight Manager以后，例如选择“集群 > *待操作集群的名称* > 服务 > KrbServer > 实例”，进入实例管理页面，包含功能区和角色实例列表。

功能区

在功能区勾选需要操作的实例后，可对角色实例执行相关维护管理任务，例如启动或停止实例等，主要操作如**表10-12**所示。

表 10-12 实例维护管理功能

操作入口	说明
“ 启动实例”	将集群中指定实例启动。适用于操作状态为“未启动”、“停止失败”或“启动失败”角色实例，以使用该角色实例。
“ 更多 > 停止实例”	将集群中指定实例停止。适用于不再使用或异常的角色实例。
“ 更多 > 重启实例”	将集群中指定实例重启。适用于状态异常的角色实例，以恢复角色实例功能。
“ 更多 > 滚动重启实例”	为集群中指定实例提供不中断业务的重启操作，具体参数配置可参考**滚动重启集群**。

操作入口	说明
“ 更多 > 入服/退服”	为集群中指定实例执行入服务或退服的操作，变更实例的业务可用状态方式，具体可参考**入服与退服实例**。 **说明** 仅HDFS的角色DataNode、Yarn的角色NodeManager、HBase的角色RegionServer支持此操作。
“*待操作实例名称* > 更多 > 同步配置”	当某个角色实例的“配置状态”为“配置过期”，表示该角色实例修改配置后还未重启生效，新的配置仅保存在FusionInsight Manager。将新的配置下发至指定实例。 **说明** • 同步角色实例配置后需要重启配置过期的角色实例。重启时对应的角色实例不可用。 • 完成同步配置后，完成后请重启实例以使配置生效。
“*待操作实例名称* > 实例配置”	具体请参考**管理实例配置**。

功能区支持按角色或运行状态进行快速筛选。

说明

单击“高级搜索”，支持指定其他筛选条件搜索指定的实例，例如主机名称、管理IP、业务IP和实例组等。

角色实例列表

角色实例列表包含了该服务中所有的角色在集群中的实例情况，列表可显示每个实例的运行状态、配置状态、实例对应的主机以及相关的IP地址信息等。

表 10-13 实例运行状态

状态	说明
良好	表示实例当前运行正常。
故障	表示实例当前无法正常工作。
已退服	表示实例处于退服状态。
未启动	表示实例已停止。
未知	表示实例的初始状态信息无法检测。
正在启动	表示实例正在执行启动过程。
正在停止	表示实例正在执行停止过程。
正在恢复	表示实例可能存在异常正在自动修复。
正在退服	表示实例正在执行退服过程。

状态	说明
正在入服	表示实例正在执行入服过程。
启动失败	表示实例启动操作失败。
停止失败	表示实例停止操作失败。

实例详情

单击实例名称可进入实例详情页面，可查看实例基本信息、配置文件、实例日志以及该实例相关的监控指标图表。

10.3.3.2 入服与退服实例

操作场景

部分角色实例以分布式并行工作的方式对外部业务提供服务，服务会单独保存每个实例是否可以使用的信息，所以需要使用FusionInsight Manager为这些实例执行入服或退服的操作，变更实例的业务可用状态方式。

不支持该此功能的实例，默认无法执行任务。

说明

当前支持退服和入服操作的角色有：HDFS的DataNode、Yarn的NodeManager、HBase的RegionServer。

- 当DataNode数量少于或等于HDFS的副本数时，不能执行退服操作。若HDFS副本数为3时，则系统中少于4个DataNode，将无法执行退服，Manager在执行退服操作时会等待30分钟后报错并退出执行。
- 由于Mapreduce任务执行时，会生成一些副本数为10的文件，此时若DataNode实例数少于10时，将无法进行退服操作。
- 如果退服前，DataNode节点的机架数（机架数由各DataNode节点所配置的“机架”的名称数量决定）大于1；而退服部分DataNode后，剩余的DataNode节点的机架数变为1，则此次退服将会失败。所以需要在退服前评估退服操作对机架数的影响，以调整退服的DataNode节点。
- 在退服多个DataNode时，如果每个DataNode存储的数据量较大，如果执行选择多个DataNode同时退服，则很有可能会因超时而退服失败。为了避免这种情况，建议每次退服仅退服1个DataNode，进行多次退服操作。

操作步骤

步骤1 DataNode节点退服前需要进行健康检查，步骤如下：

1. 使用客户端用户登录客户端安装节点，并切换到客户端安装目录。
2. 如果是安全集群，需要使用**hdfs**用户进行权限认证。

```
source bigdata_env              #配置客户端环境变量
kinit hdfs                      #设置kinit认证
Password for hdfs@HADOOP.COM:    #输入hdfs用户登录密码
```

3. 执行**hdfs fsck / -list-corruptfileblocks**，检查返回结果。
   - 如果结果是“...has 0 CORRUPT files”，执行**步骤2**。
   - 如果结果不是“...has 0 CORRUPT files”，并返回损坏的文件名称，执行**步骤1.4**。

4. 执行**hdfs dfs -rm** *损坏的文件名称*，删除损坏的文件。

说明

删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。

步骤2 登录FusionInsight Manager。

步骤3 选择“集群 > *待操作集群的名称* > 服务”。

步骤4 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤5 勾选指定的待退服角色实例。

步骤6 在“更多”选择“退服”或“入服”。

输入当前登录的用户密码确认身份，单击“确定”。

勾选“我确定退服这些实例，并接受服务性能下降的结果。”，单击“确定”，执行相应的操作。

说明

实例退服操作未完成时在其他浏览器或窗口重启集群中实例对应的服务，FusionInsight Manager将提示停止退服，实例的“操作状态”显示为“启动”。实际上后台已将该实例退服，请重新执行退服操作同步状态。

----结束

10.3.3.3 管理实例配置

操作场景

每个单独的角色实例可以修改配置参数在迁移实例到新集群场景或者重新部署相同服务的场景下，管理员可以在FusionInsight Manager中将某服务所有配置数据导入或者导出，实现配置结果的快速复制。

FusionInsight Manager支持管理单个角色实例的配置参数，修改配置参数、导出实例配置或导入实例配置时不影响其他实例。

对系统的影响

修改角色实例配置后，需要重启此实例。重启时对应的实例不可用。如果不重启，则实例“配置状态”为“配置过期”。

修改实例配置

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作的集群名称* > 服务”。

步骤3 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤4 单击指定的实例，选择“实例配置”。

默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该实例支持的所有参数分类。

步骤5 在导航树选择指定的参数分类，并在右侧修改参数值。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

步骤6 单击“保存”，并在确认对话框中单击“确定”。

等待界面提示“操作成功”，单击“完成”，配置已修改。

----结束

导出导入实例配置

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤4 单击指定的实例，选择“实例配置”。

步骤5 单击“导出”，导出配置参数文件到本地。

步骤6 在实例配置页面单击“导入”，在弹出的配置文件选择框中定位到实例的配置参数文件，即可导入所有配置。

----结束

10.3.3.4 查看实例配置文件

操作场景

FusionInsight Manager支持在管理页面上直接查看实例节点上实际的环境变量、角色配置等配置文件内容，运维人员在需要快速排查实例对应配置项是否配置错误或者需要查看部分隐藏类型的配置项时，可直接在FusionInsight Manager上进行查看，帮助用户快速分析配置问题。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作的集群名称* > 服务”。

步骤3 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤4 单击需要查看配置的实例名称，在概览页面的“配置文件”区域内，系统会显示该实例相关的配置文件列表。

图 10-3 查看实例配置文件

步骤5 单击要查看的配置文件的名称，可查看配置文件内具体的配置参数值内容。

如需获取该配置文件，可单击“下载至本地”按钮，将该配置文件内容下载到本地PC。

说明

集群内的节点故障时，将无法查看配置文件，请修复故障的节点后再查看。

----结束

10.3.3.5 实例组

10.3.3.5.1 管理实例组

操作场景

FusionInsight Manager支持对多个实例组的管理功能，即用户可以按照具有相同硬件配置的节点或者其他原则将同一角色内的多个实例进行分组。针对实例组进行的配置参数修改，将同时对组内所有的实例生效。

在大集群场景中，通过实例组将提升大集群下异构环境批量实例的管理能力，分配好实例组后，后续可反复配置，减少实例配置项的冗余，提升系统性能。

创建实例组

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 选择“实例组”。

单击，按照界面提示填写参数。

表 10-14 实例组配置参数

参数名	说明
组名称	实例组名称只能包含字母、数字、下划线（_）、中划线（-）和空格，仅以字母、数字、下划线（_）或中划线（-）开头，只能在中间包含空格，并且不能超过99个字符。
角色	表示实例组包含哪个角色的实例。
复制源	指从指定的实例组复制配置值到新组，若为空，则新组对应的各配置值为系统默认值。
描述	只可以包含汉字、英文字母、数字、中英文逗号、中英文句号、下划线（_）、空格和换行符，并且不能超过200个字符。

说明

- 每个实例必须且只能属于一个实例组，实例首次安装时默认属于的实例组为“*角色名*-DEFAULT”。
- 多余或者不再使用的实例组可以删除，删除前需要将组内的实例全部迁移至其他实例组，然后参照**删除实例组**对实例组进行删除，系统默认的实例组不可删除。

步骤5 单击“确定”完成创建实例组。

----结束

修改实例组属性

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在“实例组”页签定位到指定的实例组。

单击，按照界面提示填写参数。

步骤5 单击“确定”完成修改。

默认实例组不支持修改。

----结束

删除实例组

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在“实例组”页签定位到指定的实例组。

步骤5 单击  。

步骤6 在弹出窗口单击“确定”。

默认实例组不支持删除。

----结束

10.3.3.5.2 查看实例组信息

操作场景

管理员可以在FusionInsight Manager查看指定服务的实例组。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 单击“实例组”。

步骤5 在导航栏选择一个角色，在“基本”页签，查看该实例组的全部实例。

说明

需要将某个实例从一个实例组移动到另一个实例组中时，可以根据以下操作：

1. 勾选需要移动到新实例组的实例，然后单击“移动”。
2. 在弹出窗口选择一个目标的实例组。
   迁移时将自动继承新实例组的配置，如果该实例之前修改过配置，将以自身的配置优先。
3. 单击“确定”。
   完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

10.3.3.5.3 配置实例组参数

操作场景

在大集群场景中，用户可以在FusionInsight Manager通过实例组可批量配置多个实例的参数，减少实例配置项的冗余，提升系统性能。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 选择“实例组”。

步骤5 在导航栏选择一个角色的实例组名称，切换至“配置”页签，调整需要修改的配置参数后单击“保存”，相关配置将对实例组内所有实例生效。

----结束

10.4 主机

10.4.1 主机管理页面

10.4.1.1 查看主机列表

总览

登录FusionInsight Manager以后，单击“主机”后，打开主机管理页面，可查看主机列表及基本信息。

用户可切换视图类型，以及设置条件筛选和搜索主机。

切换视图

单击“角色视图”，可直观查看各主机上当前已部署的角色。如果该角色支持主备模式，则角色名称显示为加粗。

主机列表

主机管理页面的主机列表包含了所有集群中所有主机，并支持对主机进行相关运维操作。

在主机管理页面，可通过节点类型或所属集群筛选主机，对主机类型的筛选规则为：

- 管理节点为部署了OMS的节点，同时管理节点上也可能部署控制角色和数据角色。
- 控制节点为部署控制角色的节点，同时控制节点上也可能部署数据角色。
- 数据节点为仅部署数据角色的节点。

系统默认为“主机视图”，可显示每个主机的IP地址信息、机架规划信息、AZ信息、运行状态、所归属集群以及硬件资源等使用情况。

表 10-15 主机运行状态

状态	说明
良好	表示主机当前状态正常。
故障	表示主机当前无法正常工作。
未知	表示主机的初始状态信息无法检测。
已隔离	表示主机处于隔离的状态。
已停机	表示主机处于停机的状态。

10.4.1.2 查看主机概览

总览

登录FusionInsight Manager以后，单击“主机”，在主机列表单击指定的主机名称，可以访问主机详情页面，主要包含基本信息区、磁盘状态区、角色列表区和监控图表等。

基本信息区

主机详情页面的基本信息包含该主机的各个关键信息，例如管理IP地址、业务IP地址、主机类型、机架、防火墙、CPU核数、操作系统等信息。

磁盘状态区

磁盘状态区包含了该主机所有为集群配置的磁盘分区，并显示每个磁盘分区的使用情况。

实例列表区

实例列表区显示了该主机所有安装的角色实例，并显示每个角色实例的状态，单击角色实例名称后的日志文件，可在线查看该实例对应日志文件内容。

告警和事件的历史记录

告警和事件的历史记录区显示了当前主机上报的关键告警与事件记录，系统最大可显示20条历史记录。

图表

主机详情页面的右侧展示图表区，包含该主机的各个关键监控指标报表。

用户可以单击右上角的“ ▼ > 定制”，自定义在图表区展示的监控报表。选择时间区间后，单击“ ▼ > 导出”，可以导出指定时间区间内的详细监控指标数据。

单击监控指标标题后的可以打开监控指标的解释说明。

单击主机的“图表”页签，可直接查看该主机的全量监控图表信息。

GPU 卡状态区

主机有配置GPU卡时，GPU卡状态区显示了当前主机安装的GPU卡型号、位置及状态信息。

10.4.1.3 查看主机进程及资源

总览

登录FusionInsight Manager页面，单击“主机”，在主机列表中选择指定的主机名称，进入主机详情页面，单击“进程”和“资源”页签进入相关页面。

主机进程

进程页面显示了当前主机上已部署服务实例的角色进程信息，例如进程状态、PID、进程运行时间等，并可直接在线查看各进程的日志文件内容。

主机资源

主机资源页面显示了当前主机上已部署服务实例的详细资源使用情况，包括CPU，内存，磁盘和端口情况。

10.4.2 主机维护操作

10.4.2.1 启动、停止主机上的所有实例

操作场景

当主机发生故障状态异常时，用户可能需要停止主机上的所有角色，对主机进行维护检查。故障清除后，启动主机上的所有角色恢复主机业务。Manager支持在主机管理页面或者主机详情页面进行相关操作，以下根据主机管理页面为例进行指导。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待操作主机前的复选框。

步骤4 在“更多”选择“启动所有实例”或“停止所有实例”执行相应操作。

----结束

10.4.2.2 执行主机健康检查

操作场景

如果某个主机节点的运行状态不是良好，用户可以执行主机健康检查，快速确认某些基本功能是否存在异常。在日常运维中，管理员也可以执行主机健康检查，以保证主机上各角色实例的配置参数以及监控没有异常、能够长时间稳定运行。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待操作主机前的复选框。

步骤4 在“更多”选择“健康检查”启动任务。

健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。如果发现问题，可以单击“查看帮助”。

----结束

10.4.2.3 分配机架

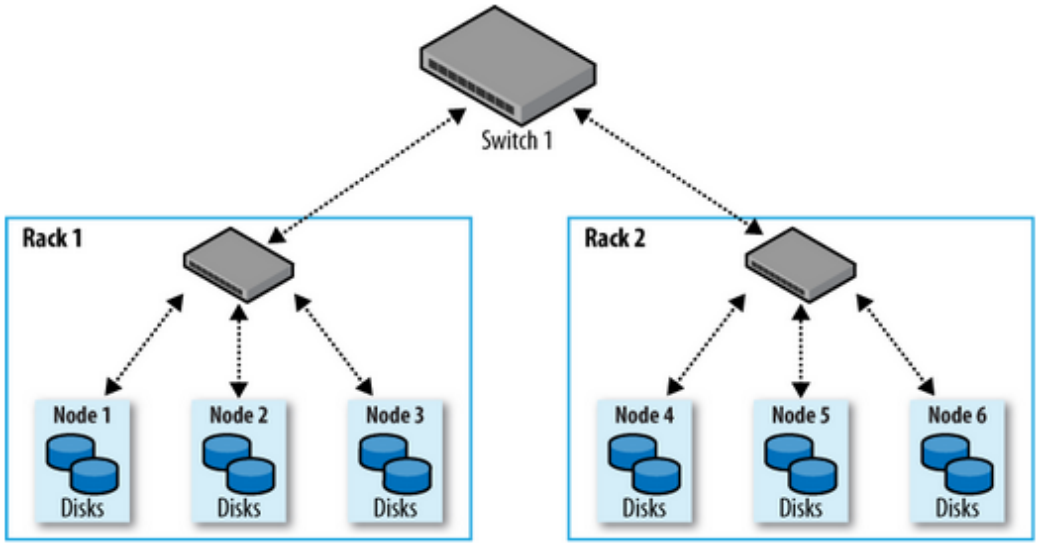
操作场景

大型集群的所有主机通常分布在多个机架上，不同机架间的主机通过交换机进行数据通信，且同一机架上的不同机器间的网络带宽要远大于不同机架机器间的网络带宽。在这种情况下网络拓扑规划应满足以下要求：

- 为了提高通信速率，希望不同主机之间的通信能够尽量发生在同一个机架之内，而不是跨机架。
- 为了提高容错能力，分布式服务的进程或数据需要尽可能存在多个机架的不同主机上。

Hadoop使用一种类似于文件目录结构的方式来表示主机。两层网络的集群如图10-4所示，Node1的Rack建议设置为**/Switch1/Rack1**，Node4的Rack建议设置为**/Switch1/Rack2**。

图 10-4 两层网络结构



由于HDFS不能自动判断集群中各个DataNode的网络拓扑情况，管理员需设置机架名称来确定主机所处的机架，NameNode才能绘出DataNode的网络拓扑图，并尽可能将DataNode的数据备份在不同机架中。同理，YARN需要获取机架信息，在可允许的范围内将任务分配给不同的NodeManager执行。

当集群网络拓扑发生变化时，需要使用FusionInsight Manager为主机重新分配机架，相关服务才会自动调整。

对系统的影响

修改主机机架名称，将影响HDFS的副本存放策略、Yarn的任务分配及Kafka的Partition存储位置。修改后需重启HDFS、Yarn和Kafka，使配置信息生效。

不合理的机架配置会导致集群的节点之间的负载（包括CPU、内存、磁盘、网络）不平衡，降低集群的可靠性，影响集群的稳定运行。所以在分配机架之前，需要进行全局的统筹，合理地设置机架。

机架分配策略

说明

物理机架：主机所在的真实的机架。

逻辑机架：在FusionInsight Manager中给主机设置的机架名称。

策略 1：每个逻辑机架包含的主机个数基本一致。

策略 2：主机所设置的逻辑机架要尽量符合其所在的物理机架。

策略 3：如果一个物理机架的主机个数很少，则需要和其他的主机较少的物理机架合并为一个逻辑机架，以满足策略1。不能将两个机房的主机合并为一个逻辑机架，否则会引起性能问题。

策略 4：如果一个物理机架的主机个数很多，则需要将其分隔为多个逻辑机架，以满足策略1。不建议物理机架中包含的主机有太大的差异，这样会降低集群的可靠性。

策略 5：建议机架的第一层为默认的“default”或其他值，但在集群中保持一致。

策略 6：每个机架所包含的主机个数不能小于3。

策略 7：一个集群的逻辑机架数，不建议多于50个（过多则不便于维护）。

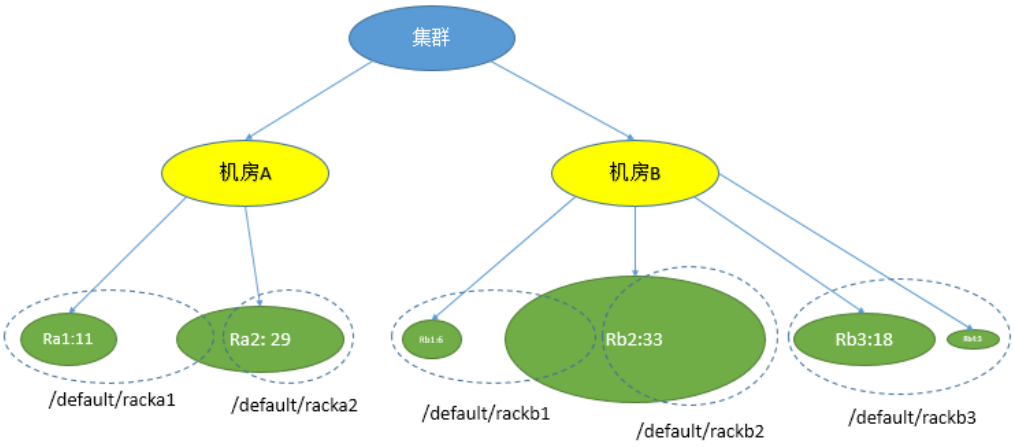
最佳实践示例

假设一个集群，共有主机100台，分别在两个机房中：机房A有40台主机，机房B有60台主机。在机房A中，物理机架Ra1有11台主机，物理机架Ra2有29台。在机房B中，物理机架Rb1有6台主机，物理机架Rb2有33台主机，物理机架Rb3有18台主机，物理机架Rb4有3台主机。

根据以上的“机架分配策略”，设置每个逻辑机架包含20个主机，具体分配如下：

- 逻辑机架 /default/racka1: 包含物理机架Ra1的11台主机，Ra2的9台主机。
- 逻辑机架 /default/racka2: 包含物理机架Ra2的剩余的20台主机。
- 逻辑机架 /default/rackb1: 包含物理机架Rb1的6台主机，Rb2的13台主机。
- 逻辑机架 /default/rackb2: 包含物理机架Rb2的剩余的20台主机。
- 逻辑机架 /default/rackb3: 包含物理机架Rb3的18台主机，Rb4的3台主机。

机架划分示例如下：



操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待操作主机前的复选框。

步骤4 在“更多”选择“设置机架”。

- 机架名称需遵循实际网络拓扑结构，以层级形式表示；各层级间以斜线“/”隔开。
- 机架命名规则为：“/level1/level2/…”，级别至少为一级，名称不能为空。机架名称由字母、数字及下划线“_”组成，且总长度不超过200个字符。

  例如“/default/rack0”。
- 如果待修改机架中所包含的主机中有DataNode实例，请确保所有DataNode实例所在主机的机架名称的层级一致。否则，会导致配置下发失败。

步骤5 单击“确定”，完成机架分配设置。

----结束

10.4.2.4 隔离主机

操作场景

某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。

说明

隔离主机仅支持隔离非管理节点。

对系统的影响

- 主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。
- 主机隔离后部分服务的实例不再工作，服务的配置状态可能过期。
- 主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。
- 待操作节点的SSH端口需保持默认（22），否则将导致本章节任务操作失败。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待隔离主机前的复选框。

步骤4 在“更多”选择“隔离”。

在弹出窗口中，输入当前登录的用户密码确认管理员身份，单击“确定”。

步骤5 在确认隔离的对话框中勾选“我确定隔离所选主机，接受可能出现的服务故障等后果。”单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功隔离，“运行状态”显示为“已隔离”。

步骤6 以**root**用户登录到被隔离主机上，执行**pkill -9 -u omm**命令终止节点上的omm用户的进程，然后执行**ps -ef | grep 'container' | grep '${BIGDATA_HOME}' | awk '{print $2}' | xargs -I '{}' kill -9 '{}'** 命令查找并终止container的进程。

步骤7 管理员已排除主机的异常或故障后，需要将主机隔离状态取消才能继续使用该主机。

在“主机”界面勾选已隔离的主机，选择“更多 > 取消隔离”。

说明

取消隔离后，主机上所有角色实例默认不启动。若需要启动主机上角色实例，可以在“主机”页面勾选目标主机，然后选择“更多 > 启动所有实例”。

----结束

10.4.2.5 导出主机信息

操作场景

管理员可以在FusionInsight Manager导出所有主机的信息。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 在右上角的下拉菜单中选择所需主机的类型，也可以通过“高级搜索”进一步筛选所需主机。

步骤4 单击“导出全部”，在“保存类型”选择“TXT”或“CSV”。单击“确定”开始导出。

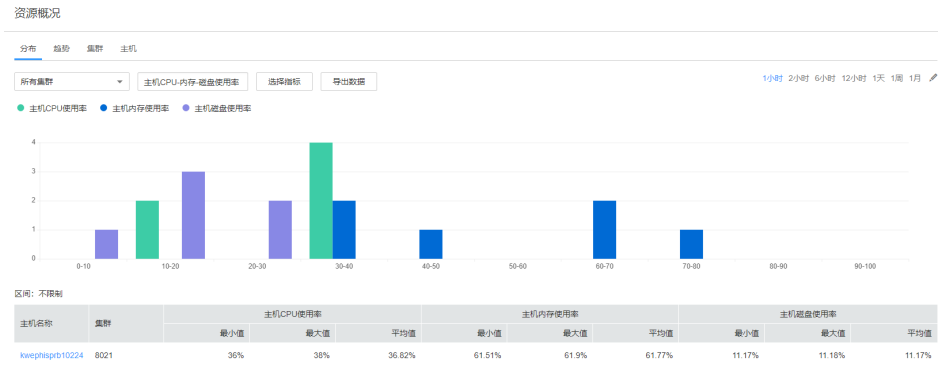
----结束

10.4.3 资源概况

10.4.3.1 分布

选择“主机 > 资源概况 > 分布”，可查看各集群的资源分布监控，如图10-5所示，默认显示1小时的监控数据。用户可单击 自定义时间区间，缺省时间区间包括：1小时、2小时、6小时、12小时、1天、1周、1月。

图 10-5 资源分布概况



- 单击“选择指标”可以自定义所需查看的指标项，详细指标项如**表10-16**所示。选择指标后，页面会显示在各个区间的主机分布图。
- 鼠标停留在某个色块时，会显示处于当前区间的主机数量，如**图10-5**所示。单击色块，页面会显示处于当前区间的主机列表。
  - 单击列表中某主机“主机名称”，会跳转至该主机的详细信息页面；
  - 单击列表中某主机“查看趋势”，会显示当前指标项整个集群的最大值、平均值、最小值、当前主机值。当前集群中，当指标为“主机CPU-内存-磁盘使用率”时，不能进行“查看趋势”操作。
- 单击“导出数据”，可以导出当前指标项集群中所有节点在选中的时间区域内的最大值、最小值、平均值。

表 10-16 指标项

指标分类	指标项
进程	• 运行的进程总数 • 进程总数 • omm进程总数 • D状态进程总数

指标分类	指标项
网络状态	• 主机网络数据包冲突数 • LAST_ACK状态数量 • CLOSING状态数量 • LISTENING状态数量 • CLOSED状态数量 • ESTABLISHED状态数量 • SYN_RECV状态数量 • TIME_WAITING状态数量 • FIN_WAIT2状态数量 • FIN_WAIT1状态数量 • CLOSE_WAIT状态数量 • DNS解析时长 • TCP临时端口使用率 • 主机网络数据包帧错误数
网络读信息	• 主机网络读包数 • 主机网络读包丢包数 • 主机网络读包错误数 • 主机网络接收速率
磁盘	• 主机磁盘写速率 • 主机磁盘已使用大小 • 主机磁盘未使用大小 • 主机磁盘读速率 • 主机磁盘使用率
内存	• 未使用内存 • 缓存内存大小 • 内核缓存的内存总量 • 共享内存大小 • 主机内存使用率 • 已使用内存
网络写信息	• 主机网络写包数 • 主机网络写包错误数 • 主机网络发送速率 • 主机网络写包丢包数

指标分类	指标项
CPU	• 改变过优先级的进程占CPU的百分比 • 用户空间占用CPU百分比 • 内核空间占用CPU百分比 • 主机CPU使用率 • CPU总时间 • CPU闲置时间
主机状态	• 主机文件句柄使用率 • 每1分钟系统平均负载 • 每5分钟系统平均负载 • 每15分钟系统平均负载 • 主机PID使用率

10.4.3.2 趋势

选择“主机 > 资源概况 > 趋势”，可查看所有集群或者单个集群的资源趋势监控页面，如图10-6所示。默认显示1小时的监控数据。用户可单击 自定义时间区间，缺省时间区间包括：1小时、2小时、6小时、12小时、1天、1周、1月。各指标趋势图默认显示整个集群的最大值、最小值、平均值。

图 10-6 资源趋势

- 单击“为图表添加主机”，可在定制显示的趋势指标图中，添加个别节点的指标趋势线，最多可添加12个主机。
- 单击 ▼ ，选择“定制”，可以自定义需要在页面上显示的指标项，详细指标项参考分布中表10-16。
- 选择“导出数据”，可以导出集群中所有节点，在所有选中的指标项下，选中时间范围内的最大值、最小值、平均值。

10.4.3.3 集群

选择“主机 > 资源概况 > 集群”，可同时查看FusionInsight Manager内各集群的资源监控页面，如图10-7所示。

系统默认显示1小时的监控数据。用户可单击 自定义时间区间，缺省时间区间包括：1小时、2小时、6小时、12小时、1天、1周、1月。

图 10-7 集群资源概况

- 单击“指定集群”，可定制需要显示的集群。
- 单击 ，选择“定制”，可以自定义需要在页面上显示的指标项，详细指标项参考分布中表10-16。
- 选择“导出数据”，可以导出各个集群在所有选中的指标项下，选中时间范围内的指标值。

10.4.3.4 主机

选择“主机 > 资源概况 > 主机”，可查看主机资源概况，分为基础配置（CPU/内存）和磁盘配置两部分，如图10-8所示。

单击“导出数据”，可导出集群中所有主机的配置列表，包括主机名称、管理IP、主机类型、核数、平台类型、内存容量、磁盘大小等。

图 10-8 主机资源概况

基础配置（CPU/内存）

鼠标放置饼图上会显示当前区域集群中各节点不同硬件配置下的配置信息及数量，格式为：*核数（平台类型）内存大小：数量*。

单击相应区域，会在下方显示相应的主机列表。

磁盘配置

横轴为节点上磁盘总容量（包含OS盘），纵轴为逻辑磁盘数量（包含OS盘）。

鼠标放置圆点上会显示处于当前配置状态下的磁盘信息，包括磁盘数量、总容量、主机数。

单击相应圆点，会在下方显示相应的主机列表。

10.5 运维

10.5.1 告警

10.5.1.1 告警与事件概述

告警

登录FusionInsight Manager，选择“运维 > 告警 > 告警”，进入如图10-9所示的界面，用户可以查看FusionInsight Manager中各集群上报的告警信息，包括告警名称、ID、级别、产生时间等信息，默认每页显示最近的十条告警。

图 10-9 FusionInsight Manager 告警管理

单击指定告警名称左侧的，展开完整告警信息参数，各项说明如表10-17所示。

表 10-17 告警参数

告警参数	说明
告警ID	告警信息的ID。
告警名称	告警信息的名称。
告警级别	包含紧急、重要、次要、提示四项级别。
来源	集群名称。

告警参数	说明
清除时间	告警检测到已清除的时间。如果未清除，则显示为“--”。
对象	触发告警的服务、进程或模块。
是否自动清除	能够在问题修复后自动清除告警。
告警状态	告警当前状态，包含自动清除、手动清除、未清除。
产生时间	产生告警的时间。
告警原因	告警可能的原因提示。
序列号	系统产生的告警计数。
附加信息	相关报错信息。
定位信息	定位告警的详细信息。主要包含以下信息： ● 集群名：产品告警的集群 ● 服务名：产生告警的服务名称 ● 角色名：产生告警的角色名称 ● 主机名：产生告警的主机名

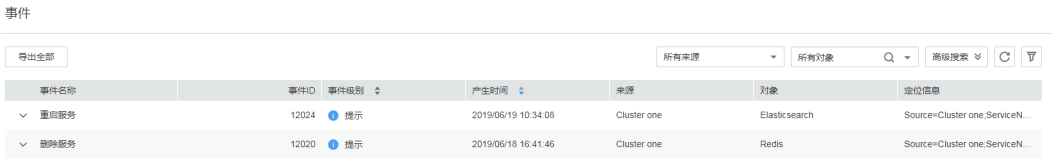
管理告警

- 单击“导出全部”可导出全部告警详情。
- 如果有多个告警已完成处理，可选中一个或多个待清除的告警，单击“清除告警”，批量清除告警。每次最多批量清除300条告警。
- 单击手动刷新当前页面，也可在 修改告警表格显示的列。
- 支持通过指定对象或集群来筛选指定的告警。
- 单击“高级搜索”显示告警搜索区域，搜索条件包括告警ID、告警名称、告警状态、告警级别、开始时间和结束时间。单击“搜索”显示过滤后的告警，再次单击“高级搜索”，会显示已经填写的搜索条件数量。
- 单个告警支持“清除”、“屏蔽”以及“查看帮助”操作。
- 告警条目较多时，可单击“归类视图”，系统会将未恢复的告警按照告警ID进行归类，方便用户查看。归类后单击告警名称后的未恢复条数，即可查看具体的告警详情。

事件

登录FusionInsight Manager，选择“运维 > 告警 > 事件”，进入事件界面，用户可以查看集群中所有事件信息，包括名称、ID、级别、产生时间、来源、对象、定位信息，每页默认显示最近的十条事件。

图 10-10 FusionInsight Manager 事件管理



单击指定事件名称左侧的，展开完整信息参数，各项说明如**表10-18**所示。

表 10-18 事件参数

事件参数	说明
事件ID	事件信息的ID。
事件名称	事件信息的名称。
事件级别	包含紧急、重要、次要、提示共4项级别。
产生时间	事件产生的时间。
对象	事件可能的原因提示。
序列号	系统产生的事件计数。
定位信息	定位事件的详细信息。主要包含以下信息： ● 来源：产生事件的集群名称 ● 服务名：产生事件的服务名称 ● 角色名：产生事件的角色名称 ● 主机名：产生事件的主机名
附加信息	相关报错信息。
事件原因	事件可能的原因提示。
来源	集群名称。

管理事件：

- 单击“导出全部”可导出全部事件详情。
- 单击手动刷新当前页面，也可在 修改事件表格显示的列。
- 支持通过指定对象或集群来筛选指定的事件。
- 单击“高级搜索”显示事件搜索区域，搜索条件包括事件ID、事件名称、事件级别、开始时间和结束时间。

10.5.1.2 配置阈值

操作场景

FusionInsight Manager支持配置监控指标阈值用于关注各指标的健康情况，如果出现异常的数据并满足预设条件后，系统将会触发一条告警信息，并在告警页面中出现此告警信息。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 告警 > 阈值设置”。

步骤3 在监控分类中选择集群内指定主机或服务的监控指标。

图 10-11 配置指标阈值

例如“主机内存使用率”，界面显示此阈值的信息：

- 发送告警开关指示为 表示将触发告警。
- “告警ID”和“告警名称”包含阈值将触发的告警信息。
- Manager会检查监控指标数值是否满足阈值条件，若连续检查且不满足的次数等于“平滑次数”设置的值则发送告警，支持自定义。
- “检查周期（秒）”表示Manager检查监控指标的时间间隔。
- 规则列表中的条目为触发告警的规则。

步骤4 单击“添加规则”，可以新增指标的监控行为。

表 10-19 监控指标规则参数

参数名	参数值	参数解释
规则名称	CPU_MAX（举例）	规则名称
告警级别	• 紧急 • 重要 • 次要 • 提示	告警级别 • 紧急 • 重要 • 次要 • 提示

参数名	参数值	参数解释
阈值类型	• 最大值 • 最小值	选择某指标的最大值或最小值，类型为“最大值”表示指标的实际值大于设置的阈值时系统将产生告警，类型为“最小值”表示指标的实际值小于设置的阈值时系统将产生告警。
日期	• 每天 • 每周 • 其他	设置规则生效的日期，即哪一天运行规则。
添加日期	09-30	仅在“日期”模式为“其他”时可见，设置规则运行的自定义日期，支持多选。
阈值设置	起止时间：00:00-8:30	设置规则运行的具体时间范围。
	阈值：10	设置规则监控指标的阈值

说明

支持单击 或 设置多个阈值时间条件。

步骤5 单击“确定”保存规则。

步骤6 在新添加规则所在的行，单击“操作”中的“应用”，此时规则的“生效状态”变成“生效”。

当前已创建的规则单击“取消应用”后，才能应用新规则。

----结束

监控指标参考

FusionInsight Manager转告警监控指标可分为节点信息指标与集群服务指标。**表10-20**表示节点中可配置阈值的指标。

表 10-20 节点信息监控指标转告警列表

监控指标组名称	监控指标名称	指标含义	默认阈值
CPU	主机CPU使用率	描述周期内当前集群的运算和控制能力，可通过观察该统计值，了解集群整体资源的使用情况。	90.0%
磁盘	磁盘使用率	描述主机磁盘的使用率。	90.0%
	磁盘inode使用率	统计采集周期内磁盘inode使用率。	80.0%
内存	主机内存使用率	统计当前时间点的内存平均使用率。	90.0%
主机状态	主机文件句柄使用率	统计采集周期内该主机的文件句柄使用率。	80.0%
	主机PID使用率	主机PID使用率。	90%
网络状态	TCP临时端口使用率	统计采集周期内该主机的TCP临时端口使用率。	80.0%
网络读信息	读包错误率	统计采集周期内该主机上该网口的读包错误率。	0.5%
	读包丢包率	统计采集周期内该主机上该网口的读包丢包率。	0.5%
	读吞吐率	统计周期内网口的平均读吞吐率（MAC层）。	80%
网络写信息	写包错误率	统计采集周期内该主机上该网口的写包错误率。	0.5%
	写包丢包率	统计采集周期内该主机上该网口的写包丢包率。	0.5%
	写吞吐率	统计周期内网口的平均写吞吐率（MAC层）。	80%
进程	D状态进程总数	统计周期内主机上D状态进程数量。	0

监控指标组名称	监控指标名称	指标含义	默认阈值
	omm进程使用率	统计周期内omm进程使用率。	90

表 10-21 集群监控指标转告警列表

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
DBService	数据库	数据库连接数使用率	数据库连接数使用率统计。	90%
		数据目录磁盘空间使用率	数据目录磁盘空间使用率统计。	80%
Flume	Agent	Flume堆内存使用率	Flume堆内存使用百分比统计。	95.0%
		Flume直接内存使用率	Flume直接内存使用百分比统计。	80.0%
		Flume非堆内存使用率	Flume非堆内存使用百分比统计。	80.0%
		Flume垃圾回收（GC）总时间	Flume垃圾回收（GC）总时间。	12000ms
HBase	GC	GC中回收old区所花时长	RegionServer的总GC时间。	5000ms
		GC中回收old区所花时长	HMaster的总GC时间。	5000ms
	CPU和内存	RegionServer直接内存使用率统计	RegionServer直接内存使用率统计。	90%
		RegionServer堆内存使用率统计	RegionServer堆内存使用率统计。	90%
		HMaster直接内存使用率统计	HMaster直接内存使用率统计。	90%
		HMaster堆内存使用率统计	HMaster堆内存使用率统计。	90%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	服务	单个RegionServer的region数目	单个RegionServer的Region数目。	2000
		处在RIT状态达到阈值时长的region数	处在RIT状态达到阈值时长的region数。	1
	容灾	容灾同步失败次数	同步容灾数据失败次数。	1
	队列	Compaction操作队列大小	Compaction操作队列大小。	100
HDFS	文件和块	HDFS缺失的块数量	HDFS文件系统中缺少副本块数量。	0
		需要复制副本的块总数	NameNode需要复制副本的块总数。	1000
	RPC	主NameNode RPC处理平均时间	NameNode RPC处理平均时间。	100ms
		主NameNode RPC队列平均时间	NameNode RPC队列平均时间。	200ms
	磁盘	HDFS磁盘空间使用率	HDFS磁盘空间使用率。	80%
		DataNode磁盘空间使用率	HDFS文件系统中DataNode可以使用的磁盘空间率。	80%
		总副本预留磁盘空间所占比率	总副本预留磁盘空间占DataNode总未使用磁盘空间的百分比。	90%
	资源	故障的DataNode总数	出故障的DataNode节点数量。	3
		NameNode非堆内存使用百分比统计	NameNode非堆内存使用百分比统计。	90%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		NameNode直接内存使用百分比统计	NameNode直接内存使用百分比统计。	90%
		NameNode堆内存使用百分比统计	NameNode堆内存使用百分比统计。	95%
		DataNode直接内存使用百分比统计	DataNode直接内存使用百分比统计。	90%
		DataNode堆内存使用百分比统计	DataNode堆内存使用百分比统计。	95%
		DataNode非堆内存使用百分比统计	DataNode非堆内存使用百分比统计。	90%
	垃圾回收	垃圾回收时间统计（GC）	NameNode每分钟的垃圾回收（GC）所占用的时间。	12000ms
		垃圾回收时间统计（GC）	DataNode每分钟的垃圾回收（GC）所占用的时间。	12000ms
Hive	HQL	Hive执行成功的HQL百分比	Hive执行成功的HQL百分比。	90.0%
	Background	Background线程使用率	Background线程使用率。	90%
	GC	MetaStore的总GC时间	MetaStore的总GC时间。	12000ms
		HiveServer的总GC时间	HiveServer的总GC时间。	12000ms
	容量	Hive已经使用的HDFS空间占可使用空间的百分比	Hive已经使用的HDFS空间占可使用空间的百分比。	85.0%
	CPU和内存	MetaStore直接内存使用率统计	MetaStore直接内存使用率统计。	95%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		MetaStore非堆内存使用率统计	MetaStore非堆内存使用率统计。	95%
		MetaStore堆内存使用率统计	MetaStore堆内存使用率统计。	95%
		HiveServer直接内存使用率统计	HiveServer直接内存使用率统计。	95%
		HiveServer非堆内存使用率统计	HiveServer非堆内存使用率统计。	95%
		HiveServer堆内存使用率统计	HiveServer堆内存使用率统计。	95%
	Session	连接到HiveServer的session数占最大允许session数的百分比	连接到HiveServer的session数占最大允许session数的百分比。	90.0%
Kafka	分区	未完全同步的Partition百分比	未完全同步的Partition数占Partition总数的百分比。	50%
	其他	Partition不可用百分比	Kafka各个Topic的Partition不可用占比。	40%
		broker上用户连接数使用率	broker上用户连接数使用率。	80%
	磁盘	Broker磁盘使用率	Broker数据目录所在磁盘的磁盘使用率。	80.0%
	进程	Broker每分钟的垃圾回收时间统计（GC）	Broker进程每分钟垃圾回收（GC）所占用的时间。	12000ms
		Kafka堆内存使用率	Kafka堆内存使用百分比统计。	95%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		Kafka直接内存使用率	Kafka直接内存使用百分比统计。	95%
Loader	内存	Loader堆内存使用率	Loader堆内存使用率。	95%
		Loader直接内存使用率统计	Loader直接内存使用率统计。	80.0%
		Loader非堆内存使用率	Loader非堆内存使用率。	80%
	GC	Loader的总GC时间	Loader的总GC时间。	12000ms
Mapreduce	垃圾回收	垃圾回收时间统计（GC）	垃圾回收时间统计（GC）。	12000ms
	资源	JobHistoryServer直接内存使用百分比统计	JobHistoryServer直接内存使用百分比统计。	90%
		JobHistoryServer非堆内存使用百分比统计	JobHistoryServer非堆内存使用百分比统计。	90%
		JobHistoryServer堆内存使用百分比统计	JobHistoryServer堆内存使用百分比统计。	95%
Oozie	内存	Oozie堆内存使用率	Oozie堆内存使用率。	95.0%
		Oozie直接内存使用率	Oozie直接内存使用率。	80.0%
		Oozie非堆内存使用率	Oozie非堆内存使用率。	80%
	GC	Oozie垃圾回收（GC）总时间	Oozie垃圾回收（GC）总时间。	12000ms
Spark2x	内存	JDBCServer2x堆内存使用率统计	JDBCServer2x堆内存使用率统计。	95%
		JDBCServer2x直接内存使用率统计	JDBCServer2x直接内存使用率统计。	95%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		JDBCServer2x非堆内存使用率统计	JDBCServer2x非堆内存使用率统计	95%
		JobHistory2x直接内存使用率统计	JobHistory2x直接内存使用率统计。	95%
		JobHistory2x非堆内存使用率统计	JobHistory2x非堆内存使用率统计。	95%
		JobHistory2x堆内存使用率统计	JobHistory2x堆内存使用率统计。	95%
		IndexServer2x直接内存使用率统计	IndexServer2x直接内存使用率统计。	95%
		IndexServer2x堆内存使用率统计	IndexServer2x堆内存使用率统计。	95%
		IndexServer2x非堆内存使用率统计	IndexServer2x非堆内存使用率统计。	95%
	GC次数	JDBCServer2x的Full GC次数	JDBCServer2x进程的Full GC次数。	12
		JobHistory2x的Full GC次数	JobHistory2x进程的Full GC次数。	12
		IndexServer2x的Full GC次数	IndexServer2x进程的Full GC次数。	12
	GC时间	JDBCServer2x的总GC时间	JDBCServer2x的总GC时间。	12000ms
		JobHistory2x的总GC时间	JobHistory2x的总GC时间。	12000ms
		IndexServer2x的总GC时间	IndexServer2x的总GC时间。	12000ms
Storm	集群	Supervisor数	统计周期内集群中可用的Supervisor数目。	1

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		已用Slot比率	统计周期内集群中可用的slot使用率。	80.0%
	Nimbus	Nimbus堆内存使用率	Nimbus堆内存使用百分比统计。	80%
Yarn	资源	NodeManager直接内存使用百分比统计	NodeManager直接内存使用百分比统计。	90%
		NodeManager堆内存使用百分比统计	NodeManager堆内存使用百分比统计。	95%
		NodeManager非堆内存使用百分比统计	NodeManager非堆内存使用百分比统计。	90%
		ResourceManager直接内存使用百分比统计	ResourceManager直接内存使用百分比统计。	90%
		ResourceManager堆内存使用百分比统计	ResourceManager堆内存使用百分比统计。	95%
		ResourceManager非堆内存使用百分比统计	ResourceManager非堆内存使用百分比统计。	90%
	垃圾回收	垃圾回收时间统计（GC）	NodeManager每分钟的垃圾回收（GC）所占用的时间。	12000ms
		垃圾回收时间统计（GC）	ResourceManager每分钟的垃圾回收（GC）所占用的时间。	12000ms
	其他	root队列下失败的任务数	root队列下失败的任务数。	50
		root队列下被杀死的任务数	root队列下被杀死的任务数。	50
	CPU和内存	挂起的内存量	挂起的内存量。	83886080MB

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	任务	正在挂起的任务	正在挂起的任务。	60
ZooKeeper	连接	ZooKeeper连接数使用率	ZooKeeper连接数使用百分比统计。	80%
	CPU和内存	ZooKeeper堆内存使用率	ZooKeeper堆内存使用百分比统计。	95%
		ZooKeeper直接内存使用率	ZooKeeper直接内存使用百分比统计。	80%
	GC	ZooKeeper每分钟的垃圾回收时间统计（GC）	ZooKeeper每分钟的垃圾回收时间统计（GC）。	12000ms
meta	OBS数据写操作	OBS数据写操作接口调用成功率	OBS数据写操作接口调用成功率。	99.0%
	OBS元数据操作	OBS元数据接口调用平均时间	OBS元数据接口调用平均时间。	500ms
		OBS元数据接口调用成功率	OBS元数据接口调用成功率。	99.0%
	OBS数据读操作	OBS数据读操作接口调用成功率	OBS数据读操作接口调用成功率。	99.0%
Ranger	GC	UserSync垃圾回收（GC）时间	UserSync垃圾回收（GC）时间。	12000ms
		RangerAdmin垃圾回收（GC）时间	RangerAdmin垃圾回收（GC）时间。	12000ms
		TagSync垃圾回收（GC）时间	TagSync垃圾回收（GC）时间。	12000ms
	CPU和内存	UserSync非堆内存使用率	UserSync非堆内存使用百分比统计。	80.0%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		UserSync直接内存使用率	UserSync直接内存使用百分比统计。	80.0%
		UserSync堆内存使用率	UserSync堆内存使用百分比统计。	95.0%
		RangerAdmin非堆内存使用率	RangerAdmin非堆内存使用百分比统计。	80.0%
		RangerAdmin堆内存使用率	RangerAdmin堆内存使用百分比统计。	95.0%
		RangerAdmin直接内存使用率	RangerAdmin直接内存使用百分比统计。	80.0%
		TagSync直接内存使用率	TagSync直接内存使用百分比统计。	80.0%
		TagSync非堆内存使用率	TagSync非堆内存使用百分比统计。	80.0%
		TagSync堆内存使用率	TagSync堆内存使用百分比统计。	95.0%
ClickHouse	集群配额	Clickhouse服务在ZooKeeper的数量配额使用率	ClickHouse服务在ZooKeeper上目录的数量配额使用百分比。	90%
		Clickhouse服务在ZooKeeper的容量配额使用率	ClickHouse服务在ZooKeeper上目录的容量配额使用百分比。	90%

10.5.1.3 配置告警屏蔽状态

操作场景

如果如下特定场景中不希望看到FusionInsight Manager上报指定的告警，可以手动设置屏蔽。

- 使用过程中，不想关注某些不重要的告警，屏蔽次要告警。
- 第三方产品集成FusionInsight产品时，部分告警与产品自身的告警信息重复，屏蔽重复告警。
- 部署环境特殊时，可能存在特定告警误报，屏蔽误报的告警。

某种告警被屏蔽后，与该告警ID相同的新告警将不再出现在“告警管理”页面中，也不会被统计。已经上报的告警仍然显示。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 告警 > 屏蔽设置”。

步骤3 在“屏蔽设置”区域，选择指定的服务或模块。

步骤4 在告警列表中选择指定的告警。

图 10-12 屏蔽告警

界面显示此告警的信息，包含名称、ID、级别、屏蔽状态和操作：

- 屏蔽状态包含：“屏蔽”和“显示”。
- 操作包含：“屏蔽”和“查看帮助”。

说明

在屏蔽列表上方可筛选指定的告警。

步骤5 设置已选中告警的屏蔽状态：

- 单击“屏蔽”后在弹出的对话框中单击“确定”，修改告警的屏蔽状态为“屏蔽”。
- 单击“取消屏蔽”后在弹出的对话框中单击“确定”，修改告警的屏蔽状态为“显示”。

----结束

10.5.2 日志

10.5.2.1 在线检索日志

操作场景

FusionInsight Manager支持在线检索并显示组件的日志内容，用于问题定位等其他日志查看场景。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 日志 > 在线检索”。

步骤3 根据所需查询日志分别填写表10-22各参数，用户可以根据需要选择所需查阅日志时长，缺省时间区间包括：半小时、1小时、2小时、6小时、12小时、1天、1周、1月，也可以单击自定义“开始时间”和“结束时间”：

表 10-22 日志检索参数

参数名	说明
检索内容	检索的关键字或正则表达式。
服务	选择所需查询日志的服务或模块。
文件	当且仅当选择服务中一个角色时，支持选择指定日志文件进行搜索。
最低日志级别	选择所需查询日志的最低日志级别，选择某一级别后会显示从本级别到更高日志级别的日志。 级别从低到高依次为： TRACE < DEBUG < INFO < WARN < ERROR < FATAL
主机范围	• 单击…可勾选所需主机。 • 请输入所需查询日志的节点主机名或管理平面的IP地址。 • 各IP地址间用“,”隔开，例如：192.168.10.10,192.168.10.11。 • 如果IP地址连续，用“-”连接。例如：192.168.10.[10-20]。 • 如果IP地址分段连续，连续时用“-”连接，各IP地址段间用“,”隔开，例如：192.168.10.[10-20,30-40]。 **说明** – 如不指定，默认选择所有主机。 – 一次性输入最多10个表达式。 – 所有表达式一次性最多匹配2000个主机。
高级配置	• 最大数量：一次性显示的最大日志条数，如果检索到的日志数量超过设定值，时间较早的将被忽略。不配表示不限制。 • 检索超时：用于限制每个节点上的最大检索时间，超时后会中止搜索，已经搜索到的结果仍会显示。

步骤4 单击“检索”开始搜索，结果包含字段如表10-23所示。

表 10-23 检索结果

参数名	说明
时间	该行日志产生的具体时间点。
来源	产生日志的集群。
主机名称	记录该行日志的日志文件所在节点的主机名。
位置	该行日志所在的日志文件的具体路径。 单击位置信息可进入在线日志浏览页面。默认显示该日志所在行前后各100条日志，可单击页首或页尾的“更多”显示更多日志信息。单击“下载”可以下载该日志文件到本地。
行号	该行日志在日志文件中所在的行数。
级别	该行日志的级别。
日志	日志的具体内容。

说明

在检索过程中可单击“停止”强制停止当前检索进度，并在列表中显示已检索出的结果。

步骤5 单击“过滤”，可以针对界面上已经显示的日志信息进行二次筛选，具体字段如表10-24所示。填写完毕后，单击“过滤”进行检索，单击“重置”可清空已填写信息。

表 10-24 过滤

参数名	说明
关键字	需要检索的日志关键字。
主机名称	需要检索的主机名。
位置	所需检索的日志文件路径。
开始时间	所需检索日志信息的开始时间。
结束时间	所需检索日志信息的结束时间。
来源集群	需要检索的集群。

----结束

10.5.2.2 下载日志

操作场景

FusionInsight Manager支持批量导出各个服务角色所有实例生成的日志，无需手工登录单个节点获取。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 日志 > 下载”。

步骤3 选择日志下载范围。

1. “服务”：单击 ··· 勾选所需服务。
2. “主机”：填写服务所部署主机的IP，也可单击 ··· 勾选所需主机。
3. 单击右上角的 设置日志的起始收集时间“开始时间”和“结束时间”。

步骤4 单击“下载”完成日志下载。

下载的日志压缩包中会包括对应开始时间和结束时间的拓扑信息，方便查看与定位。

拓扑文件以“topo_<拓扑结构变化时间点>.txt”命名。文件内容包括节点IP、节点主机名以及节点所安装的服务实例（OMS节点以“Manager:Manager”标识）。

例如：

```
192.168.204.124|suse-124|
DBService:DBServer;KrbClient:KerberosClient;LdapClient:SlapdClient;LdapServer:SlapdServer;Manager:Manager;meta:meta
```

----结束

10.5.3 健康检查

10.5.3.1 查看健康检查任务

操作场景

管理员可以在健康检查的管理中心查看所有健康检查任务，便于在修改某些配置之后的场景对比修改前后是否对集群产生影响。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 健康检查”。

默认显示所有已保存的检查报告，以列表形式展示，包含如下所示的信息：

表 10-25 健康检查报告记录

项目	说明
检查对象	表示被检查的对象，可打开下拉菜单查看详情。
状态	表示检查的结果状态，包含未发现问题、发现问题和检查中。
检查类型	表示执行检查的主体，包含“系统”、“集群”、“主机”、“服务”和“OMS”五个检查维度。集群维度默认包含所有检查。

项目	说明
启动方式	表示此次检查的属性，是否自动触发或手动执行。
开始时间	表示此次检查的开始时间。
结束时间	表示此次检查的结束时间。
操作	支持“导出报告”和“查看帮助”。

说明

- 在检查记录列表右上方，可以筛选指定的检查对象和结果状态。
- 如果检查类型为集群时，“查看帮助”在“检查对象”的下拉菜单中。
- 系统执行健康检查时，涉及检查对象的监控指标数据时，并非以当前实时的监控数据进行判断，而是收集近期的历史数据，因此存在时间延迟。

----结束

10.5.3.2 管理健康检查报告

操作场景

用户可以在FusionInsight Manager对已保存的所有健康检查报告进行管理，即下载和删除历史健康检查报告。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 健康检查”。

步骤3 在目标健康检查报告所在行，单击“导出报告”，下载报告文件。

----结束

10.5.3.3 修改健康检查配置

操作场景

管理员可以启用自动健康检查减少手工操作时间。自动健康检查默认会对整个集群进行检查。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 健康检查 > 配置”。

“定期健康检查”表示是否启用自动执行健康检查，选择“启用”表示启用，默认“不启用”表示不启用。

启用后根据运维需要选择检查周期为：“每天”、“每周”或“每月”。

步骤3 单击“确定”保存配置。

----结束

10.5.4 备份恢复设置

10.5.4.1 创建备份任务

操作场景

FusionInsight Manager支持在界面上创建备份任务，运行备份任务将对指定的数据进行备份。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 备份恢复 > 备份管理 > 创建”。

步骤3 设置“备份对象”为“OMS”或需要备份数据的集群。

步骤4 在“任务名称”输入参数值。

步骤5 在“备份类型”选择任务执行属性。

表 10-26 备份类型说明

类型	参数	说明
周期备份	开始时间	表示周期备份任务第一次启动的时间
	周期	表示任务下次启动，与上一次运行的时间间隔，支持“小时”或“天”
	备份策略	可以选择下策略： ● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份
手动备份	无	需要手动运行任务才能进行备份

步骤6 在“备份配置”指定需要备份的数据。

- 支持备份元数据和业务数据。
- 各组件不同数据的备份任务操作请参考**备份恢复管理**。

步骤7 单击“确定”保存。

步骤8 在备份任务列表，可以查看刚创建的备份任务。

在指定的备份任务“操作”列，选择“更多 > 即时备份”，可以立即运行备份任务。

----结束

10.5.4.2 创建恢复任务

操作场景

FusionInsight Manager支持在界面上创建恢复任务，运行恢复任务将把指定的备份数据恢复到集群中。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 备份恢复 > 恢复管理 > 创建”。

步骤3 设置“恢复对象”为“OMS”或需要恢复数据的集群。

步骤4 在“任务名称”输入参数值。

步骤5 在“恢复配置”指定需要恢复的数据。

- 支持恢复元数据和业务数据。
- 各组件不同数据的恢复任务操作请参考**备份恢复管理**。

步骤6 单击“确定”保存。

步骤7 在恢复任务列表，可以查看刚创建的恢复任务。

在指定的备份任务“操作”列，单击“执行”，可以立即运行恢复任务。

----结束

10.5.4.3 其他任务管理说明

操作场景

FusionInsight Manager还支持对备份恢复进行不同的维护管理功能。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 备份恢复 > 备份管理”或“运维 > 备份恢复 > 恢复管理”。

步骤3 在任务列表指定任务的“操作”列，选择需要执行的操作。

表 10-27 更多维护管理功能

操作入口	说明
“配置”	修改备份任务的参数。
“恢复”	部分业务数据的备份任务执行成功后，可以直接使用此功能快速恢复数据。
“更多 > 即时备份”	立即运行备份任务。
“更多 > 停止”	可以停止正在运行的任务。

操作入口	说明
“更多 > 删除”或“删除”	删除任务。
“更多 > 挂起”	禁用自动备份任务。
“更多 > 重新执行”	启用自动备份任务。
“更多 > 查询历史”或“查询历史”	打开任务运行日志窗口，查看运行详细情况以及备份路径。
“查看”	检查恢复任务的参数设置。
“执行”	运行恢复任务。

----结束

10.6 审计

10.6.1 审计管理页面概述

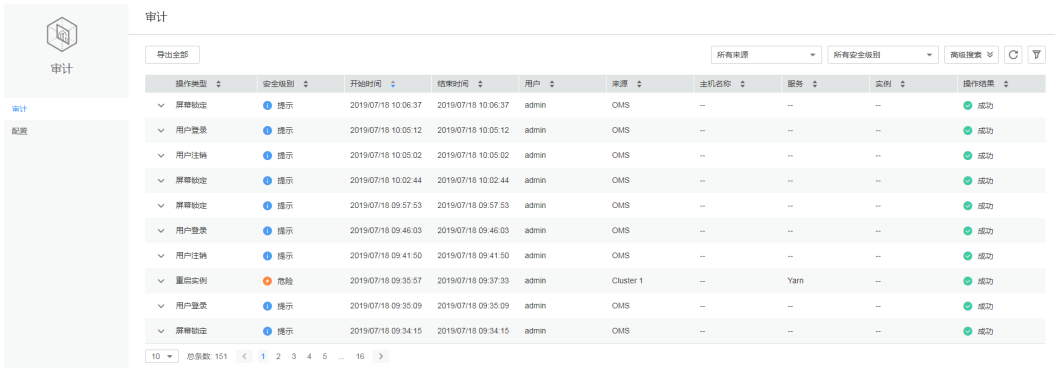
操作场景

“审计”页面记录用户对集群Manager页面操作信息。管理员可通过该页面查看用户在Manager上的历史操作记录。审计管理包含的审计内容信息，请参考**审计日志**。

概述

登录FusionInsight Manager，单击“审计”，界面展示如**图10-13**所示FusionInsight Manager审计信息，包括操作类型、安全级别、开始时间、结束时间、用户、主机名称、服务、实例、操作结果等。

图 10-13 审计信息列表



- 用户可以在“所有安全级别”中选择高危、危险、一般和提示级别的审计日志。
- 在高级搜索中，用户可设置过滤条件来查询审计日志。
  a. 在“操作类型”中，用户可根据用户管理、集群、服务、健康检查等来指定操作类型查询对应的审计日志。

b. 在“服务”中，用户可选择相应的服务来查询审计日志。

说明

在服务中选择“--”，表示除服务以外其他类型的审计日志。

c. 在“操作结果”中，用户可选择成功、失败和未知来查询审计日志。

- 单击手动刷新当前页面，也可在 修改审计表格显示的列。
- 单击“导出全部”，可一次性导出所有审计信息，可导出“TXT”或者“CSV”格式。

10.6.2 配置审计日志转储

操作场景

Manager的审计日志默认保存在数据库中，如果长期保留可能引起数据目录的磁盘空间不足问题，管理员如果需要将审计日志保存到其他归档服务器，可以在FusionInsight Manager设置转储参数及时自动转储，便于管理审计日志信息。

若用户未配置审计日志转储，当审计日志达到十万条，系统自动将这十万条审计日志保存到文件中。保存路径为主管理节点“${BIGDATA_DATA_HOME}/dbdata_om/dumpData/iam/operatelog”，保存的文件名格式为“OperateLog_store_YY_MM_DD_HH_MM_SS.csv”，保存的审计日志历史文件数最大为50。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“审计 > 配置”。

步骤3 单击“审计日志转储”右侧的开关。

“审计日志转储”默认为不启用，开关显示为表示启用。

步骤4 根据表10-28填写转储参数。

表 10-28 审计日志转储参数

参数名	参数解释	参数值
SFTP IP 模式	目标IP的IP地址模式，可选择“IPv4”或者“IPv6”。	IPv4
SFTP IP	必选参数，指定审计日志转储后存放的SFTP服务器，建议使用基于SSH v2的SFTP服务，否则存在安全风险。	192.168.10.51（举例）
SFTP端口	必选参数，指定审计日志转储后存放的SFTP服务器连接端口。	22（举例）
保存路径	必选参数，指定SFTP服务器上保存审计日志的路径。	/opt/omm/oms/auditLog（举例）

参数名	参数解释	参数值
SFTP用户名	必选参数，指定登录SFTP服务器的用户名。	root（举例）
SFTP密码	必选参数，指定登录SFTP服务器的密码。	*SFTP服务器的密码*
SFTP公共密钥	可选参数，指定SFTP服务器的公共密钥，建议配置SFTP的公共密钥，否则可能存在安全风险。	-
转储模式	必选参数，指定转储模式 ● “按数量”：日志到达指定条数（默认10万条）时开始转储 ● “按时间”：指定某一日期开始转储，转储频率为一年一次。	● 按数量 ● 按时间
转储日期	必选参数，当选择“按时间”转储模式时可用。选择一个转储日期后，系统将在此日期开始转储。转储的日志范围为当前年份1月1日0时之前的所有审计日志。	11月06（举例）

说明

SFTP公共密钥为空时，系统将进行安全风险提示，确定安全风险后再保存配置。

步骤5 单击“确定”，设置完成。

说明

审计日志转储文件关键字段参考：

- “USERTYPE”表示用户类型，“0”表示“人机”用户，“1”表示“机机”用户。
- “LOGLEVEL”表示安全级别，“0”表示高危，“1”表示危险，“2”表示一般，“3”表示提示。
- “OPERATERESULT”表示操作结果，“0”表示成功，“1”表示失败。

----结束

10.7 租户资源

10.7.1 多租户介绍

10.7.1.1 简介

定义

多租户是MRS集群中的多个资源集合（每个资源集合是一个租户），具有分配和调度资源的能力。资源包括计算资源和存储资源。

背景

现代企业的数据集群在向集中化和云化方向发展，企业级大数据集群需要满足：

- 不同用户在集群上运行不同类型的应用和作业（分析、查询、流处理等），同时存放不同类型和格式的数据。
- 某些类型的用户（例如银行、政府单位等）对数据安全非常关注，很难容忍将自己的数据与其他用户的放在一起。

这给大数据集群带来了以下挑战：

- 合理地分配和调度资源，以支持多种应用和作业在集群上平稳运行。
- 对不同的用户进行严格的访问控制，以保证数据和业务的安全。

多租户将大数据集群的资源隔离成一个个资源集合，彼此互不干扰，用户通过“租用”需要的资源集合，来运行应用和作业，并存放数据。在大数据集群上可以存在多个资源集合来支持多个用户的不同需求。

对此，MRS企业级大数据集群提供了完整的企业级大数据多租户解决方案。

优势

- 合理配置和隔离资源

  租户之间的资源是隔离的，一个租户对资源的使用不影响其它租户，保证了每个租户根据业务需求去配置相关的资源，可提高资源利用效率。
- 测量和统计资源消费

  系统资源以租户为单位进行计划和分配，租户是系统资源的申请者和消费者，其资源消费能够被测量和统计。
- 保证数据安全和访问安全

  多租户场景下，分开存放不同租户的数据，以保证数据安全；控制用户对租户资源的访问权限，以保证访问安全。

10.7.1.2 技术原理

10.7.1.2.1 多租户管理页面概述

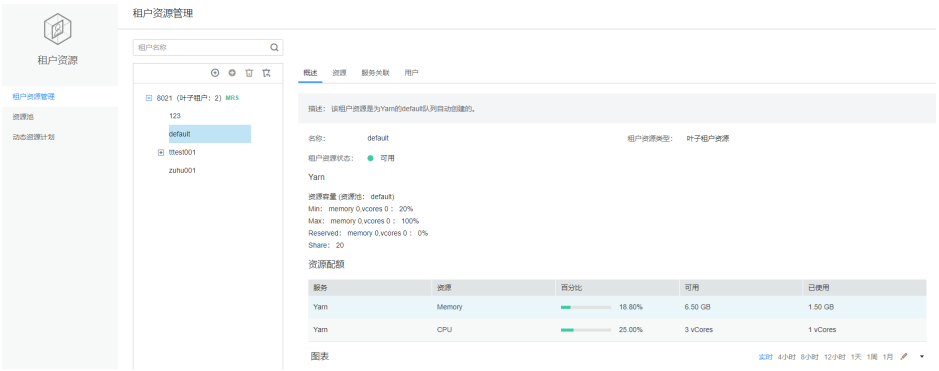
统一的多租户管理

登录FusionInsight Manager，选择“租户资源 > 租户资源管理”，可以查看到FusionInsight Manager作为统一的多租户管理平台，集成了租户生命周期管理、租户资源配置、租户服务关联和租户资源使用统计等功能，为企业提供了成熟的多租户管理模型，实现集中式的租户和业务管理。

图形化的操作界面

FusionInsight Manager实现全图形化的多租户管理界面：通过树形结构实现多级租户的管理和操作，将当前租户的基本信息和资源配额集成在一个界面中，方便运维和管理，如图10-14所示。

图 10-14 多租户管理



层级式的租户管理

FusionInsight Manager支持层级式的租户管理，可以为租户进一步添加子租户，实现资源的再次配置。一级租户下一级的子租户属于二级租户，以此类推。为企业提供了成熟的多租户管理模型，实现集中式的租户和业务管理。

简化的权限管理

FusionInsight Manager对普通用户封闭了租户内部的权限管理细节，对管理员简化了权限管理的操作方法，提升了租户权限管理的易用性和用户体验。

- 使用RBAC方式，在多租户管理时，可根据业务场景为各用户分别配置不同权限。
- 租户的管理员，具有租户的管理权限，包括：查看当前租户的资源和服务、在当前租户中添加/删除子租户并管理子租户资源的权限。支持定义单个租户的管理员，可以将租户的管理权限委托给系统管理员之外的其它用户。
- 租户对应的角色，具有租户的计算资源和存储资源的全部权限。创建租户时，系统自动创建租户对应的角色，可以添加用户并绑定该角色为其他用户授权，以使用该租户的资源。

清晰的资源管理

- **资源自主配置**

  FusionInsight Manager支持在创建租户时配置计算资源和存储资源，和进一步添加、修改、删除租户内资源。

  修改租户的计算资源或存储资源，当前租户对应的角色所关联的权限将自动更新。

- **资源使用统计**

  资源使用统计是管理员获取当前集群应用和服务的运行状态，提高集群运维效率，做出运维决策的重要依据。FusionInsight Manager通过“资源配额”展示租户的资源统计，包括租户动态计算资源VCores和Memory， HDFS存储资源（Space）的使用统计。

说明

- “资源配额”视图动态计算租户资源使用情况。

服务	资源	百分比	可用	已使用
Yarn	Memory	0.00%	24.00 GB	0 MB
Yarn	CPU	0.00%	12 vCores	0 vCores

  Superior调度器可用资源计算方式分别如下：

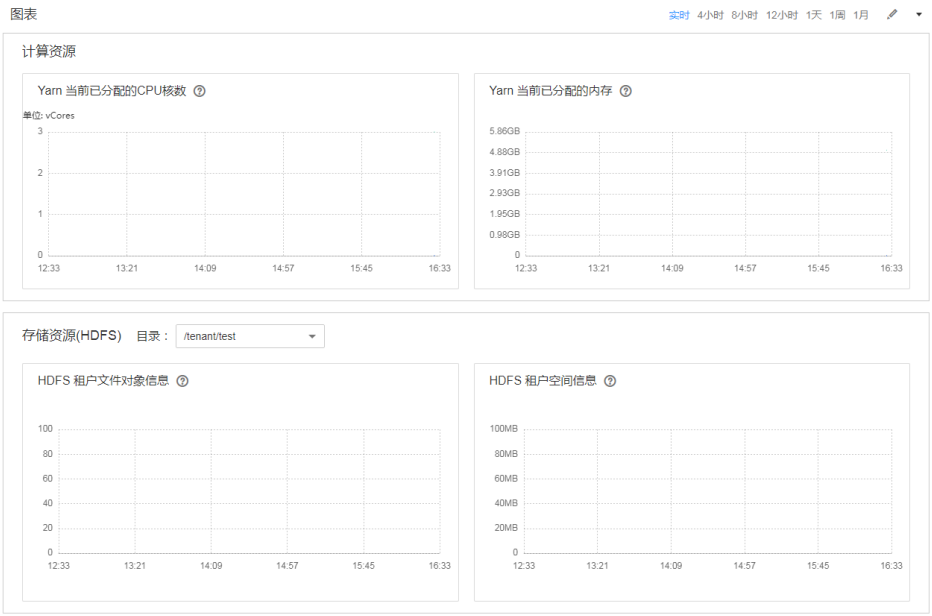
  - Superior

    Yarn可用资源（Memory、CPU）为资源池容量按照队列权重按比例分配。
- 当租户管理员绑定一个租户角色时，租户管理员将拥有该租户的管理权限，以及该租户全部资源的权限。

- **资源图形化监控**

  资源图形化监控支持表10-29中监控项图形化显示，如图10-15所示。

图 10-15 精细化监控



默认显示实时的监控数据，用户可单击 自定义时间区间，缺省时间区间包括：4小时、8小时、12小时、1天、1周、1月，单击 ，在弹出菜单中选择“导出”，导出对应的监控项信息。

表 10-29 监控项

所属服务	监控指标项	说明
HDFS	HDFS租户空间信息 • 分配的空间大小 • 已使用的空间大小	HDFS可选择指定的存储目录进行监控。存储目录与当前租户在“资源”中添加的目录一致。
	HDFS租户文件对象信息 • 已使用的文件对象个数	

所属服务	监控指标项	说明
Yarn	Yarn当前已分配的CPU核数 • AM分配的最大CPU核数 • 已分配的CPU核数 • AM已使用的CPU核数	当前租户的监控信息。如某租户未配置相应子项，则不显示。 监控数据取自Yarn原生WebUI中“Scheduler > Application Queues > Queue:*租户名*”。
	Yarn当前已分配的内存 • AM分配的最大内存 • 已分配的内存 • AM已使用的内存	

10.7.1.2.2 相关模型

多租户相关模型

多租户相关模型如下图所示。

图 10-16 多租户相关模型

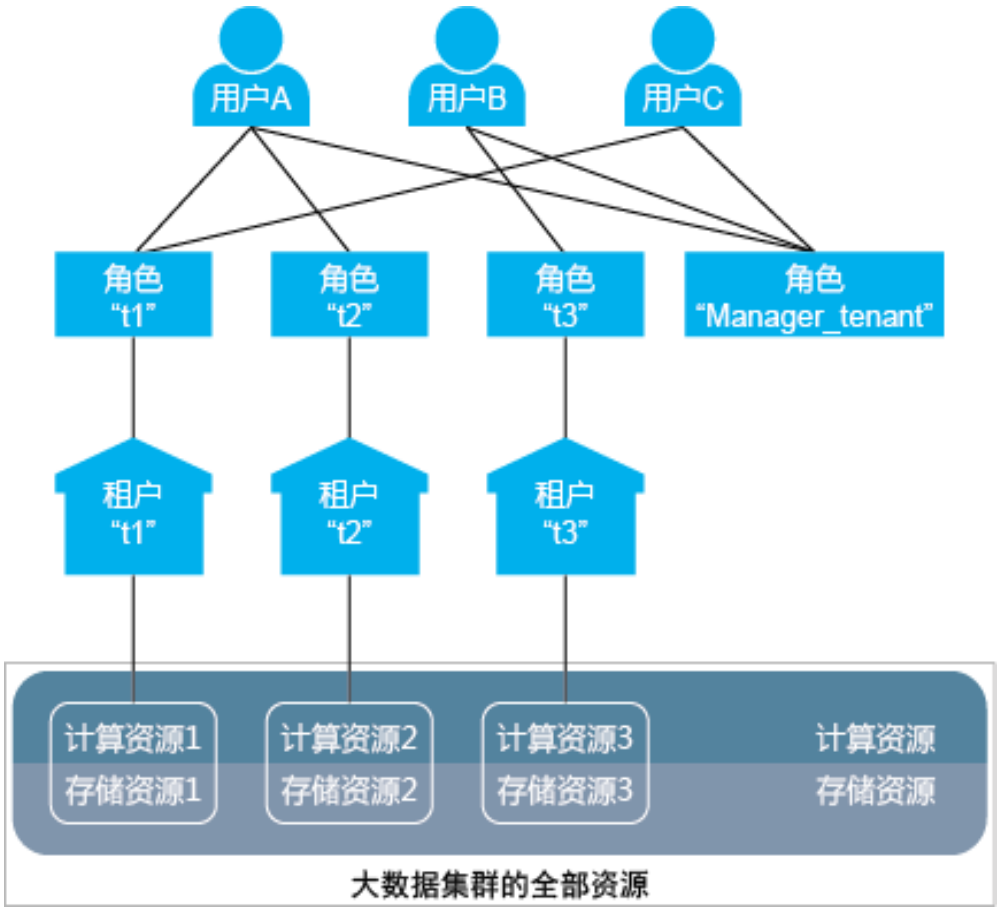


图10-16中涉及的概念如表10-30所示。

表 10-30 相关概念说明

概念	说明
用户	用户是拥有用户名，密码等帐号信息的自然人，是大数据集群的使用者。 图10-16中有三个不同的用户：用户A、用户B和用户C。
角色	角色是承载一个或多个权限的载体。权限是限定在具体对象上的，例如对HDFS中的“/tenant”目录的访问权限，这里权限就限定在“/tenant”目录这个具体对象上。 图10-16中有四个不同的角色：角色“t1”、角色“t2”、角色“t3”和角色“Manager_tenant”。 • 角色“t1”、角色“t2”和角色“t3”为创建租户时，集群自动生成的角色，角色名和租户名相同，分别对应租户“t1”、租户“t2”和租户“t3”，不能单独使用。 • 角色“Manager_tenant”为集群中本身存在的角色，不能单独使用。
租户	租户是从大数据集群中划分出的资源集合。多个不同的租户统称为多租户，租户内部进一步划分出的资源集合是子租户。 图10-16中有三个不同的租户：租户“t1”、租户“t2”和租户“t3”。
资源	• 计算资源包括CPU和内存。 租户的计算资源是从集群总计算资源中划分出的，租户之间不可以互占计算资源。 图10-16中：计算资源1、计算资源2和计算资源3分别是租户“t1”、租户“t2”和租户“t3”从集群中划分出的计算资源。 • 存储资源包括磁盘或第三方存储系统。 租户的存储资源是从集群总存储资源中划分出的，租户之间不可以互占存储资源。 图10-16中：存储资源1、存储资源2和存储资源3分别是租户“t1”、租户“t2”和租户“t3”从集群中划分出的存储资源。

若用户想要使用租户资源或为租户添加/删除子租户，则需要同时绑定该租户对应的角色和角色“Manager_tenant”。在图10-16中，各用户绑定的角色如表10-31所示。

表 10-31 各用户绑定的角色

用户	绑定的角色	权限
用户A	• 角色“t1” • 角色“t2” • 角色“Manager_tenant”	• 使用租户“t1”和租户“t2”的资源。 • 为租户“t1”和租户“t2”添加/删除子租户。

用户	绑定的角色	权限
用户B	• 角色“t3” • 角色“Manager_tenant”	• 使用租户“t3”的资源。 • 为租户“t3”添加/删除子租户。
用户C	• 角色“t1” • 角色“Manager_tenant”	• 使用租户“t1”的资源。 • 为租户“t1”添加/删除子租户。

用户和角色是多对多的关系，一个用户可以绑定多个角色，一个角色可以被多个用户绑定。用户通过绑定角色和租户建立关系，因此用户和租户也是多对多的关系。一个用户可以使用多个租户的资源，多个用户也可以使用同一个租户的资源，例如图10-16中，用户A使用租户“t1”和租户“t2”的资源，用户A和用户C都使用租户“t1”的资源。

说明

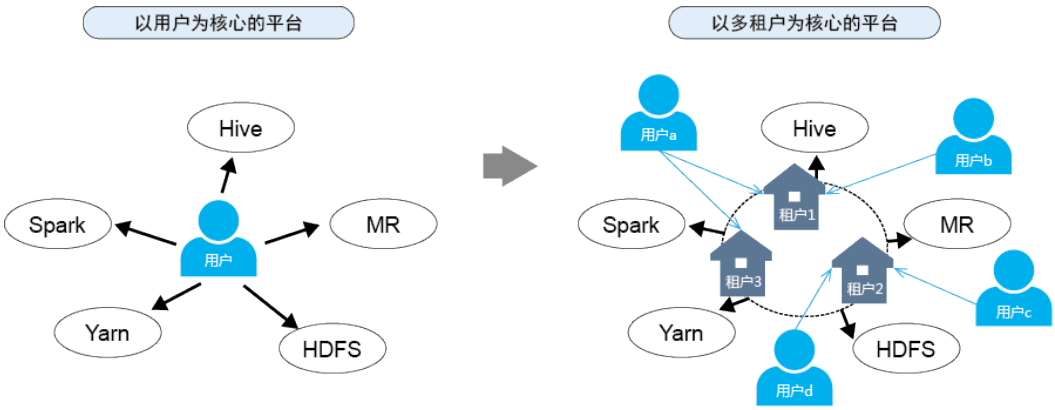
“父租户”、“子租户”、“一级租户”、“二级租户”的概念是针对客户的多租户业务场景设定的，注意与FusionInsight Manager上的“叶子租户资源”和“非叶子租户资源”的概念区别。

- 一级租户：按照租户所在层级确定名称，如最初创建的租户为一级租户，一级租户的子租户为二级租户。
- 父租户与子租户：用来表述租户间上下层级关系的称呼。
- 非叶子租户资源：创建租户时，选择的租户类型，该类型可以创建子租户。
- 叶子租户资源：创建租户时，选择的租户类型，该类型不可以创建子租户。

多租户平台

租户是FusionInsight大数据平台的核心概念，使传统的以用户为核心的大数据平台向以多租户为核心的大数据平台转变，更好的适应现代企业多租户应用环境，如图10-17所示。

图 10-17 以用户为核心的平台和以多租户为核心的平台



对于以用户为核心的大数据平台，用户直接访问并使用全部的资源和服务。

- 用户的应用可能只用到集群的部分资源，资源利用效率低。
- 不同用户的数据可能存放在一起，难以保证数据安全。

对于以租户为核心的大数据平台，用户通过访问租户来使用需要的资源和服务。

- 按照应用需求分配和调度出需要的资源，以租户来统一使用，资源利用效率高。
- 用户通过分配不同的角色获得使用不同租户资源的权限，以保障访问安全。
- 不同的租户之间数据隔离，以保证数据安全。

10.7.1.2.3 资源概述

MRS集群的资源分为计算资源和存储资源。多租户可实现资源的隔离：

- **计算资源**

  计算资源包括CPU和内存。租户之间不可以相互占用计算资源，私有计算资源独立。
- **存储资源**

  存储资源包括磁盘或第三方存储系统。租户之间不可以相互访问数据，私有存储资源独立。

计算资源

计算资源可分为静态服务资源和动态资源：

- **静态服务资源**

  静态服务资源是集群分配给各个服务的计算资源，每个服务的计算资源总量固定，不与其他服务共享，是静态的。这些服务包括Flume、HBase、HDFS和Yarn。
- **动态资源**

  动态资源是分布式资源管理服务Yarn动态调度给任务队列的计算资源。Mapreduce、Spark2x、Flink和Hive的任务队列由Yarn来动态调度资源。

说明

大数据集群为Yarn分配的资源是静态服务资源，可以由Yarn动态分配给任务队列计算使用。

存储资源

存储资源是分布式文件存储服务HDFS中可分配的数据存储空间资源。目录是HDFS存储资源分配的基本单位，租户通过指定HDFS文件系统的目录来获取存储资源。

10.7.1.2.4 动态资源

简介

Yarn是大数据集群中的分布式资源管理服务，大数据集群为Yarn分配资源，资源总量可配置。Yarn内部为任务队列进一步分配和调度计算资源。对于Mapreduce、Spark、Flink和Hive的任务队列，计算资源完全由Yarn来分配和调度。

Yarn任务队列是计算资源分配的基本单位。

对于租户，通过Yarn任务队列申请到的资源是动态资源。用户可以动态创建并修改任务队列的配额，可以查看任务队列的使用状态和使用统计。

资源池

现代企业IT经常会面对纷繁复杂的集群环境和上层需求。例如以下业务场景：

- 集群异构，集群中各个节点的计算速度、存储容量和网络性能存在差异，需要把复杂应用的所有任务按照需求，合理地分配到各个计算节点上。
- 计算分离，多个部门需要数据共享，但是需要把计算完全分离在不同的计算节点上。

这就要求对计算资源的节点进一步分区。

资源池用来指定动态资源的配置。Yarn任务队列和资源池关联，可实现资源的分配和调度。

一个租户只能设置一个默认资源池。用户通过绑定租户相关的角色，来使用该租户资源池的资源。若需要使用多个资源池的资源，可通过绑定多个租户相关的角色实现。

调度机制

Yarn动态资源支持标签调度（Label Based Scheduling）策略，此策略通过为计算节点（Yarn NodeManager）创建标签（Label），将具有相同标签的计算节点添加到同一个资源池中，Yarn根据任务队列对资源的需求，将任务队列和有相应标签的资源池动态关联。

例如，集群中有40个以上的节点，根据各节点的硬件和网络配置，分别用Normal、HighCPU、HighMEM、HighIO为四类节点创建标签，添加到四个资源池中，资源池中的各节点性能如**表10-32**所示。

表 10-32 不同资源池中的各节点性能

标签名	节点数	硬件和网络配置	添加到	关联
Normal	10	一般	资源池A	普通的任务队列
HighCPU	10	高性能CPU	资源池B	计算密集型的任务队列
HighMEM	10	大量内存	资源池C	内存密集型的任务队列
HighIO	10	高性能网络	资源池D	IO密集型的任务队列

任务队列只能使用所关联的资源池里的计算节点。

- 普通的任务队列关联资源池A，使用硬件和网络配置一般的Normal节点。
- 计算密集型的任务队列关联资源池B，使用具有高性能CPU的HighCPU节点。
- 内存密集型的任务队列关联资源池C，使用具有大量内存的HighMEM节点。
- IO密集型的任务队列关联资源池D，使用具有高性能网络的HighIO节点。

Yarn任务队列与特定的资源池关联，可以更有效地使用资源，保证节点性能充足且互不影响。

FusionInsight Manager中最多支持添加50个资源池。系统默认包含一个默认资源池。

调度器介绍

MRS集群默认即启用了Superior调度器。

- Superior调度器为增强型，Superior取名源自苏必利尔湖，意指由该调度器管理的数据足够大。

为满足企业需求，克服Yarn社区在调度上遇到的挑战与困难，Superior调度器做了以下增强：

- 增强资源共享策略

  Superior调度器支持队列层级，在同集群集成开源调度器的特性，并基于可配置策略进一步共享资源。针对实例，管理员可通过Superior调度器为队列同时配置绝对值或百分比的资源策略计划。Superior调度器的资源共享策略将Yarn的标签调度增强为资源池特性，Yarn集群中的节点可根据容量或业务类型不同，进行分组以使队列更有效地利用资源。

- 基于租户的资源预留策略

  部分租户可能在某些时间中运行关键任务，租户所需的资源应保证可用。Superior调度器构建了支持资源预留策略的机制，在这些租户队列运行的任务可立即获取到预留资源，以保证计划的关键任务可正常执行。

- 租户和资源池的用户公平共享

  Superior调度器提供了队列内用户间共享资源的配置能力。每个租户中可能存在不同权重的用户，高权重用户可能需要更多共享资源。

- 大集群环境下的调度性能优势

  Superior调度器接收到各个NodeManager上报的心跳信息，并将资源信息保存在内存中，使得调度器能够全局掌控集群的资源使用情况。Superior调度器采用了push调度模型，令调度更加精确、高效，大大提高了大集群下的资源使用率。另外，Superior调度器在NodeManager心跳间隔较大的情况下，调度性能依然优异，不牺牲调度性能，也能避免大集群环境下的“心跳风暴”。

- 优先策略

  当某个服务在获取所有可用资源后还无法满足最小资源的要求，则会发生优先抢占。抢占功能默认关闭。

10.7.1.2.5 存储资源

简介

HDFS是大数据集群中的分布式文件存储服务，存放大数据集群上层应用的所有用户数据，例如写入HBase表或Hive表的数据。

目录是HDFS存储资源分配的基本单位。HDFS支持传统的层次型文件组织结构。用户或者应用程序可以创建目录，在目录中创建、删除、移动或重命名文件。租户通过指定HDFS文件系统的目录来获取存储资源。

调度机制

系统支持将HDFS目录存储到指定标签的节点上，或存储到指定硬件类型的磁盘上。例如以下业务场景：

- 实时查询与数据分析共集群时，实时查询只需部署在部分节点上，其数据也应尽可能的只存储在这些节点上。

- 关键数据根据实际业务需要保存在具有高度可靠性的节点中。

管理员可以根据实际业务需要，通过数据特征灵活配置HDFS数据存储策略，将数据保存在指定的节点上。

对于租户，存储资源是各租户所占用的HDFS资源。可以通过将指定目录的数据存储到租户配置的存储路径中，实现存储资源调度，保证租户间的数据隔离。

用户可以添加/删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额来管理存储资源。

10.7.1.3 多租户使用

10.7.1.3.1 使用说明

租户主要用于资源控制、业务隔离的场景。在实际业务中，管理员需要先明确使用集群资源的业务场景，规划租户。

说明

- 新安装集群的Yarn组件默认使用的是Superior调度器，参见**使用Superior调度器的租户业务**。

多租户使用包含三类操作：创建租户、管理租户和管理资源。各操作的具体动作如**表10-33**所示。

表 10-33 使用租户的各种操作

操作	具体动作	说明
创建租户	• 添加租户 • 添加子租户 • 添加用户并绑定租户的角色	创建租户时，便可根据业务需求，为租户配置计算资源、存储资源和关联服务；为租户添加用户，并为用户绑定需要的角色。 创建一级租户的用户，需要绑定“Manager_administrator”或“System_administrator”角色。 创建子租户的用户，至少需要绑定父租户对应的角色。
管理租户	• 管理租户目录 • 恢复租户数据 • 清除租户非关联队列 • 删除租户	管理租户是随着业务变化对租户进行的编辑操作。 管理或删除一级租户的用户，以及恢复租户数据的用户，需要绑定“Manager_administrator”或“System_administrator”角色。 管理或删除子租户的用户，至少需要绑定父租户对应的角色。

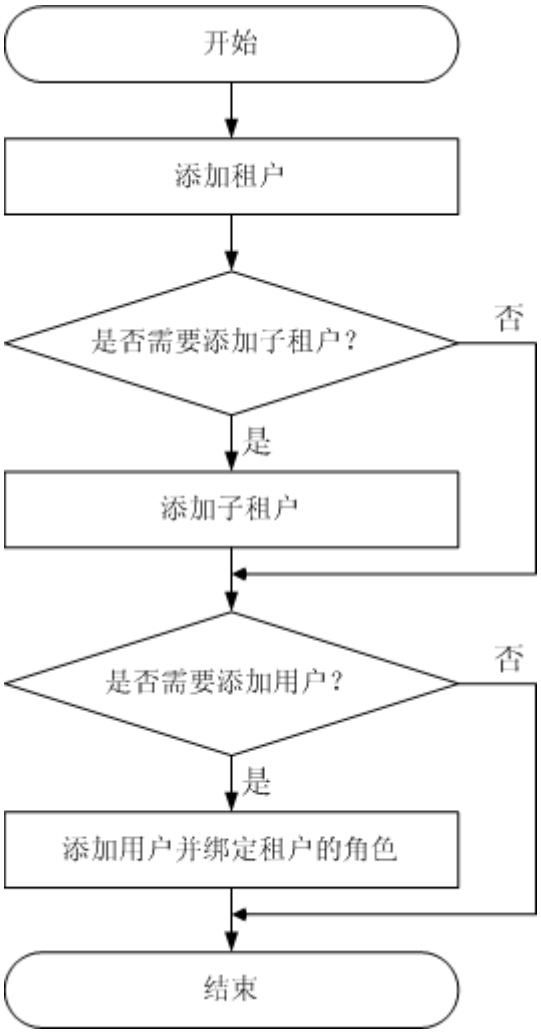
操作	具体动作	说明
管理资源	• 添加资源池 • 修改资源池 • 删除资源池 • 配置队列 • 配置资源池的队列容量策略 • 清除队列配置	管理资源是随着业务变化对租户再次配置资源的操作。 管理资源的用户，需要绑定“Manager_administrator”或“System_administrator”角色。

10.7.1.3.2 流程概述

在实际业务中，管理员需要先明确使用集群资源的业务场景，规划租户。然后在FusionInsight Manager界面添加租户，并配置租户的动态资源、存储资源以及所关联的服务。

创建租户的具体操作流程如**流程概述**所示。

图 10-18 创建租户流程



创建租户的操作说明如表10-34所示。

表 10-34 创建租户的操作说明

操作	说明
添加租户	可配置待添加租户的计算资源、存储资源和关联服务。
添加子租户	可配置待添加子租户的计算资源、存储资源和关联服务。
添加用户并绑定租户的角色	若一个用户想要使用“tenant1”租户包含的资源，或为“tenant1”租户添加/删除子租户，则需要同时绑定“Manager_tenant”和“tenant1_*集群ID*”两个角色。

10.7.2 使用 Superior 调度器的租户业务

10.7.2.1 创建租户

10.7.2.1.1 添加租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建租户，以满足实际使用场景。

前提条件

- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 单击⊕，打开添加租户的配置页面，参见表10-35为租户配置属性。

表 10-35 租户参数一览

参数名	描述
集群	选择要创建租户的集群。

参数名	描述
名称	● 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（ _ ）。 ● 根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
租户资源类型	指定租户是否是一个叶子租户： ● 选择“叶子租户资源”：当前租户为叶子租户，不支持添加子租户。 ● 选择“非叶子租户资源”：当前租户为非叶子租户，支持添加子租户。
计算资源	为当前租户选择动态计算资源。 ● 选择“Yarn”时，系统自动在Yarn中以租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“Default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 ● 不选择“Yarn”时，系统不会自动创建任务队列。
配置模式	计算资源参数配置模式。 ● 选择“基础”时，只需配置“默认资源池容量（%）”参数即可。 ● 选择“高级”时，可手动配置资源分配权重，租户的最小/最大/预留资源。
默认资源池容量 （%）	配置当前租户在默认资源池中使用的计算资源百分比，取值范围0~100%。
权重	资源分配权重，取值范围从0到100。
最小资源	保证租户资源能获得的资源（有抢占支持）。取值可以是父租户资源的百分比或绝对值。当租户资源作业量比较少时，资源会自动借给其他租户资源，当租户资源能使用的资源不满足最小资源时，可以通过抢占来要回之前借出的资源。
最大资源	租户资源最多能使用的资源，租户资源不能得到比最大资源设定更多的资源。取值可以是父租户资源的百分比或绝对值。
预留资源	租户资源预留资源。即使租户资源内没有作业，预留的资源也不能给别的租户资源使用。取值可以是父租户资源的百分比或绝对值。

参数名	描述
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将分配存储资源。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。
存储空间配额	配置当前租户使用的HDFS存储空间配额。 ● 取值范围：当存储空间配额单位设置为MB时，范围为1～8796093022208。当存储空间配额单位设置为GB时，范围为1～8589934592。 ● 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 ● 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。
存储路径	配置租户在HDFS中的存储目录。 ● 系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“/tenant/ta1”。 ● 第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
服务	是否需要关联使用其他服务的资源，参见**步骤4**。
描述	配置当前租户的描述信息。

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 创建租户时系统会自动创建一个Yarn任务队列，并自动以租户名称命名该队列。如果已经存在同名队列，新队列命名为“租户名称-N”。“N”表示从1开始的自然数，存在同名队列的时候N会自动累加以区别已有队列。例如“saletenant”、“saletenant-1”和“saletenant-2”。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：
   - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
   - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

10.7.2.1.2 添加子租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建子租户，将当前租户的资源进一步分配以满足实际使用场景。

前提条件

- 已添加父租户，且属于非叶子租户。
- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择父租户节点然后单击，打开添加子租户的配置页面，参见**表10-36**为子租户配置属性。

表 10-36 子租户参数一览

参数名	描述
集群	显示上级父租户所在集群。
父租户资源	显示上级父租户的名称。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划子租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。

参数名	描述
租户资源类型	指定租户是否是一个叶子租户： ● 选择“叶子租户资源”：当前租户为叶子租户，不支持添加子租户。 ● 选择“非叶子租户资源”：当前租户为非叶子租户，支持添加子租户，但租户层级不能超过5层。
计算资源	为当前租户选择动态计算资源。 ● 选择“Yarn”时，系统自动在Yarn中以子租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“Default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 ● 不选择“Yarn”时，系统不会自动创建任务队列。
配置模式	计算资源参数配置模式。 ● 选择“基础”时，只需配置“默认资源池容量（%）”参数即可。 ● 选择“高级”时，可手动配置资源分配权重，租户的最小/最大/预留资源。
默认资源池容量 （%）	配置当前租户使用的计算资源百分比，基数为父租户的资源总量。
权重	资源分配权重，取值范围从0到100。
最小资源	保证租户资源能获得的资源（有抢占支持）。取值可以是父租户资源的百分比或绝对值。当租户资源作业量比较少时，资源会自动借给其他租户资源，当租户资源能使用的资源不满足最小资源时，可以通过抢占来要回之前借出的资源。
最大资源	租户资源最多能使用的资源，租户资源不能得到比最大资源设定更多的资源。取值可以是父租户资源的百分比或绝对值。
预留资源	租户资源预留资源。即使租户资源内没有作业，预留的资源也不能给别的租户资源使用。取值可以是父租户资源的百分比或绝对值。
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。

参数名	描述
存储空间配额	配置当前租户使用的HDFS存储空间配额。 ● 当存储空间配额单位设置为MB时，范围为1～8796093022208，当“存储空间配额单位”设置为GB时，范围为1～8589934592。 ● 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 ● 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 ● 如果此配额大于父租户的配额，实际存储量不超过父租户配额。
存储路径	配置租户在HDFS中的存储目录。 ● 系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“/tenant/ta1”，系统默认自动配置此参数值为“/tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。 ● 支持在父目录中自定义存储路径。
服务	是否需要关联使用其他服务的资源，参见**步骤4**。
描述	配置当前租户的描述信息

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 子租户可以将当前租户的资源进一步分配。每一级别父租户下，直接子租户的资源百分比之和不能超过100%。所有一级租户的计算资源百分比之和也不能超过100%。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：
   - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
   - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

10.7.2.1.3 添加用户并绑定租户的角色

操作场景

创建好的租户不能直接登录集群访问资源，管理员需要通过FusionInsight Manager为已有租户创建新用户，通过绑定租户的角色继承其操作权限，以满足业务使用。

前提条件

管理员已明确业务需求，并已创建了租户。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 用户”。

步骤2 若在系统中添加新的用户，请单击“添加用户”，打开添加用户的配置页面。

若为系统中已有的用户绑定租户权限，请单击该用户所在行的“修改”，打开修改用户的配置页面。

参见**表10-37**为用户配置属性。

表 10-37 用户参数一览

参数名	描述
用户名	指定当前的用户名，长度为3~32个字符，可包含数字、字母、下划线（_）、中划线（-）和空格。 • “用户名”不能与集群各节点所有操作系统用户名相同，否则此用户无法正常使用。 • 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。使用“User1”时请输入正确的用户名。

参数名	描述
用户类型	可选值包括“人机”和“机机”。 ● “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。 ● “机机”用户：用于应用开发的场景。选择该值用户密码随机生成，无需填写。
密码	选择“人机”用户需填写“密码”。 密码必须包含8~64个字符，至少包含以下类型字符中的四种：大写字母、小写字母、数字、特殊字符和空格。不能与用户名或倒序的用户名相同。
确认密码	再次输入密码。
用户组	单击“添加”，选择对应用户组将用户添加进去。 ● 如果用户组添加了角色，则用户可获得对应角色中的权限。 ● 例如，为新用户分配Hive的权限，请将用户加入Hive组。
主组	选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。
角色	单击“添加”为用户绑定租户的角色。 **说明** ● 若一个用户想要获取使用“tenant1”租户包含的资源，且能够为“tenant1”租户添加/删除子租户，则需要同时绑定“Manager_tenant”和“tenant1_*集群ID*”两个角色。 ● 如果租户关联了HBase服务且当前集群启用了Ranger鉴权，用户需要通过Ranger界面配置HBase相关执行权限。
描述	配置当前用户的描述信息。

步骤3 单击“确定”完成用户创建。

----结束

10.7.2.2 管理租户

10.7.2.2.1 管理租户目录

操作场景

管理员通过FusionInsight Manager管理指定租户使用的HDFS存储目录，能根据业务需求对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

已添加具有HDFS存储资源的租户。

查看租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击目标的租户。

步骤3 单击“资源”页签。

步骤4 查看“HDFS存储”表格。

- 指定租户目录的“文件目录数上限”列表示文件和目录数量配额。
- 指定租户目录的“存储空间配额 ”列表示租户目录的存储空间大小。

----结束

添加租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，单击“添加目录”。

- “父目录”，表示当前租户对应父租户的存储目录。

  说明

  当前租户不是子租户则不显示此参数。

- “路径”，填写租户目录的路径。

  说明

  当前租户不是子租户则新路径将在HDFS的根目录下创建。

- “文件\目录数上限”填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  当前已使用的文件数的数据采集周期为1个小时，因此超过文件数阈值的告警上报会存在延迟。

- “存储空间配额”，填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  已使用的存储空间的数据采集周期为1个小时，因此超过存储空间阈值的告警上报会存在延迟。

步骤5 单击“确定”完成租户目录添加。

----结束

修改租户目录属性

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。

- “文件\目录数上限”，填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。
- “存储空间配额”填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

步骤5 单击“确定”完成租户目录修改。

----结束

删除租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“删除”。

说明

不支持删除创建租户时系统创建的租户目录。

步骤5 单击“确定”完成租户目录删除。

----结束

10.7.2.2.2 恢复租户数据

操作场景

租户默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，管理员需要通过FusionInsight Manager手动恢复配置数据。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“租户资源状态”，绿色表示租户可用，灰色表示租户不可用。

2. 单击“资源”，查看“Yarn”或者“HDFS存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击，在弹出的确认窗中输入当前登录的用户密码确认身份，单击“确定”。

步骤5 在“恢复租户资源数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

10.7.2.2.3 删除租户

操作场景

根据业务需求，对于当前不再使用的租户，管理员可以通过FusionInsight Manager删除租户，释放租户占用的资源。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择待删除的租户，单击。

说明

- 根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户资源的数据。”，否则将自动删除租户对应的存储空间。

步骤3 单击“确定”，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

租户删除后，Yarn中对应的租户任务队列不会被删除。同时Yarn角色管理中，此租户任务队列不再显示。

----结束

10.7.2.3 管理资源

10.7.2.3.1 添加资源池

操作场景

在集群中，管理员可从逻辑上对所有Yarn的节点进行分区，使多个NodeManager形成一个Yarn资源池。每个NodeManager只能属于一个资源池。管理员通过FusionInsight Manager添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池，便于指定的队列利用这些计算资源。

系统中默认包含了一个名为“default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 单击“添加资源池”。

步骤4 设置资源池的属性。

- “集群”：选择待添加资源池的集群名称。
- “名称”：填写资源池的名称。长度为1~50个字符，可包含数字、字母或下划线（_），且不能以下划线（_）开头。
- “资源标签”：配置资源池的资源标签，包括数字、字母、下划线（_）或减号（-），长度为1~50个字符，且只能以数字或者字母开头。
- “资源”：在界面左边可用主机列表中，勾选指定的主机，单击 >> ，将选中的主机加入已选主机列表。只支持选择本集群中的主机。资源池中的主机列表可以为空。

  说明

  根据业务需求，可以通过主机名称 、CPU、内存、操作系统和平台类型，筛选需要选取的资源主机。

步骤5 单击“确定”保存。

完成资源池创建后，管理员可以在资源池的列表中查看资源池的名称、成员、类型。已加入自定义资源池的主机，不再是“default”资源池的成员。

----结束

10.7.2.3.2 修改资源池

操作场景

根据业务需要，资源池的主机需要调整时，管理员可以通过FusionInsight Manager修改已有资源池中的成员。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“编辑”。

步骤4 在“资源”修改“主机”。

- 增加主机：在界面左边主机列表，选择指定的主机名称，单击 >> ，将选中的主机加入资源池。
- 删除主机：在界面右边主机列表，选择指定的主机名称，单击 << ，将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

10.7.2.3.3 删除资源池

操作场景

根据业务需要，资源池不再使用时，管理员可以通过FusionInsight Manager进行删除资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列容量配置**。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

步骤4 在弹出窗口中单击“确定”。

----结束

10.7.2.3.4 配置队列

操作场景

根据业务需求，管理员可以在FusionInsight Manager修改指定租户的队列配置。

前提条件

已添加使用Superior调度器的租户。

操作步骤

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 单击“动态资源计划”页签。

步骤3 单击“队列配置”页签。

步骤4 “集群”参数选择待操作的集群名称，然后指定租户资源名的“操作”列，单击“修改”。

说明

- 在“租户资源管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击“队列配置（*队列名*）”后面的也能打开修改队列配置页面。
- 一个队列只能绑定一个非default资源池。
- 对于“每个YARN容器最多分配核数”、“每个YARN容器最大分配内存（MB）”、“最多运行任务数”、“每个用户最多运行任务数”和“最多挂起任务数”等参数，为便于操作，当子租户值为-1时，父租户值可以设置为具体限制值；当父租户设置为具体限制值时，子租户可以设置为-1。
- “每个YARN容器最多分配核数”和“每个YARN容器最大分配内存（MB）”需要同时修改为非-1的值才会生效。

表 10-38 队列配置参数

参数名	描述
AM最多占有资源（%）	表示当前队列内所有Application Master所占的最大资源百分比。
每个YARN容器最多分配核数	表示当前队列内单个YARN容器可分配的最多核数，默认为-1，表示取值范围内不限制。
每个YARN容器最大分配内存（MB）	表示当前队列内单个YARN容器可分配的最大内存，默认为-1，表示取值范围内不限制。
最多运行任务数	表示当前队列最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
每个用户最多运行任务数	表示每个用户在当前队列中最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
最多挂起任务数	表示当前队列最多同时可挂起任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可挂起任务。取值范围为-1～2147483647。
资源分配规则	表示单个用户任务间的资源分配规则，包括FIFO和FAIR。 一个用户若在当前队列上提交了多个任务，FIFO规则代表一个任务完成后再执行其他任务，按顺序执行。FAIR规则代表各个任务同时获取到资源并平均分配资源。
默认资源标签	表示在指定资源标签（Label）的节点上执行任务。
Active状态	• ACTIVE表示当前队列可接受并执行任务。 • INACTIVE表示当前队列可接受但不执行任务，若提交任务，任务将处于挂起状态。

参数名	描述
Open状态	● OPEN表示当前队列处于打开状态。 ● CLOSED表示当前队列处于关闭状态，若提交任务，任务直接会被拒绝。
故障时是否队列迁移	集群开启单集群跨AZ高可用时，如果AZ故障后，需要当该租户正在运行的队列重新提交至其他AZ，可设置“故障时是否队列迁移”参数为“是”。

步骤5 单击“确定”完成配置。

----结束

10.7.2.3.5 配置资源池的队列容量策略

操作场景

添加资源池后，需要为Yarn任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。

该任务指导系统管理员通过FusionInsight Manager配置队列策略。使用Superior调度器的租户队列，可以配置使用不同资源池的资源。

前提条件

- 已登录FusionInsight Manager。
- 已添加资源池。
- 任务队列不与其他队列相关联资源池，除了默认资源池。

操作步骤

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 单击“动态资源计划”页签。

步骤3 单击“资源分布策略”页签。

步骤4 “集群”参数选择待操作的集群名称，然后在“资源池”选择指定的资源池。

步骤5 在“资源分配”列表指定队列的“操作”列，单击“修改”。

步骤6 在“修改资源分配”窗口的“资源配置策略”页签设置任务队列在此资源池中的资源配置策略。

- “权重”：表示租户能获得的资源。其初始值与最小资源百分比值一致。
- “最小资源”：表示租户能获得的最少资源。
- “最大资源”：表示租户能获得的最多资源。
- “预留资源”：表示保留给租户自身队列，且不能借用给其他租户队列的资源。

步骤7 在“修改资源分配”窗口的“用户策略”页签设置用户策略。

说明

defaultUser(built-in)表示如果一个用户未配置策略，则默认使用defaultUser所指定的策略。该策略不可删除。

- 单击“添加用户策略”添加用户策略。
  - “用户名”：表示用户的名称。
  - “权重”：表示用户能获得的资源。
  - “最多核数”：表示用户最多可以使用的虚拟核数。
  - “最大内存”：表示用户最大可以使用的内存。
- 单击“操作”列的“修改”修改现有用户策略。
- 单击“操作”列的“删除”删除现有用户策略。

步骤8 单击“确定”保存配置。

----结束

10.7.2.3.6 清除队列容量配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，管理员可以在FusionInsight Manager清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

操作步骤

步骤1 登录FusionInsight Manager界面。

步骤2 选择“租户资源 > 动态资源计划”。

步骤3 “集群”参数选择待操作的集群名称，然后在“资源池”，选择待操作的资源池。

步骤4 在资源分配表格，指定租户资源名的“操作”列，单击“清除”。

步骤5 在弹出的对话框中单击“确定”，清除队列在当前资源池的配置。

----结束

10.7.2.4 管理全局用户策略

操作场景

如果租户配置使用Superior调度器，那么系统可以控制具体用户使用资源调度器的行为，包含：

- 最大运行任务数
- 最大挂起任务数

- 默认队列

操作步骤

- 添加策略
  a. 在FusionInsight Manager，单击“租户资源”。
  b. 单击“动态资源计划”页签。
  c. 单击“全局用户策略”页签。

     说明

     defaults(default setting)表示如果一个用户未配置全局用户策略，则默认使用defaults所指定的策略。该策略不可删除。

  d. 单击“添加全局用户策略”，在弹出窗口中填写以下参数。
     - 集群：选择需要操作的集群。
     - 用户名：表示需要控制资源调度的用户，请输入当前集群中已存在用户的名称。
     - 最大运行任务数：表示该用户在当前集群中能运行的最大任务数量。
     - 最大挂起任务数：表示该用户在当前集群中能挂起的最大任务数量。
     - 默认队列：表示用户的队列，请输入当前集群中已存在队列的名称。
- 修改策略
  a. 在FusionInsight Manager，单击“租户资源”。
  b. 单击“动态资源计划”页签。
  c. 单击“全局用户策略”页签。
  d. 在指定用户策略所在行，单击“操作”列中的“修改”。
  e. 调整相关参数后，单击“确定”。
- 删除策略
  a. 在FusionInsight Manager，单击“租户资源”。
  b. 单击“动态资源计划”页签。
  c. 单击“全局用户策略”页签。
  d. 在指定用户策略所在行，单击“操作”列中的“删除”。
     在弹出窗口单击“确定”。

10.7.3 使用 Capacity 调度器的租户业务

10.7.3.1 创建租户

10.7.3.1.1 添加租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建租户，以满足实际使用场景。

前提条件

- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 单击⊕，打开添加租户的配置页面，参见**表10-39**为租户配置属性。

表 10-39 租户参数一览

参数名	描述
集群	选择要创建租户的集群。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
租户类型	指定租户是否是一个叶子租户： • 选择“叶子租户”：当前租户为叶子租户，不支持添加子租户。 • 选择“非叶子租户”：当前租户为非叶子租户，支持添加子租户。
计算资源	为当前租户选择动态计算资源。 • 选择“Yarn”时，系统自动在Yarn中以租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“Default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 • 不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量（%）	配置当前租户在“Default”资源池中使用的计算资源百分比，取值范围0~100%。
默认资源池最大容量（%）	配置当前租户在“Default”资源池中使用的最大计算资源百分比，取值范围0~100%。
存储资源	为当前租户选择存储资源。 • 选择“HDFS”时，系统将分配存储资源。 • 不选择“HDFS”时，系统不会分配存储资源。

参数名	描述
文件\目录数上限	配置文件和目录数量配额。
存储空间配额	配置当前租户使用的HDFS存储空间配额。 • 取值范围：当存储空间配额单位设置为MB时，范围为1～8796093022208。当存储空间配额单位设置为GB时，范围为1～8589934592。 • 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 • 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。
存储路径	配置租户在HDFS中的存储目录。 • 系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“/tenant/ta1”。 • 第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
描述	配置当前租户的描述信息。

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 创建租户时系统会自动创建一个Yarn任务队列，并自动以租户名称命名该队列。如果已经存在同名队列，新队列命名为“租户名称-N”。“N”表示从1开始的自然数，存在同名队列的时候N会自动累加以区别已有队列。例如“saletenant”、“saletenant-1”和“saletenant-2”。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：
   - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
   - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

10.7.3.1.2 添加子租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建子租户，将当前租户的资源进一步分配以满足实际使用场景。

前提条件

- 已添加父租户，且属于非叶子租户。
- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择父租户节点然后移单击，打开添加子租户的配置页面，参见表10-40为子租户配置属性。

表 10-40 子租户参数一览

参数名	描述
集群	显示上级父租户所在集群。
父租户资源	显示上级父租户的名称。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划子租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。

参数名	描述
租户类型	指定租户是否是一个叶子租户： ● 选择“叶子租户”：当前租户为叶子租户，不支持添加子租户。 ● 选择“非叶子租户”：当前租户为非叶子租户，支持添加子租户，但租户层级不能超过5层。
计算资源	为当前租户选择动态计算资源。 ● 选择“Yarn”时，系统自动在Yarn中以子租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“Default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 ● 不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量（%）	配置当前租户使用的计算资源百分比，基数为父租户的资源总量。
默认资源池最大容量（%）	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。
存储空间配额	配置当前租户使用的HDFS存储空间配额。 ● 当存储空间配额单位设置为MB时，范围为1～8796093022208，当“存储空间配额单位”设置为GB时，范围为1～8589934592。 ● 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 ● 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 ● 如果此配额大于父租户的配额，实际存储量不超过父租户配额。

参数名	描述
存储路径	配置租户在HDFS中的存储目录。 • 系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“/tenant/ta1”，系统默认自动配置此参数值为“/tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。 • 支持在父目录中自定义存储路径。
描述	配置当前租户的描述信息

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 子租户可以将当前租户的资源进一步分配。每一级别父租户下，直接子租户的资源百分比之和不能超过100%。所有一级租户的计算资源百分比之和也不能超过100%。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：
   - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
   - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

10.7.3.1.3 添加用户并绑定租户的角色

操作场景

创建好的租户不能直接登录集群访问资源，管理员需要通过FusionInsight Manager为已有租户创建新用户，通过绑定租户的角色继承其操作权限，以满足业务使用。

前提条件

管理员已明确业务需求，并已创建了租户。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 用户”。

步骤2 若在系统中添加新的用户，请单击“添加用户”，打开添加用户的配置页面。

若为系统中已有的用户绑定租户权限，请单击该用户所在行的“修改”，打开修改用户的配置页面。

参见**表10-41**为用户配置属性。

表 10-41 用户参数一览

参数名	描述
用户名	指定当前的用户名，长度为3~32个字符，可包含数字、字母、下划线（_）、中划线（-）或空格。 ● “用户名”不能与集群各节点所有操作系统用户名相同，否则此用户无法正常使用。 ● 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。使用“User1”时请输入正确的用户名。
用户类型	可选值包括“人机”和“机机”。 ● “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。 ● “机机”用户：用于应用开发的场景。选择该值用户密码随机生成，无需填写。
密码	选择“人机”用户需填写“密码”。 密码必须包含8~64个字符，至少包含以下类型字符中的四种：大写字母、小写字母、数字、特殊字符和空格。不能与用户名或倒序的用户名相同。
确认密码	再次输入密码。

参数名	描述
用户组	单击“添加”，选择对应用户组将用户添加进去。 ● 如果用户组添加了角色，则用户可获得对应角色中的权限。 ● 例如，为新用户分配Hive的权限，请将用户加入Hive组。
主组	选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。
角色	单击“添加”为用户绑定租户的角色。 **说明** 若一个用户想要获取使用“tenant1”租户包含的资源，且能够为“tenant1”租户添加/删除子租户，则需要同时绑定“Manager_tenant”和“tenant1_*集群ID*”两个角色。
描述	配置当前用户的描述信息。

步骤3 单击“确定”完成用户创建。

----结束

10.7.3.2 管理租户

10.7.3.2.1 管理租户目录

操作场景

管理员通过FusionInsight Manager管理指定租户使用的HDFS存储目录，能根据业务需求对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

已添加具有HDFS存储资源的租户。

查看租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击目标的租户。

步骤3 单击“资源”页签。

步骤4 查看“HDFS存储”表格。

- 指定租户目录的“文件目录数上限”列表示文件和目录数量配额。
- 指定租户目录的“存储空间配额 ”列表示租户目录的存储空间大小。

----结束

添加租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，单击“添加目录”。

- “父目录”，表示当前租户对应父租户的存储目录。

  说明

  当前租户不是子租户则不显示此参数。

- “路径”，填写租户目录的路径。

  说明

  当前租户不是子租户则新路径将在HDFS的根目录下创建。

- “文件\目录数上限”填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  当前已使用的文件数的数据采集周期为1个小时，因此超过文件数阈值的告警上报会存在延迟。

- “存储空间配额”，填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  已使用的存储空间的数据采集周期为1个小时，因此超过存储空间阈值的告警上报会存在延迟。

步骤5 单击“确定”完成租户目录添加。

----结束

修改租户目录属性

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。

- “文件\目录数上限”，填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。
- “存储空间配额”填写租户目录的存储空间大小。

- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

步骤5 单击“确定”完成租户目录修改。

----结束

删除租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“删除”。

说明

不支持删除创建租户时系统创建的租户目录。

步骤5 单击“确定”完成租户目录删除。

----结束

10.7.3.2.2 恢复租户数据

操作场景

租户默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，管理员需要通过FusionInsight Manager手动恢复配置数据。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“租户资源状态”，绿色表示租户可用，灰色表示租户不可用。
2. 单击“资源”，查看“Yarn”或者“HDFS存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击，在弹出的确认窗中输入当前登录的用户密码确认身份，单击“确定”。

步骤5 在“恢复租户资源数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

10.7.3.2.3 删除租户

操作场景

根据业务需求，对于当前不再使用的租户，管理员可以通过FusionInsight Manager删除租户，释放租户占用的资源。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择待删除的租户，单击。

说明

- 根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户的数据。”，否则将自动删除租户对应的存储空间。
- 如果使用不属于supergroup组的用户执行删除租户操作，并且不保留租户数据，需要使用属于supergroup组的用户登录HDFS客户端，手动清理租户对应的存储空间，以免数据残留。

步骤3 单击“确定”，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

租户删除后，Yarn中对应的租户任务队列不会被删除。同时Yarn角色管理中，此租户任务队列不再显示。

----结束

10.7.3.2.4 Capacity Scheduler 模式下清除租户非关联队列

操作场景

在Yarn Capacity Scheduler模式下，删除租户的时候，只是把租户队列的容量设置为0，并且把状态设为“STOPPED”，但是队列在Yarn的服务里面仍然残留。由于Yarn的机制，无法动态删除队列，管理员可以执行命令手动清除残留的队列。

对系统的影响

- 脚本运行过程中会重启controller服务，同步Yarn的配置，并重启主备ResourceManager实例。
- 重启controller服务时，无法登录和操作FusionInsight Manager。
- 重启主备ResourceManager实例后，Yarn组件以及依赖Yarn的组件会出现短暂的服务不可用告警。

前提条件

已删除某个租户，但该租户对应的队列依然存在。

操作步骤

步骤1 确定该租户对应的队列依然存在。

1. 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn”。通过“ResourceManager(主)”链接进入ResourceManager WebUI界面。
2. 单击左侧“Scheduler”界面，可以查看租户对应的队列依然存在，且状态为“STOPPED”，“Configured Capacity”值为0。

步骤2 以**omm**用户登录主管理节点。

步骤3 执行以下目录，执行“cleanQueuesAndRestartRM.sh”脚本。

cd ${BIGDATA_HOME}/om-server/om/sbin

./cleanQueuesAndRestartRM.sh -c *集群ID*

说明

“集群ID”为需执行操作集群ID号，可在FusionInsight Manager的“集群 > *待操作集群的名称* > 集群属性”中查看。

在脚本运行过程中，需输入**yes**及管理员密码。

```
Running the script will restart Controller and restart ResourceManager.
Are you sure you want to continue connecting (yes/no)?yes
Please input admin password:
Begin to backup queues ...
...
```

步骤4 脚本运行成功后，在FusionInsight Manager界面，选择“集群 > *待操作集群名称* > 服务 > Yarn”。通过“ResourceManager(主)”链接进入ResourceManager WebUI界面。

步骤5 单击左侧“Scheduler”界面，确认被删除租户的队列已经清除。

----结束

10.7.3.3 管理资源

10.7.3.3.1 添加资源池

操作场景

在集群中，管理员可从逻辑上对所有Yarn的节点进行分区，使多个NodeManager形成一个Yarn资源池。每个NodeManager只能属于一个资源池。管理员通过FusionInsight Manager添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池，便于指定的队列利用这些计算资源。

系统中默认包含了一个名为“Default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 单击“添加资源池”。

步骤4 设置资源池的属性。

- “集群”：选择待添加资源池的集群名称。
- “名称”：填写资源池的名称。长度为1~50个字符，可包含数字、字母或下划线（_），且不能以下划线（_）开头。
- “资源”：在界面左边可用主机列表中，勾选指定的主机，单击 >> ，将选中的主机加入已选主机列表。只支持选择本集群中的主机。资源池中的主机列表可以为空。

  说明

  根据业务需求，可以通过主机名称 、CPU、内存、操作系统和平台类型，筛选需要选取的资源主机。

步骤5 单击“确定”保存。

完成资源池创建后，管理员可以在资源池的列表中查看资源池的名称、成员、类型。已加入自定义资源池的主机，不再是“Default”资源池的成员。

----结束

10.7.3.3.2 修改资源池

操作场景

根据业务需要，资源池的主机需要调整时，管理员可以通过FusionInsight Manager修改已有资源池中的成员。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“编辑”。

步骤4 在“资源”修改“主机”。

- 增加主机：在界面左边主机列表，选择指定的主机名称，单击 >> ，将选中的主机加入资源池。
- 删除主机：在界面右边主机列表，选择指定的主机名称，单击 << ，将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

10.7.3.3.3 删除资源池

操作场景

根据业务需要，资源池不再使用时，管理员可以通过FusionInsight Manager进行删除资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列容量配置**。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

步骤4 在弹出窗口中单击“确定”。

----结束

10.7.3.3.4 配置队列

操作场景

根据业务需要，管理员可以通过FusionInsight Manager修改指定租户的队列配置。

前提条件

已添加使用Capacity调度器的租户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 动态资源计划”。

默认显示“资源分布策略”。

步骤3 单击“队列配置”页签。

步骤4 “集群”参数选择待操作的集群名称，然后在指定租户资源名的“操作”列，单击“修改”。

说明

- 在“租户资源管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击“队列配置（*队列名*）”名后面的也能打开修改队列配置窗口。
- 一个队列只能绑定一个非Default资源池，即新添加的资源池只能绑定一个队列，作为这个队列的默认资源池。

表 10-42 队列配置参数

参数名	描述
租户资源名（队列）	租户及队列名称。
最大应用数量	表示最大应用程序数量。
AM最大资源百分比	表示集群中可用于运行application master的最大资源占比。
用户资源最小上限百分比（%）	表示每个用户最低资源保障（百分比）。任何时刻，一个队列中每个用户可使用的资源量均有一定的限制。当一个队列中同时运行多个用户的应用程序时，每个用户的使用资源量在一个最小值和最大值之间浮动，其中，最小值取决于正在运行的应用程序数目，而最大值则由此参数决定。 比如，假设此参数的值设置为25。当两个用户向该队列提交应用程序时，每个用户可使用资源量不能超过50%，如果三个用户提交应用程序，则每个用户可使用资源量不能超多33%，如果四个或者更多用户提交应用程序，则每个用户可用资源量不能超过25%。
用户资源上限因子	表示用户使用的最大资源限制因子，与当前租户在集群中实际资源百分比相乘，可计算出用户使用的最大资源百分比。
状态	表示资源计划当前的状态，“运行”为运行状态，“停止”为停止状态。
默认资源池	表示队列使用的资源池，默认为“Default”。 如果需要修改为其他资源池，需要先配置队列容量，请参见**配置资源池的队列容量策略**。

步骤5 单击“确定”完成配置。

----结束

10.7.3.3.5 配置资源池的队列容量策略

操作场景

添加资源池后，需要为Yarn任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。每个队列只能配置一个资源池的队列容量策略。

管理员可以在任何一个资源池中查看队列并配置队列容量策略。配置队列策略后，Yarn任务队列与资源池形成关联关系。

前提条件

已添加队列，即已创建关联了计算资源的租户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 动态资源计划”。

默认显示“资源分布策略”。

步骤3 “集群”参数选择待操作的集群名称，然后在“资源池”，选择待操作的资源池。

步骤4 在“资源分配”表格，指定租户资源名队列的“操作”列，单击“修改”。

步骤5 在“修改资源分配”窗口设置任务队列在此资源池中的资源容量策略。

- “资源容量（%）”：表示当前租户计算资源使用的资源百分比。
- “最大资源容量（%）”：表示当前租户计算资源使用的最大资源百分比。

步骤6 单击“确定”保存配置。

说明

删除队列的资源容量值并保存，可以取消队列在此资源池中的资源容量策略，表示解除队列与资源池的关联关系。需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

----结束

10.7.3.3.6 清除队列容量配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，管理员可以在FusionInsight Manager清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

操作步骤

步骤1 登录FusionInsight Manager界面。

步骤2 选择“租户资源 > 动态资源计划”。

步骤3 “集群”参数选择待操作的集群名称，然后在“资源池”，选择待操作的资源池。

步骤4 在资源分配表格，指定租户资源名的“操作”列，单击“清除”。

步骤5 在弹出的对话框中单击“确定”，清除队列在当前资源池的配置。

----结束

10.7.4 切换调度器

操作场景

新安装的MRS集群默认即使用了Superior调度器，如果是历史版本升级的集群，管理员可以根据以下指导，将Yarn的调度器从Capacity调度器一键式切换到Superior调度器。

前提条件

- 确保集群网络通畅，网络环境安全，Yarn服务状态正常。
- 在切换调度器期间，不允许做添加、删除、修改租户，以及启停服务等操作。

对系统的影响

- 调度器切换过程中，由于要重启Resource Manager，因此切换期间向Yarn提交任务会失败。
- 调度器切换过程中，正在Yarn上面执行的Job的Task任务会继续执行，但不会启动新的Task。
- 调度器切换完成后，在Yarn上面执行的任务有可能会失败进而导致业务中断。
- 调度器切换完成后，在租户管理中将使用Superior的相关参数。
- 调度器切换完成后，Capacity调度器中“资源容量”为“0”的租户队列在Superior调度器中分配不到资源，提交到该租户队列的任务会执行失败。建议在Capacity调度器中不要将租户队列的“资源容量”配置为“0”。
- 调度器切换完成后，在观察期内，不允许对资源池、Yarn节点标签（Label）和租户做添加、删除的操作。若添加或者删除了资源池、Yarn节点标签（Label）和租户的操作，将不支持回退到Capacity调度器。

  说明

  - 切换调度器观察期建议为一周，如果对资源池、Yarn节点标签（Label）和租户做添加、删除的操作，将视为观察期结束。
- 回退可能会丢失部分或者所有的Yarn任务信息。

从 Capacity 调度器切换到 Superior 调度器

步骤1 确保Yarn服务状态正常。

1. 使用管理员帐号，登录FusionInsight Manager系统。
2. 选择“集群 > *待操作的集群名称* > 服务”，查看Yarn服务的状态是否正常。

步骤2 使用**omm**用户登录主管理节点。

步骤3 执行调度器切换。

调度器切换分为三种模式：

0：将Capacity调度器配置转换到Superior，然后将Capacity调度器切换到Superior。

1：只将Capacity调度器配置转换到Superior。

2：只将Capacity调度器切换到Superior。

- 集群环境相对简单，租户数小于20的情况下，建议执行模式0，将Capacity调度器配置转换到Superior的同时切换调度器。

  执行以下命令。

  sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 0**

  说明

  “集群ID”为需执行操作集群ID号，可在FusionInsight Manager的“集群 > *待操作集群的名称* > 集群属性”中查看。

  ```
  Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
  Start to convert Capacity scheduler configurations to Superior. Please wait...
  Convert configurations successfully.
  Start to switch the Yarn scheduler to Superior. Please wait...
  Switch the Yarn scheduler to Superior successfully.
  ```

- 集群环境相对复杂，租户信息复杂，且要求将capacity调度器队列配置信息保留到Superior调度器，建议先执行模式1，将Capacity调度器配置信息转化成Superior配置信息，对转换过来的配置信息做检查后，再执行模式2，将Capacity调度器切换到Superior。

  a. 执行以下命令，将Capacity调度器配置信息转化成Superior配置信息。

     sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 1**

     ```
     Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
     Start to convert Capacity scheduler configurations to Superior. Please wait...
     Convert configurations successfully.
     ```

  b. 执行以下命令，将Capacity调度器切换到Superior。

     sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 2**

     ```
     Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
     Start to switch the Yarn scheduler to Superior. Please wait...
     Switch the Yarn scheduler to Superior successfully.
     ```

- 不保存Capacity调度器队列配置，建议直接执行模式2，只切换调度器，不转换配置。

  a. 登录FusionInsight Manager，删除除了default租户的所有租户。

  b. 登录FusionInsight Manager，删除除了default资源池的所有资源池。

     执行以下命令，将Capacity调度器切换到Superior。

     sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 2**

     ```
     Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
     Start to switch the Yarn scheduler to Superior. Please wait...
     Switch the Yarn scheduler to Superior successfully.
     ```

说明

登录主管理节点，可查看调度器切换的日志信息。

- ${BIGDATA_LOG_HOME}/controller/aos/switch_scheduler.log
- ${BIGDATA_LOG_HOME}/controller/aos/aos.log

----结束

10.8 系统设置

10.8.1 权限设置

10.8.1.1 用户管理

10.8.1.1.1 创建用户

操作场景

FusionInsight Manager最大支持50000个用户（包括系统内置用户）。默认情况下，系统只有一个用户“admin”具有FusionInsight Manager最高操作权限。管理员应根据实际业务场景需要，通过FusionInsight Manager创建新用户并指定其操作权限以满足业务使用。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在用户列表上方，单击“添加用户”。

步骤4 填写“用户名”。用户名由数字、字母、下划线、中划线（-）或空格组成，不区分大小写，不能与系统或操作系统中已有的用户名相同。

步骤5 设置“用户类型”，可选值包括“人机”和“机机”。

- “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。
- “机机”用户：用于组件应用开发的场景。选择该值则用户密码随机生成，无需填写。

步骤6 根据业务实际需要，在“用户组”，单击“添加”，选择一个或多个用户组添加到列表中。

说明

- 如果选中的用户组绑定了角色或者在Ranger中配置了权限策略，用户将获得对应的权限。
- 安装FusionInsight Manager后默认生成的部分用户组包含特殊权限，请根据界面上用户组描述信息选择正确的用户组。
- 如果已有的用户组无法满足使用，可以单击“创建新用户组”先创建用户组，参见**添加用户组**。

步骤7 根据业务实际需要，在“用户组”添加的所有组中选择一个组作为用户创建目录和文件的主组。

下拉列表包含“用户组”中添加的全部组。

说明

由于一个用户可以属于多个组（包括主组和附属组，主组只有一个，附属组可以有多个），设置用户的主组是为便于维护以及遵循hadoop社区的权限机制。此外用户的主组和其他组在权限控制方面，作用一致。

步骤8 根据业务实际需要，在“角色”，单击“添加”，为单个用户绑定角色。

说明

- 创建用户时添加角色可细化用户的权限。
- 创建用户时，如果用户从用户组获得的权限还不满足业务需要，则可以再分配其他已创建的角色。也可以单击“创建新角色”先创建角色，参见**添加角色**。

  为新用户分配角色授权，最长可能需要3分钟时间生效，如果从用户组获得的权限已满足使用，则无需再添加角色。
- 组件启用Ranger鉴权后，除系统默认用户组或角色的权限外，其他权限需要通过配置Ranger策略为用户赋权。
- 若用户既没有加入用户组也没有设置角色，通过此用户登录FusionInsight Manager后，用户将无权查看或操作。

步骤9 根据业务实际需要填写“描述”。

步骤10 单击“确定”完成用户创建。

“人机”用户创建成功后，通常需要修改初始密码后才可以正常使用，可以使用该用户登录FusionInsight Manager，按照界面提示重置密码即可。

----结束

10.8.1.1.2 修改用户信息

操作场景

管理员可以在FusionInsight Manager修改已创建的用户信息，包括修改用户组、主组、角色分配权限和描述。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要修改信息的用户所在行，单击“修改”。

根据实际情况，修改对应参数。

说明

修改用户的用户组，或者修改用户的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

10.8.1.1.3 导出用户信息

操作场景

管理员可以在FusionInsight Manager导出所有已创建的用户信息。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 单击“导出全部”，可一次性导出所有用户信息。

用户信息包含以下几个字段：用户名、创建时间、描述、用户类型（0表示人机帐号，1表示机机帐号）、主组、用户组列表、绑定的角色列表。

步骤4 在“保存类型”选择“TXT”或“CSV”。单击“确定”开始导出。

----结束

10.8.1.1.4 锁定用户

操作场景

由于业务变化，用户可能长期暂停使用，为了保证安全，管理员可以锁定用户。

锁定用户的方法包含以下两种方式：

- 自动锁定：通过设置密码策略中的“密码连续错误次数”，将超过登录失败次数的用户自动锁定。具体操作请参见**配置密码策略**。
- 手动锁定：由管理员通过手动的方式将用户锁定。

以下将具体介绍手动锁定。不支持锁定“机机”用户。

对系统的影响

用户被锁定后，不能在FusionInsight Manager重新登录或在集群中重新进行身份认证。锁定后的用户需要管理员手动解锁或者等待锁定时间结束才能恢复使用。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要锁定用户所在行，单击“锁定”。

步骤4 在弹出的窗口勾选“我已阅读此信息并了解其影响。”，单击“确定”完成锁定操作。

----结束

10.8.1.1.5 解锁用户

操作场景

在用户输入错误密码次数大于允许输入错误次数，造成用户被锁定的场景下，管理员可以通过FusionInsight Manager为锁定的用户解锁。仅支持解锁使用FusionInsight Manager创建的用户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要解锁用户所在行，单击“解锁”。

步骤4 在弹出的窗口勾选“我已阅读此信息并了解其影响。”，单击“确定”完成解锁操作。

----结束

10.8.1.1.6 删除用户

操作场景

根据业务需要，管理员应在FusionInsight Manager删除不再使用的系统用户。

说明

- 用户删除后，已经发放的TGT在24小时内仍然有效，用户可以使用该TGT继续进行安全认证并访问系统。
- 如新建用户与已删除用户同名，则会继承已删除用户的拥有的所有Owner权限。建议根据实际业务需求决定是否删除该用户持有的资源。例如HDFS上的文件。
- 默认的**admin**用户无法删除。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要删除用户所在行，选择“更多 > 删除”。

说明

如果需要批量删除多个用户，勾选需要删除的用户后直接单击“删除”即可。

步骤4 在弹出的窗口单击“确定”完成删除操作。

----结束

10.8.1.1.7 修改用户密码

操作场景

出于安全的考虑，“人机”类型系统用户密码必须定期修改。

如果用户具备使用FusionInsight Manager的权限时，可以通过FusionInsight Manager完成修改自身密码工作。

如果用户不具备使用FusionInsight Manager的权限时，可以通过客户端修改自身密码。

前提条件

- 从管理员获取当前的密码策略。
- 已在集群内的任一节点安装了客户端，并获取此节点IP地址。请联系管理员获取客户端安装用户密码。

使用 FusionInsight Manager 修改密码

步骤1 登录FusionInsight Manager。

步骤2 移动鼠标到界面右上角的用户名。

在弹出菜单中单击“修改密码”。

步骤3 在“密码修改界面”分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];',<.>/\?。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。

----结束

使用客户端修改密码

步骤1 以客户端安装用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，修改系统用户密码。此操作对所有服务器生效。

kpasswd *系统用户名称*

例如，修改系统用户“test1”，执行**kpasswd test1**。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];',<.>/\?。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。

说明

如果kpasswd命令运行出错，可以尝试：

- 关闭ssh会话再重新打开。
- 执行**kdestroy**命令后再执行**kpasswd**。

----结束

10.8.1.1.8 初始化用户密码

操作场景

用户如果遗忘密码或公共帐号密码需要定期修改时，管理员可通过FusionInsight Manager初始化密码。初始化密码后系统用户首次使用帐号需要修改密码。

说明

此操作仅支持“人机”用户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要初始化密码用户所在行，选择“更多 > 初始化密码”。在弹出窗口中输入当前登录的管理员用户密码确认身份，单击“确定”，在确认对话框单击“确定”。

步骤4 填写“新密码”和“确认新密码”，单击“确定”。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];',<.>/\?。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。

----结束

10.8.1.1.9 导出认证凭据文件

操作场景

用户为安全模式集群进行应用开发的场景下，需要获取用户keytab文件用于安全认证。管理员可以通过FusionInsight Manager导出keytab文件。

说明

修改用户密码后，之前导出的keytab将失效，需要重新导出。

前提条件

下载“人机”用户的认证凭据文件前，需要使用Manager界面或者客户端修改过一次此用户的密码，否则下载获取的keytab文件无法使用。请参见**修改用户密码**。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在需导出文件的用户所在行，选择“更多 > 下载认证凭据”，待文件自动生成后指定保存位置，并妥善保管该文件。

认证凭据中会携带kerberos服务的“krb5.conf”文件。

解压认证凭据文件后可以获取两个文件：

- “krb5.conf”文件包含认证服务连接信息。
- “user.keytab”文件包含用户认证信息。

----结束

10.8.1.2 用户组管理

操作场景

FusionInsight Manager最大支持5000个用户组（包括系统内置用户组）。根据不同业务场景需要，管理员使用FusionInsight Manager创建并管理不同用户组。用户组通过绑定角色获取操作权限，用户加入用户组后，可获得用户组具有的操作权限。用户组同时可以达到对用户进行分类并统一管理多个用户。

前提条件

- 管理员已明确业务需求，并已创建业务场景需要的角色。
- 已登录FusionInsight Manager。

添加用户组

步骤1 选择“系统 > 权限 > 用户组”。

步骤2 在组列表上方，单击“添加用户组”。

步骤3 填写“组名”和“描述”。

“组名”由数字、字母、或下划线、中划线（-）或空格组成，不区分大小写，长度为1～64位，不能与系统中已有的用户组名相同。

步骤4 在“角色”，单击“添加”选择指定的角色并添加。

说明

- 对于已启用Ranger授权的组件（HDFS与Yarn除外），Manager上非系统默认角色的权限将无法生效，需要通过配置Ranger策略为用户组赋权。
- HDFS与Yarn的资源请求在Ranger中的策略条件未能覆盖的情况下，组件ACL规则仍将生效。

步骤5 在“用户”，单击“添加”选择指定的用户并添加。

步骤6 单击“确定”完成用户组创建。

----结束

查看用户组信息

用户组列表默认显示所有用户。单击指定用户组名称左侧的箭头展开详细信息，可以查看此用户组中的用户数、用户以及绑定的角色。

修改用户组信息

在要修改信息用户组所在的行，单击“修改”，修改用户组信息。

导出用户组信息

单击“导出全部”，可一次性导出所有用户组信息，可导出“TXT”或者“CSV”格式。

用户组信息包含以下几个字段：用户组名、描述、用户列表、角色列表。

删除用户组

在要删除用户组所在行，单击“删除”。如果需要批量删除多个用户组，勾选需要删除的用户组后再单击列表上方“删除”即可。用户组中包含用户时，不允许删除。如需删除，请先通过修改用户组删除其包含的所有用户，再删除该用户组。

10.8.1.3 角色管理

操作场景

FusionInsight Manager最大支持5000个角色（包括系统内置角色，不包括租户自动创建的角色）。根据不同业务场景需要，管理员使用FusionInsight Manager创建并管理不同角色，通过角色对Manager和组件进行授权管理。

前提条件

- 管理员已明确业务需求。
- 登录FusionInsight Manager。

添加角色

步骤1 选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

“角色名称”由数字、字母、或下划线组成，长度为3～50位，不能与系统中已有的角色名相同。

步骤3 在“配置资源权限”列表，选择待增加权限的集群，为角色选择服务权限。

在设置组件的权限时，可通过右上角的“搜索”框输入资源名称，然后单击搜索图标显示搜索结果。

搜索范围仅包含当前权限目录，无法搜索子目录。搜索关键字支持模糊搜索，不区分大小写。

说明

- 对于已启用Ranger授权的组件（HDFS与Yarn除外），Manager上非系统默认角色的权限将无法生效，需要通过配置Ranger策略为用户组赋权。
- HDFS与Yarn的资源请求在Ranger中的策略条件未能覆盖的情况下，组件ACL规则仍将生效。
- 设置组件的权限时，每次最大支持1000条权限。

步骤4 单击“确定”完成。

----结束

修改角色信息

在要修改信息角色所在的行，单击“修改”。

导出角色信息

单击“导出全部”，可一次性导出所有角色信息，可导出“TXT”或者“CSV”格式文件。

角色信息包含以下几个字段：角色名、描述、是否默认角色。

删除角色

在要删除角色所在行，单击“删除”。如果需要批量删除多个角色，勾选需要删除的角色后单击列表上方“删除”即可。角色被用户绑定时不可删除；如需删除，请先通过修改用户解除角色和用户之间的关联，再删除该角色。

任务示例（创建 Manager 角色）

步骤1 选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，在“角色名称”和“描述”输入角色名字与描述。

步骤3 在“配置资源权限”区域选择“Manager”，按照以下说明设置角色“权限”。

Manager权限：

- Cluster：
  - 查看权限：“集群”页面查看权限、“运维 > 告警”页面下“告警”、“事件”的查看权限。
  - 管理权限：“集群”、“运维”页面的管理权限。
- User：
  - 查看权限：“系统”页面下“权限”区域中内容的查看权限。
  - 管理权限：“系统”页面下“权限”区域中内容的管理权限。
- Audit：

  管理权限：“审计”页面信息的管理权限。
- Tenant：

  管理权限：“租户”页面管理权限；“运维 > 告警”页面下“告警”、“事件”的查看权限。
- System：

  管理权限：“系统”页面除“权限”区域外，其他区域的管理权限；“运维 > 告警”页面下“告警”、“事件”的查看权限。

步骤4 单击“确定”完成。

----结束

10.8.1.4 安全策略

10.8.1.4.1 配置密码策略

操作场景

根据业务安全需要，管理员可以在FusionInsight Manager设置密码安全规则、用户登录安全规则及用户锁定规则。

须知

- 密码策略涉及用户管理的安全性，请根据企业安全要求谨慎修改，否则会有安全性风险。
- 修改密码策略之后，再修改用户密码，此时新的密码策略才会生效。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 密码策略”。

步骤3 具体参数参见**表10-43**。

表 10-43 密码策略参数说明

参数名称	描述
最小密码长度	密码包含的最小字符个数，取值范围为8～64。默认值为“8”。
字符类型的数目	密码字符包含大写字母、小写字母、数字和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=和空格）的最小种类。可选择数值为“4”和“5”。默认值“4”表示可使用大写字母、小写字母、数字、特殊符号，选择“5”表示可使用全部。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。默认值为“5”。
用户锁定时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。默认值为“5”。
密码有效期（天）	密码有效使用天数：取值范围0～90，0表示永久有效，默认值为“90”。
重复使用规则	修改密码时，不允许使用最近N次使用过的密码，N=1～5，默认为“1”。此策略只影响“人机”用户。

参数名称	描述
密码失效提前提醒天数	密码失效提前提醒天数：表示提醒密码失效到密码真正失效的天数。提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录FusionInsight Manager界面时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒，默认值为“5”。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔，取值范围为0～1440。“0”表示永远有效，“1440”表示1天。默认值为“5”。

步骤4 单击“确定”保存配置。修改密码策略之后，再修改用户密码，此时新的密码策略才会生效。

----结束

10.8.1.4.2 配置私有属性

操作场景

admin用户或绑定Manager_administrator角色的管理员用户，可以在FusionInsight Manager配置私有属性功能开关，用于支持用户（集群中所有业务用户）设置或取消自己的私有（Independent）属性。

开启私有属性开关后，需要业务用户登录后设置Independent属性，完成用户私有属性配置。

限制约束

- 管理员不能设置或取消业务用户的Independent属性。
- 管理员不能获取私有用户的认证凭据。

前提条件

已获取要求权限的管理员用户和密码。

操作步骤

配置私有属性开关

步骤1 以admin用户或绑定Manager_administrator角色的用户登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 配置Independent”。

步骤3 打开或关闭Independent属性，根据提示输入密码，单击“确认”完成身份验证。

步骤4 身份验证通过后，等待修改OMS配置完成，单击“完成”结束操作。

说明

关闭Independent属性功能后：

- 已拥有这个属性的业务用户可以在右上角用户名下取消Independent属性，取消后无法重新设置。取消后已创建的私有表继续保持私有属性，取消后无法继续创建私有表。
- 没有这个属性的业务用户无法在右上角用户名下进行设置和取消操作。

配置用户私有属性

步骤5 以业务用户登录FusionInsight Manager。

须知

设置Independent属性后，管理员不能初始化私有用户（业务用户设置了Independent属性后，即为私有用户）的密码；如果忘记此用户密码，密码将无法找回。

admin用户无法设置Independent属性。

步骤6 移动鼠标到界面右上角的用户名。

步骤7 在弹出的菜单栏中单击“设置Independent”或“取消Independent”。

说明

- 私有属性功能开关已开启，业务用户当前已设置私有属性时，菜单栏显示“取消Independent”。
- 私有属性功能开关已开启，业务用户当前已取消私有属性时，菜单栏显示“设置Independent”。
- 私有属性功能开关已关闭，业务用户当前已设置私有属性时，菜单栏显示“取消Independent”。
- 私有属性功能开关已关闭，业务用户当前已取消私有属性时，菜单栏不显示。

步骤8 根据界面提示，输入密码，单击“确定”完成身份验证。

步骤9 身份验证通过后，在确认对话框中单击“确定”。

----结束

10.8.2 对接设置

10.8.2.1 配置 SNMP 北向参数

操作场景

如果用户需要在统一的运维网管平台查看集群的告警、监控数据，管理员可以在FusionInsight Manager使用SNMP服务将相关数据上报到网管平台。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 对接 > SNMP”。

步骤3 单击“SNMP服务”右侧的开关。

“SNMP服务”默认为不启用，开关显示为 表示启用。

步骤4 根据表10-44所示的说明填写对接参数。

表 10-44 对接参数

参数名称	参数说明
版本	SNMP协议版本号，取值范围： ● V2C：低版本，安全性较低。 ● V3：高版本，安全性更高。 推荐使用V3版本。
本地端口	本地端口，默认值“20000”，取值范围“1025”到“65535”。
读团体名	该参数仅在设置“版本”为v2c时可用，用于设置只读团体名。
写团体名	该参数仅在设置“版本”为v2c时可用，用于设置可写团体名。
安全用户名	该参数仅在设置“版本”为v3时可用，用于设置协议安全用户名。
认证协议	该参数仅在设置“版本”为v3时可用，用于设置认证协议，推荐选择SHA。
认证密码	该参数仅在设置“版本”为v3时可用，用于设置认证密钥。
确认认证密码	该参数仅在设置“版本”为v3时可用，用于确认认证密钥。
加密协议	该参数仅在设置“版本”为v3时可用，用于设置加密协议，推荐选择AES256。
加密密码	该参数仅在设置“版本”为v3时可用，用于设置加密密钥。
确认加密密码	该参数仅在设置“版本”为v3时可用，用于确认加密密钥。

说明

- “安全用户名”中禁止出现以64的公因子（1、2、4、8等）为单位长度的重复字符串，例如abab，abcdabcd。
- “认证密码”和“加密密码”密码长度为8到16位，至少需要包含大写字母、小写字母、数字、特殊字符中的3种类型字符。两个密码不能相同。两个密码不可和安全用户名相同或安全用户名的倒序字符相同。
- 使用SNMP协议从安全方面考虑，需要定期修改“认证密码”和“加密密码”密码。
- 使用SNMP v3版本时，安全用户在5分钟之内连续鉴权失败5次将被锁定，5分钟后自动解锁。

步骤5 单击“添加Trap目标”，在弹出的“添加Trap目标”对话框中填写以下参数：

- 目标标识：Trap目标标识，一般指接收Trap的网管或主机标识。长度限制1～255字节，一般由字母或数字组成。
- 目标IP模式：目标IP的IP地址模式，可选择“IPV4”或者“IPV6”。
- 目标IP：目标IP，要求可与管理节点的管理平面IP地址互通。
- 目标端口：接收Trap的端口，要求与对端保持一致，取值范围“0～65535”。
- Trap团体名：该参数仅在设置版本为V2C时可用，用于设置主动上报团体名。

单击“确定”，设置完成，退出“添加Trap目标”对话框。

步骤6 单击“确定”，设置完成。

----结束

10.8.2.2 配置 Syslog 北向参数

操作场景

如果用户需要在统一的告警平台查看集群的告警和事件，管理员可以在FusionInsight Manager使用Syslog协议将相关数据上报到告警平台。

须知

Syslog协议未做加密，传输数据容易被窃取，存在安全风险。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 对接 > Syslog”。

步骤3 单击“Syslog服务”右侧的开关。

“Syslog服务”默认为不启用，开关显示为 表示启用。

步骤4 根据表10-45所示的说明填写北向参数。

表 10-45 Syslog 对接参数

参数区域	参数名称	参数说明
Syslog协议	服务IP模式	设置对接服务器IP地址模式，可选择“IPV4”或者“IPV6”。
	服务IP	设置对接服务器IP地址。
	服务端口	设置对接端口。
	协议	设置协议类型，可选值： • TCP • UDP
	安全级别	设置上报消息的严重程度，取值范围： • Emergency • Alert • Critical • Error • Warning • Notice • Informational（默认值） • Debug **说明** “安全级别”和“Facility”共同组成发出消息的优先级（**Priority**）。 优先级（**Priority**）=“Facility” × 8 +“安全级别” “安全级别”和“Facility”各项对应的数值请参考**表10-46**。
	Facility	设置产生日志的模块。可选项参考**表10-46**，推荐使用默认值“local use 0 (local0)”。
	标识符	设置产品标识，默认为“FusionInsight Manager”。 标识符可以包含字母、数字、下划线、 空格、|、 $、{、 }、 点、中划线，并且不能超过256个字符。
报告信息	报文格式	设置告警报告的消息格式，具体要求请参考界面帮助。 报文格式可以包含字母、数字、下划线、 空格、|、 $、{、 }、 点、中划线，并且不能超过1024个字符。 **说明** 报文格式中信息域的说明请参考**表10-47**。
	报告信息类型	设置需要上报的告警类型。
	上报消息级别	设置需要上报的告警级别。

参数区域	参数名称	参数说明
未恢复告警上报设置	周期上报未恢复告警	设置是否按指定周期上报未清除的告警。打开开关表示启用此功能，关闭开关表示不启用。开关默认为关闭。
	间隔时间（分钟）	设置周期上报告警的时间间隔，当“周期上报未恢复告警”开关设置为打开时启用。单位为分钟，默认值为“15”，支持范围为“5”到“1440”（1天）。
心跳设置	上报心跳	设置是否开启周期上报Syslog心跳消息。打开开关表示开启此功能，关闭开关表示不启用。开关默认为关闭。
	心跳周期（分钟）	设置周期上报心跳的时间间隔，当“上报心跳”开关设置为打开时启用。单位为分钟，默认值为“15”，支持范围为“1”到“60”。
	心跳报文	设置心跳上报的内容，当“上报心跳”开关设置为打开时启用，不能为空。支持数字、字母、下划线、竖线、冒号、空格、英文逗号和句号字符，长度小于等于256。

说明

设置周期上报心跳报文后，在某些集群容错自动恢复的场景下（例如主备OMS倒换）可能会出现报文上报中断的现象，此时等待自动恢复即可。

步骤5 单击“确定”，设置完成。

----结束

参考信息

表 10-46 “安全级别”和“Facility”**字段数值编码**

安全级别	Facility	数值编码
Emergency	kernel messages	0
Alert	user-level messages	1
Critical	mail system	2
Error	system daemons	3
Warning	security/authorization messages (note 1)	4
Notice	messages generated internally by syslogd	5
Informational	line printer subsystem	6
Debug	network news subsystem	7
-	UUCP subsystem	8

安全级别	Facility	数值编码
-	clock daemon (note 2)	9
-	security/authorization messages	10
-	FTP daemon	11
-	NTP subsystem	12
-	log audit (note 1)	13
-	log alert (note 1)	14
-	clock daemon	15
-	local use 0~7 (local0 ~ local7)	16~23

表 10-47 报文格式信息域表

信息域	描述
dn	集群名称
id	告警ID
name	告警名称
serialNo	告警序列号 **说明** 故障告警及其对应的恢复告警的告警序列号相同。
category	告警类型，取值范围： ● 0：故障告警 ● 1：恢复告警 ● 2：事件
occurTime	告警产生时间
clearTime	告警清除时间
isAutoClear	告警是否自动清除，取值范围： ● 1：是 ● 0：否
locationInfo	告警位置信息
clearType	告警清除类型，取值范围： ● -1：未清除 ● 0：自动清除 ● 2：手动清除

信息域	描述
level	告警级别，取值范围： • 1：紧急告警 • 2：重要告警 • 3：次要告警 • 4：提示告警
cause	告警原因
additionalInfo	附加信息
object	告警对象

10.8.2.3 配置监控指标数据转储

操作场景

监控数据上报功能可以将系统中采集到的监控数据写入到文本文件，并以FTP或SFTP的形式上传到指定的服务器中。

使用该功能前，管理员需要在FusionInsight Manager页面进行相关配置。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 对接 > 监控数据上传”。

步骤3 单击“监控数据上传”右边的开关。

“监控数据上传”默认为不启用，开关显示为 表示启用。

步骤4 根据表10-48所示的说明填写上传参数：

表 10-48 上传配置参数

参数名称	参数说明
FTP IP地址模式	必选参数，指定服务器IP地址模式，可选择“IPV4”或“IPV6”。
FTP IP地址	必选参数，指定监控指标数据对接后存放监控文件的FTP服务器。
FTP端口	必选参数，指定连接FTP服务器的端口。
FTP用户名	必选参数，指定登录FTP服务器的用户名。
FTP密码	必选参数，指定登录FTP服务器的密码。
保存路径	必选参数，指定监控文件在FTP服务器保存的路径。

参数名称	参数说明
转储时间间隔（秒）	必选参数，指定监控文件在FTP服务器保存的周期，单位为秒。
转储模式	必选参数，指定监控文件发送时使用的协议。可选协议为“FTP”和“SFTP”。建议使用基于SSH v2的SFTP模式，否则可能存在安全风险。
SFTP服务公钥	可选参数，指定FTP服务器的公共密钥，“模式”选择“SFTP”时此参数生效。

步骤5 单击“确定”，设置完成。

说明

选择转储模式为SFTP，当SFTP服务公钥为空时，先进行安全风险提示，确定安全风险后再保存配置。

----结束

数据格式

配置完成后，监控上报功能会将集群中监控数据周期性的写入到文本文件中，并根据用户配置的上报周期，将这些文件上报到对应的FTP/SFTP服务中。

- 监控文件产生规则
  - 按照指标的采集周期，监控指标会被分别写入到每30s，60s，以及300s产生的文件

    30s周期：默认采集周期为30s的实时指标。

    60s周期：默认采集周期为60s的实时指标。

    300s周期：非30s、60s采集的所有指标。
  - 文件名格式：*metirc_{周期}_{文件创建时间YYYYMMDDHHMMSS }.log*

    例如：metric_60_20160908085915.log

    metric_300_20160908085613.log
- 监控文件内容
  - 监控写入文件格式：

    “集群ID|集群名称|显示名称|服务名称|指标ID|采集时间|采集主机|单位|指标值”，其中：各字段间以“|”分隔，例如：

    ```
    1|xx1|Host|Host|10000413|2019/06/18 10:05:00|189-66-254-146|KB/s|309.910
    1|xx1|Host|Host|10000413|2019/06/18 10:05:00|189-66-254-152|KB/s|72.870
    2|xx2|Host|Host|10000413|2019/06/18 10:05:00|189-66-254-163|KB/s|100.650
    ```

    说明：实际的文件中不存在对应的文件格式标题。
  - 监控文件上传间隔：

    监控文件上传时间间隔可以在页面通过“转储时间间隔（秒）”配置，目前支持30s-300s之间均可。配置完成后，系统会按照指定的时间间隔，将文件定期上传到对应的FTP/SFTP服务器。
- 监控指标说明文件
  - 指标全集文件

指标全集文件all-shown-metric-zh_CN包括了所有指标的详细信息。第三方系统从上报的文件内容中解析出指标id后，可以通过查询指标全集文件获取指标详细信息。

指标全集文件位置：

主备OMS节点：{FusionInsight安装路径}/om-server/om/etc/om/all-shown-metric-zh_CN

指标全集文件内容参考：

```
实时指标ID,5分钟指标ID,指标名称,指标采集周期(秒),是否默认采集,指标所属服务,指标所属角色
00101,10000101,JobHistoryServer非堆内存使用量,30,false,Mapreduce,JobHistoryServer
00102,10000102,JobHistoryServer非堆内存分配量,30,false,Mapreduce,JobHistoryServer
00103,10000103,JobHistoryServer堆内存使用量,30,false,Mapreduce,JobHistoryServer
00104,10000104,JobHistoryServer堆内存分配量,30,false,Mapreduce,JobHistoryServer
00105,10000105,阻塞线程数,30,false,Mapreduce,JobHistoryServer
00106,10000106,运行线程数,30,false,Mapreduce,JobHistoryServer
00107,10000107,GC时间,30,false,Mapreduce,JobHistoryServer
00110,10000110,JobHistoryServer的CPU使用率,30,false,Mapreduce,JobHistoryServer
...
```

- 重要指标字段说明

  实时指标ID：指标的采集周期为30s/60s的指标ID，一个独立的指标项只可能存在30s或者60s的实时指标项。

  5分钟指标ID：指标对应的5分钟（300s）的指标ID。

  指标采集周期(秒)：主要是针对实时指标的采集周期，可选值为30或60。

  指标所属服务：指标所属的服务名名称，标明指标所属的服务类型，如HDFS、HBase等。

  指标所属角色：指标所属的角色名名称，标明指标所属的实际角色类型，如JobServer、RegionServer等。

- 解析说明

  针对采集周期为30s/60s的指标，参考该指标说明文件的是第1列，即**实时指标ID**即可找到对应的指标说明。

  针对采集周期为300s的指标，参考该指标说明文件对应的第2列，即**5分钟指标ID**即可找到对应的指标说明。

10.8.3 导入证书

操作场景

CA证书用于FusionInsight Manager各个模块、集群的组件客户端与服务端在通信过程中加密数据，实现安全通信。FusionInsight Manager支持快速导入CA证书，以确保产品安全使用。适用于以下场景：

- 首次安装好集群以后，需要更换企业证书。
- 企业证书有效时间已过期或安全性加强，需要更换为新的证书。

对系统的影响

- 更换证书过程中集群需要重启，此时系统无法访问且无法提供服务。
- 更换证书以后，所有组件和Manager的模块使用的证书将自动更新。
- 更换证书以后，还未信任该证书的本地环境，需要重新安装证书。

前提条件

- 证书文件和密钥文件可向企业证书管理员申请或由管理员生成。
- 获取需要导入到集群的CA证书文件（*.crt）、密钥文件（*.key）以及保存访问密钥文件密码的文件（password.property）。证书名称和密钥名称支持大小写字母和数字。以上文件在生成以后需要打包成tar格式压缩包。
- 准备一个访问密钥文件的密码用于访问密钥文件。

  密码复杂度要求如下，如果密码复杂度不满足如下要求，可能存在安全风险：
  - 密码字符长度最小为8位。
  - 至少需要包含大写字母、小写字母、数字、特殊字符~`!?,.;-_'(){}[]/<>@#$%^&*+|\=中的4种类型字符。
- 向证书管理员申请证书时，请提供访问密钥文件的密码并申请crt、cer、cert和pem格式证书文件，以及key和pem格式密钥文件。申请的证书需要有签发功能。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 证书”。

步骤2 在“上传证书”右侧单击 ···，在文件窗口中浏览已获取的证书文件tar压缩包并确认选择此文件。

步骤3 单击上传文件，Manager将上传压缩包并自动执行导入操作。

步骤4 导入完成后提示同步集群配置并重启WEB服务使新证书生效，单击“确定”。

步骤5 在弹出窗口输入当前登录用户密码验证身份，单击“确定”自动同步集群配置并重启WEB服务。

步骤6 重启完成后在浏览器地址栏中，输入并访问FusionInsight Manager的网络地址，验证能否正常打开页面。

步骤7 登录FusionInsight Manager。

步骤8 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 重启”。

步骤9 输入当前登录的用户密码确认身份，单击“确定”。

----结束

10.8.4 OMS 管理

10.8.4.1 OMS 维护页面概述

总览

登录FusionInsight Manager以后，选择“系统 > OMS”后，打开OMS维护页面，管理员可以在此页面对OMS进行维护操作，包含查看基本信息、查看OMS业务模块的服务状态，也可以手工触发健康检查。

基本信息

FusionInsight Manager支持显示当前OMS的关联信息，包含如**表10-49**所示内容：

表 10-49 OMS 信息说明

项目	说明
版本	表示OMS版本，与FusionInsight Manager版本相同。
IP模式	表示当前集群网络的IP地址模式。
HA模式	表示OMS工作模式，由安装FusionInsight Manager时的配置文件指定。
当前主用	表示OMS主进程节点主机名，即主管理节点主机名。单击主机名可进入对应的主机详情页面。
当前备用	表示OMS备进程节点主机名，即备管理节点主机名。单击主机名可进入对应的主机详情页面。
持续时间	表示OMS进程启动持续的时间。

OMS 服务状态

FusionInsight Manager支持显示OMS所有业务模块的运行状态，每个业务模块的状态显示为●表示运行正常。

健康检查

管理员可以在OMS维护页面单击“健康检查”开始为OMS的状态进行检查。如果某些检查项存在问题，可直接打开检查说明进行处理。

进入/退出维护模式

配置OMS进入或退出维护模式。

系统参数

在大集群场景下对接DMPS集群。

10.8.4.2 修改 OMS 服务配置参数

操作场景

根据用户环境的安全要求，管理员可以在FusionInsight Manager修改OMS中Kerberos与LDAP配置。

对系统的影响

修改OMS的服务配置参数后，需要重启对应的OMS模块，此时FusionInsight Manager将无法正常使用。

操作步骤

修改okerberos配置

步骤1 登录FusionInsight Manager，选择“系统 > OMS”。

步骤2 在okerberos所在行，单击“修改配置”。

步骤3 根据表10-50所示的说明修改参数。

表 10-50 okerberos 参数配置一览表

参数名	说明
连接KDC最大时延（毫秒）	应用连接到Kerberos的超时时间，单位为毫秒，请填写整数值。
最大尝试次数	应用连接到Kerberos的最大重试次数，请填写整数值。
操作Ldap最大时延（毫秒）	Kerberos连接LDAP的超时时间，单位为毫秒。
搜索Ldap最大时延（毫秒）	Kerberos在LDAP查询用户信息的超时时间，单位为毫秒。
Kadmin监听端口	kadmin服务的端口。
KDC监听端口	kinit服务的端口。
Kpasswd监听端口	kpasswd服务的端口。

步骤4 单击“确定”。

在弹出窗口输入当前登录用户密码验证身份，单击“确定”，在确认重启的对话框中单击“确定”。

修改oldap配置

步骤5 在oldap所在行，单击“修改配置”。

步骤6 根据表10-51所示的说明修改参数。

表 10-51 oldap 参数配置一览表

参数名	说明
Ldap服务监听端口	LDAP服务端口号。

步骤7 单击“确定”。

在弹出窗口输入当前登录用户密码验证身份，单击“确定”，在确认重启的对话框中单击“确定”。

说明

如果重置LDAP帐户密码需要重启ACS，操作步骤如下：

1. 使用PuTTY，以**omm**用户登录主管理节点，执行以下命令更新域配置：
   sh ${BIGDATA_HOME}/om-server/om/sbin/restart-RealmConfig.sh
   提示以下信息表示命令执行成功：
   Modify realm successfully. Use the new password to log into FusionInsight again.
2. 执行**sh $CONTROLLER_HOME/sbin/acs_cmd.sh stop**，停止ACS。
3. 执行**sh $CONTROLLER_HOME/sbin/acs_cmd.sh start**，启动ACS。

重启集群

步骤8 登录FusionInsight Manager，参考**滚动重启集群**章节，重启集群。

----结束

10.8.5 部件管理

10.8.5.1 查看部件包

操作场景

完整的MRS集群由多个部件包组成，FusionInsight Manager单独安装某些服务前需要检查此服务对应的部件包是否已安装。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 部件”。

步骤2 在“已安装部件”查看所有部件列表。

说明

在“平台类型”列可查看部件已注册的OS及平台类型。

步骤3 单击部件名称左侧的，可查看部件包含的服务及其版本号。

----结束

10.9 集群管理

10.9.1 配置客户端

10.9.1.1 安装客户端

操作场景

该操作指导安装工程师安装MRS集群所有服务（不包含Flume）的客户端。MRS针对不同服务提供了Shell脚本，供开发维护人员在不同场景下登录其对应的服务维护客户端完成对应的维护任务。

说明

- 通过Manager界面修改服务端配置或系统升级后，建议重新安装客户端，否则客户端与服务端版本将不一致。

前提条件

- 安装目录可以不存在，会自动创建。但如果存在，则必须为空。目录路径不能包含空格。
- 客户端节点为集群外部服务器时，必须能够与集群业务平面网络互通，否则安装会失败。
- 客户端必须启用NTP服务，并保持与服务端时间一致，否则安装会失败。
- 对于下载所有组件客户端的情况，HDFS与Mapreduce是合一目录（“*客户端目录*/HDFS/”）。
- 安装和使用客户端可以使用任意用户进行操作，用户名和密码请从管理员处获取，本章节以“user_client”进行举例。要求“user_client”用户为服务器文件目录（如“/opt/Bigdata/client”）和客户端安装目录（如“/opt/Bigdata/hadoopclient”）的“owner”，两个目录的权限为“755”。
- 使用客户端需要已从管理员处获取“组件业务用户”（默认用户或新增用户）和“密码”。
- 使用**omm**和**root**以外的用户安装客户端时，若“/var/tmp/patch”目录已存在，需将此目录权限修改为“777”，将此目录内的日志权限修改为“666”。

操作步骤

步骤1 获取软件包。

登录FusionInsight Manager，具体请参考**访问FusionInsight Manager（MRS 3.x及之后版本）**，在“集群”下拉列表中单击需要操作的集群名称。

选择“更多 > 下载客户端”，弹出“下载集群客户端”信息提示框。

说明

在只安装单个服务的客户端的场景中，选择“集群 > *待操作集群的名称* > 服务 > *服务名称* > 更多 > 下载客户端”，弹出“下载客户端”信息提示框。

步骤2 “选择客户端类型”中选择“完整客户端”。

“仅配置文件”下载的客户端配置文件，适用于应用开发任务中，完整客户端已下载并安装后，管理员通过Manager界面修改了服务端配置，开发人员需要更新客户端配置文件的场景。

平台类型包括x86_64和aarch64两种：

- x86_64：可以部署在X86平台的客户端软件包。
- aarch64：可以部署在TaiShan服务器的客户端软件包。

说明

集群支持下载x86_64和aarch64两种类型客户端，但是客户端类型必须待安装节点的架构匹配，否则客户端会安装失败。

步骤3 是否在集群的节点中生成客户端文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。单击“确定”，等待下载完成后，使用**omm**用户或**root**用户将获取的软件包复制到将要安装客户端的服务器文件目录，例如“/opt/Bigdata/client”。然后执行**步骤5**。

  说明

  当用户无法获取**root**用户权限，需要用**omm**用户操作。

- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，执行**步骤4**。

步骤4 上传软件包。

使用WinSCP工具，以准备安装客户端的用户（如“user_client”），将获取的软件包上传到将要安装客户端的服务器文件目录，例如“/opt/Bigdata/client”。

客户端软件包名称格式为：“FusionInsight_Cluster_*<集群ID>*_Services_Client.tar”。

后续步骤及章节以FusionInsight_Cluster_1_Services_Client.tar进行举例。

说明

客户端所在主机可以是集群内节点，也可以是集群外节点。当该节点为集群外部服务器时，必须能够与集群网络互通，并启用NTP服务以保持与服务端时间一致。

例如可以为外部服务器配置与集群一样的NTP时钟源，配置之后可以执行**ntpq -np**命令检查时间是否同步。

- 如果显示结果的NTP时钟源IP地址前有“*”号，表示同步正常，如下：

```
remote refid st t when poll reach delay offset jitter
==============================================================================
*10.10.10.162 .LOCL. 1 u 1 16 377 0.270 -1.562 0.014
```

- 如果显示结果的NTP时钟源IP前无“*”号，且“refid”项内容为“.INIT.”，或者回显异常，表示同步不正常，请联系技术支持。

```
remote refid st t when poll reach delay offset jitter
==============================================================================
10.10.10.162 .INIT. 1 u 1 16 377 0.270 -1.562 0.014
```

也可以为外部服务器配置与集群一样的chrony时钟源，配置之后可以执行**chronyc sources**命令检查时间是否同步。

- 如果显示结果的主OMS节点chrony服务IP地址前有“*”号，表示同步正常，如下：

```
MS Name/IP address         Stratum Poll Reach LastRx Last sample
===============================================================================
^* 10.10.10.162              10  10   377   626    +16us[  +15us] +/-  308us
```

- 如果显示结果的主OMS节点NTP服务IP前无“*”号，且“**Reach**”项内容为“0”，表示同步不正常。

```
MS Name/IP address         Stratum Poll Reach LastRx Last sample
===============================================================================
^? 10.1.1.1                   0  10    0     -    +0ns[   +0ns] +/-    0ns
```

步骤5 以**user_client**用户登录将要安装客户端的服务器。

步骤6 解压软件包。

进入安装包所在目录，例如“/opt/Bigdata/client”。执行如下命令解压安装包到本地目录。

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤7 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig.tar: OK
```

步骤8 解压获取的安装文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

步骤9 配置客户端网络连接。

1. 确保客户端所在主机能与解压目录下“hosts”文件（例如“/opt/Bigdata/client/FusionInsight_Cluster_*<集群ID>*_Services_ClientConfig/hosts”）中所列出的各主机在网络上互通。
2. 当客户端所在主机不是集群中的节点时，需要在客户端所在节点的“/etc/hosts”文件（更改此文件需要**root**用户权限）中设置集群所有节点主机名和业务平面IP地址映射，主机名和IP地址请保持一一对应，可执行以下步骤在hosts文件中导入集群的域名映射关系。
   a. 切换至**root**用户或者其他具有修改hosts文件权限的用户。

      su - root

   b. 进入客户端解压目录。

      cd /opt/Bigdata/client/FusionInsight_Cluster_1_Services_ClientConfig

   c. 执行**cat realm.ini >> /etc/hosts**，将域名映射关系导入到hosts文件中。

说明

- 当客户端所在主机不是集群中的节点时，配置客户端网络连接，可避免执行客户端命令时出现错误。
- 如果采用yarn-client模式运行Spark任务，请在“*客户端安装目录*/Spark/spark/conf/spark-defaults.conf”文件中添加参数“spark.driver.host”，并将参数值设置为客户端的IP地址。
- 当采用yarn-client模式时，为了Spark WebUI能够正常显示，需要在Yarn的主备节点（即集群中的ResourceManager节点）的hosts文件中，配置客户端的IP地址及主机名对应关系。

步骤10 进入安装包所在目录，执行如下命令安装客户端到指定目录（绝对路径），例如安装到“/opt/hadoopclient”目录。

cd /opt/Bigdata/client/FusionInsight_Cluster_1_Services_ClientConfig

执行**./install.sh /opt/hadoopclient**命令，等待客户端安装完成（以下只显示部分屏显结果）。

```
The component client is installed successfully
```

说明

- 如果已经安装的全部服务或某个服务的客户端使用了“/opt/hadoopclient”目录，再安装其他服务的客户端时，需要使用不同的目录。
- 卸载客户端请删除客户端安装目录。
- 如果要求安装后的客户端仅能被该安装用户（如“user_client”）使用，请在安装时加“-o”参数，即执行**./install.sh /opt/hadoopclient -o**命令安装客户端。
- 如果安装NTP服务器为chrony模式，请在安装时加“chrony”参数，即执行**./install.sh /opt/hadoopclient -o chrony**命令安装客户端。
- 由于HBase使用的Ruby语法限制，如果安装的客户端中包含了HBase客户端，建议客户端安装目录路径只包含大写字母、小写字母、数字以及_-?.@+=字符。
- 客户端节点为集群外部服务器且此节点无法与主oms节点的业务平面IP互通时或者无法访问主节点的20029端口时，客户端可以正常安装成功，但无法注册到集群中，无法在界面上进行展示。

步骤11 检查客户端是否安装成功，请登录客户端。

1. 执行**cd /opt/hadoopclient**命令进入客户端安装目录。
2. 执行**source bigdata_env**命令配置客户端环境变量。
3. 如果集群为安全模式，执行以下命令，设置**kinit**认证，输入客户端用户登录密码；普通模式集群无需执行用户认证。

   kinit admin

   ```
   Password for admin@HADOOP.COM: #输入admin用户登录密码（与登录集群的用户密码一致）
   ```

4. 输入**klist**命令查询并确认权限内容。

   ```
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: admin@HADOOP.COM

   Valid starting     Expires            Service principal
   04/09/2021 18:22:35  04/10/2021 18:22:29  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   说明

   - 使用kinit认证时，票据默认会存放到“/tmp/krb5cc_*uid*”目录中。

     *uid*表示当前登录操作系统的用户id，例如root用户的uid为0，那么root用户登录系统后使用kinit认证的票据会默认存放在“/tmp/krb5cc_0”。

     若当前用户对于“/tmp”目录没有读写权限，则会将票据缓存路径修改为“客户端安装目录/tmp/krb5cc_uid”，例如客户端安装目录为“/opt/hadoopclient”，则kinit认证的票据会存放在“/opt/hadoopclient/tmp/krb5cc_uid”。
   - 使用kinit认证时，如果使用相同的用户登录操作系统，则存在票据相互覆盖的风险。可使用**-c** *cache_name*参数指定票据缓存位置，或者通过设置KRB5CCNAME环境变量避免该问题。

步骤12 集群重装后，之前安装的客户端将不再可用，需要重新部署客户端。

1. 以**root**用户登录客户端所在节点。
2. 使用以下命令查看客户端所在目录（下例中“/opt/hadoopclient”为客户端所在目录）。

   ll /opt

   ```
   drwxr-x---. 6 root root       4096 Dec 11 19:00 hadoopclient
   drwxr-xr-x. 3 root root       4096 Dec  9 02:04 godi
   drwx------. 2 root root      16384 Nov  6 01:03 lost+found
   drwxr-xr-x. 2 root root       4096 Nov  7 09:49 rh
   ```

3. 使用**mv**命令移除所有客户端程序所在文件夹内的文件（例如移除“/opt/hadoopclient”文件夹）。

   mv /opt/hadoopclient */tmp/clientbackup*

4. 重新安装客户端。

----结束

10.9.1.2 使用客户端

操作场景

客户端安装后，用户可以通过客户端在运维场景或业务场景中使用shell命令，也可以在应用程序开发场景中使用客户端中的样例工程。

该任务指导用户在运维场景或业务场景中使用客户端。

前提条件

- 已安装客户端。
  例如安装目录为“/opt/Bigdata/client”。
- 各组件业务用户由系统管理员根据业务需要创建。
  “机机”用户需要下载keytab文件，“人机”用户第一次登录时需修改密码。

操作步骤

步骤1 以客户端安装用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/Bigdata/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 根据实际业务需要，执行shell命令。

----结束

10.9.1.3 更新已安装客户端的配置

操作场景

集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户如果在Manager修改了服务配置参数并重启了服务，已安装的客户端需要重新下载并安装，或者使用配置文件更新客户端。

前提条件

已安装客户端。

操作步骤

方法一：

步骤1 登录FusionInsight Manager，在“集群”下拉列表中单击需要操作的集群名称。

步骤2 选择“更多 > 下载客户端 > 仅配置文件”。

此时生成的压缩文件包含所有服务的配置文件。

步骤3 是否在集群的节点中生成配置文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。然后执行**步骤4**。
- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，然后执行**步骤4**。

步骤4 使用WinSCP工具，以客户端安装用户将压缩文件保存到客户端安装的目录，例如“/opt/hadoopclient”。

步骤5 解压软件包。

例如下载的客户端文件为“FusionInsight_Cluster_1_Services_Client.tar”执行如下命令进入客户端所在目录，解压文件到本地目录。

cd /opt/hadoopclient

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤6 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

sha256sum -c
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar: OK
```

步骤7 解压获取配置文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar

步骤8 在客户端安装目录下执行如下命令，使用配置文件更新客户端。

sh refreshConfig.sh *客户端安装目录 配置文件所在目录*

例如，执行以下命令：

sh refreshConfig.sh /opt/hadoopclient /opt/hadoopclient/
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 以**root**用户登录客户端安装节点。

步骤2 进入客户端安装的目录，例如“/opt/Bigdata/client”，执行以下命令更新配置文件：

cd /opt/Bigdata/client

sh autoRefreshConfig.sh

步骤3 按照提示输入FusionInsight Manager管理员用户名，密码以及FusionInsight Manager界面浮动IP。

步骤4 输入需要更新配置的组件名，组件名之间使用“,”分隔。如需更新所有组件配置，可直接单击回车键。

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

10.9.2 集群互信管理

10.9.2.1 集群互信概述

功能介绍

默认情况下，安全模式下的大数据集群用户只能访问本集群中的资源，无法在其他安全模式集群中进行身份认证并访问资源。

特性描述

- **域**

  每个系统用户安全使用的范围定义为“域”，不同的Manager系统需要定义唯一的域名。跨Manager访问实际上就是用户跨域使用。

- **用户加密**

  配置跨Manager互信，当前Kerberos服务端仅支持并使用“aes256-cts-hmac-sha1-96:normal”和“aes128-cts-hmac-sha1-96:normal”加密类型加密跨域使用的用户，不支持修改。

- **用户认证**

  配置跨Manager集群互信后，两个系统中只要存在同名用户，且对端系统的同名用户拥有访问自身系统中某个资源的对应权限，则可以使用当前系统用户访问远程资源。

- **直接互信**

  系统在配置互信的两个集群分别保存对端系统的互信票据，通过互信票据访问对端系统。

10.9.2.2 修改 Manager 系统域名

操作场景

每个系统用户安全使用的范围定义为“域”，不同的系统需要定义唯一的域名。FusionInsight Manager的域名在安装过程中生成，如果需要修改为特定域名，管理员可通过FusionInsight Manager进行配置。

须知

- 修改系统域名为高危操作，在执行本章节操作前，请确认已参考**备份OMS数据**章节成功备份了OMS数据。

对系统的影响

- 修改Manager系统域名时，需要重启所有集群，集群在重启期间无法使用。
- 修改域名后，Kerberos管理员与OMS Kerberos管理员的密码将重新初始化，请使用默认密码并重新修改。组件运行用户的密码是系统随机生成的，如果用于身份认证，请参见**导出认证凭据文件**，重新下载keytab文件。
- 修改域名后，“admin”用户、组件运行用户和系统管理员在修改域名以前添加的“人机”用户，密码会重置为相同密码，请重新修改。重置后的密码由两部分组成：系统生成部分和用户设置部分，系统生成部分为Admin@123，用户设置部分规则参照**表10-53**中“密码后缀”参数的说明，默认值为Admin@123。例如：系统生成部分为Admin@123，用户设置部分为Test#$%@123，则此时重置后的密码为Admin@123Test#$%@123。
- 重置后的密码必需满足当前用户密码策略，使用**omm**用户登录主OMS节点后，执行如下工具脚本可以获取到修改域名后的“人机”用户密码。

  sh ${BIGDATA_HOME}/om-server/om/sbin/get_reset_pwd.sh *密码后缀 user_name*

  - *密码后缀*为用户设置的参数，默认值为“Admin@123”。
  - *user_name*为可选参数，默认取值为“admin”。

  例如：

  sh ${BIGDATA_HOME}/om-server/om/sbin/get_reset_pwd.sh Test#$%@123

  ```
  To get the reset password after changing cluster domain name.
    pwd_min_len     : 8
    pwd_char_types  : 4
  The password reset after changing cluster domain name is: "Admin@123Test#$%@123"
  ```

  “pwd_min_len”和“pwd_char_types”分别表示当前用户密码策略“最小密码长度”和“密码字符类型数目”，“Admin@123Test#$%@123”为修改系统域名后的“人机”用户密码。
- 修改系统域名后，重置后的密码由系统生成部分和用户设置部分组成，且必需满足当前用户密码策略，长度不足时在Admin@123和用户设置部分中间，使用一个或多个@补全；字符种类为5时，在Admin@123后补充一个空格。

  当用户设置部分为Test@123，使用默认用户密码策略时，新密码为“Admin@123Test@123”，长度为17字符种类为4。需满足当前用户密码策略时，新密码处理如**表10-52**所示。

  表 10-52 满足不同密码策略时的新密码

最小密码长度	字符种类	对比用户密码策略结果	重置后的密码
8到17位	4	已满足用户密码策略	Admin@123Test@123

最小密码长度	字符种类	对比用户密码策略结果	重置后的密码
18位	4	需补充一个@	Admin@123@Test@123
19位	4	需补充两个@	Admin@123@@Test@123
8到18位	5	需补充一个空格	Admin@123 Test@123
19位	5	需补充一个空格和一个@	Admin@123 @Test@123
20位	5	需补充一个空格和两个@	Admin@123 @@Test@123

- 修改系统域名后，系统管理员在修改域名以前添加的“机机”用户，请重新下载keytab文件。
- 修改系统域名后，请重新下载并安装集群客户端。

前提条件

- 管理员已明确业务需求，并规划好不同系统的域名。
  域名只能包含大写字母、数字、圆点（.）及下划线（_），且只能以字母或数字开头。
- Manager内所有集群全部组件的运行状态均为“良好”。
- Manager内所有集群的ZooKeeper服务的“acl.compare.shortName”参数需确保为默认值“true”。否则请修改该参数为“true”后重启ZooKeeper服务。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 域和互信”。

步骤3 修改相关参数。

表 10-53 相关参数

参数名	描述
本端域	填写本系统规划好的域名。

参数名	描述
密码后缀	“人机”用户密码重置后的用户设置密码部分，默认值为Admin@123。 **说明** 该参数只有在修改了“本端域”参数后，才会生效。且需满足以下条件： • 密码字符长度为8到16位。 • 至少需要包含大写字母、小写字母、数字、特殊字符中的三种类型字符。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];:',<.>/?和空格。

步骤4 单击“确定”，等待修改配置完成后再继续执行后续步骤，完成前请勿提前执行后续步骤。

步骤5 以**omm**用户登录主管理节点。

步骤6 执行以下命令，重启更新域配置。

sh ${BIGDATA_HOME}/om-server/om/sbin/restart-RealmConfig.sh

提示以下信息表示命令执行成功。

```
Modify realm successfully. Use the new password to log into FusionInsight again.
```

说明

重启后部分主机与服务可能无法访问并触发告警，执行“restart-RealmConfig.sh”后大约需要1分钟自动恢复。

步骤7 使用重置后的admin用户及密码（例如Admin@123Admin@123）登录FusionInsight Manager，单击主页上待操作集群名称后的 ，单击“重启”，重启集群。

在弹出窗口中输入当前登录的用户密码确认身份，然后单击“确定”。

在确认重启集群的对话框中单击“确定”，等待界面提示“操作成功。”，单击“完成”。

步骤8 退出FusionInsight Manager，重新登录正常表示配置已成功。

步骤9 使用**omm**用户登录主管理节点，执行以下命令刷新作业提交客户端配置：

sh /opt/executor/bin/refresh-client-config.sh

----结束

10.9.2.3 配置跨 Manager 集群互信

操作场景

当不同的两个Manager系统下安全模式的集群需要互相访问对方的资源时，管理员可以设置互信的系统，使外部系统的用户可以在本系统中使用。

每个系统用户安全使用的范围定义为“域”，不同的Manager系统需要定义唯一的域名。跨Manager访问实际上就是用户跨域使用。

说明

最多支持配置500个互信集群。

对系统的影响

- 配置跨集群互信后，外部系统的用户可以在本系统中使用，请管理员根据企业业务与安全要求，定期检视Manager系统中用户的权限。
- 配置跨集群互信时需要停止所有集群，会造成业务中断。
- 配置跨集群互信后，互信的集群中均会增加Kerberos内部用户“krbtgt/*本集群域名@外部集群域名*”、“krbtgt/*外部集群域名@本集群域名*”，用户不能删除。请管理员根据企业安全要求，及时且定期修改密码，需同时修改互信系统中4个用户且密码保持一致。具体请参见**修改组件运行用户密码**。修改密码期间可能影响跨系统业务应用的连接。
- 配置跨集群互信后，各个集群都需要重新下载并安装客户端。
- 配置跨集群互信后，验证配置后是否可以正常工作，且如何使用本系统用户访问对端系统资源，请参见**配置跨集群互信后的用户权限**。

前提条件

- 管理员已明确业务需求，并规划好不同系统的域名。域名只能包含大写字母、数字、圆点（.）及下划线（_），且只能以字母或数字开头。
- 配置跨集群互信前，两个Manager系统的域名必须不同。MRS创建ECS/BMS集群时会随机生成唯一系统域名，通常无需修改。
- 配置跨集群互信前，两个集群中不能存在有相同的主机名，也不能存在相同的IP地址。
- 配置互信的两个集群系统时间必须一致，且系统上的NTP服务必须使用同一个时间源。
- 配置互信的两个集群系统内所有集群全部组件的运行状态均为“良好”。
- Manager内所有集群的ZooKeeper服务的“acl.compare.shortName”参数需确保为默认值“true”。否则请修改该参数为“true”后重启ZooKeeper服务。

操作步骤

步骤1 登录其中一个FusionInsight Manager。

步骤2 在主页中停止所有集群。

单击主页上待操作集群名称后的 ，单击“停止”，输入管理员密码后在弹出的“停止集群”窗口中单击“确定”，等待集群停止成功。

步骤3 选择“系统 > 权限 > 域和互信”。

步骤4 修改配置参数“互信对端域”。

表 10-54 相关参数

参数名	描述
“realm_name”	填写对端系统的域名。

参数名	描述
“ip_port”	填写对端系统的KDC地址。 *参数值格式为：对端系统内要配置互信集群的Kerberos服务部署的节点IP地址:端口。* ● 如果是双平面组网，需填写业务平面IP地址。 ● 采用IPv6地址时，IP地址应写在中括号“[]”中。 ● 部署主备Kerberos服务或者对端系统内有多个集群需要与本端建立互信时，多个KDC地址使用逗号分隔。 ● 端口值可通过查看KrbServer服务的“kdc_ports”参数获取，默认值为“21732”。部署服务的节点IP可通过在KrbServer服务页面选择“实例”页签，查看KerberosServer角色的“业务IP”获取。 例如，Kerberos服务部署在10.0.0.1和10.0.0.2上，与本端系统建立互信，则对应参数值为“10.0.0.1:21732,10.0.0.2:21732”。

说明

如果需要配置与多个Manager系统的互信关系，请单击 添加新项目，并填写参数值。最多支持16个系统。删除多余的配置请单击 。

步骤5 单击“确定”。

步骤6 以**omm**用户登录主管理节点，执行以下命令更新域配置。

sh ${BIGDATA_HOME}/om-server/om/sbin/restart-RealmConfig.sh

提示以下信息表示命令执行成功。

```
Modify realm successfully. Use the new password to log into FusionInsight again.
```

重启后部分主机与服务可能无法访问并触发告警，执行“restart-RealmConfig.sh”后大约需要1分钟自动恢复。

步骤7 登录FusionInsight Manager，启动所有集群。

单击主页上待操作集群名称后的 ，单击“启动”，在“启动集群”窗口单击“确定”，等待集群启动成功。

步骤8 登录另外一个系统的FusionInsight Manager，重复以上操作。

----结束

10.9.2.4 配置跨集群互信后的用户权限

操作场景

配置完跨Manager集群互信后，需要在互信的系统上设置访问用户的权限，这样指定的用户才能在互信系统上进行对应的业务操作。

前提条件

两个系统已完成互信配置。

操作步骤

步骤1 登录本端系统的FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”，检查本次业务操作的用户是否在已存在：

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 单击指定用户左侧的，检查该用户所在的用户组和角色分配的权限是否满足本次业务需求。若不满足，参见**权限设置**创建新角色并绑定用户，也可以直接修改用户的用户组或角色权限。

步骤4 参见**创建用户**，创建本次业务所需要的用户，同时关联业务所需要的用户组或者角色信息。

步骤5 登录互信系统的FusionInsight Manager，重复**步骤2**～**步骤4**，创建相同名字的用户并设置权限。

----结束

10.9.3 配置定时备份告警与审计信息

操作场景

管理员可通过修改配置文件，实现定时备份FusionInsight Manager的告警信息、Manager审计信息以及所有服务的审计信息到指定的存储位置。

备份支持使用SFTP协议或FTP协议，FTP协议未加密数据可能存在安全风险，建议使用SFTP。

操作步骤

步骤1 以**omm**用户登录主管理节点。

说明

用户只需在主管理节点执行此操作，不支持在备管理节点上配置定时备份。

步骤2 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/sbin

步骤3 执行以下命令，配置定时备份Manager告警、审计或者服务审计信息。

./setNorthBound.sh -t *信息类型* **-i** *远程服务器IP* **-p** *服务器使用的SFTP或FTP端口* **-u** *用户名* **-d** *保存信息的路径* **-c** *时间间隔（分钟）* **-m** *每个保存文件的信息记录数* **-s** *备份启停开关* **-e** *指定的协议*

例如：

./setNorthBound.sh -t alarm -i 10.0.0.10 -p 22 -u sftpuser -d /tmp/ -c 10 -m 100 -s true -e sftp

此脚本将修改告警信息备份配置文件“alarm_collect_upload.properties”。文件存储路径为“${BIGDATA_HOME}/om-server/tomcat/webapps/web/WEB-INF/classes/config”。

./setNorthBound.sh -t audit -i 10.0.0.10 -p 22 -u sftpuser -d /tmp/ -c 10 -m 100 -s true -e sftp

此脚本将修改审计信息备份配置文件“audit_collect_upload.properties”。文件存储路径为“${BIGDATA_HOME}/om-server/tomcat/webapps/web/WEB-INF/classes/config”。

./setNorthBound.sh -t service_audit -i 10.0.0.10 -p 22 -u sftpuser -d /tmp/ -c 10 -m 100 -s true -e sftp

此脚本将修改服务审计信息备份配置文件“service_audit_collect_upload.properties”。文件存储路径为“${BIGDATA_HOME}/om-server/tomcat/webapps/web/WEB-INF/classes/config”。

步骤4 根据界面提示输入用户的密码。密码将加密保存在配置文件中。

```
Please input sftp/ftp server password:
```

步骤5 显示如下结果，说明修改成功。备管理节点将自动同步配置文件。

```
execute command syncfile successfully.
Config Succeed.
```

----结束

10.9.4 修改 FusionInsight Manager 添加的路由表

操作场景

安装FusionInsight Manager时系统会自动在主管理节点上创建2条路由信息，执行**ip rule list**可以查看，如下示例：

```
0:from all lookup local
32764:from all to 10.10.100.100 lookup ntp_rt  #FusionInsight Manager创建的ntp路由信息（未配置外部NTP时钟源时无此信息）
32765:from 192.168.0.117 lookup om_rt  #FusionInsight Manager创建的om路由信息
32766:from all lookup main
32767:from all lookup default
```

说明

没有配置ntp外部服务器时只会有一条om路由信息“om_rt”。

如果FusionInsight Manager创建的路由信息与企业网络规划配置的路由信息发生冲突时，管理员可以使用“autoroute.sh”工具禁用或启用Manager创建的路由信息。

对系统的影响

禁用Manager创建的路由信息后，在设置新的路由信息之前，FusionInsight Manager页面无法登录，集群运行不受影响。

前提条件

已经成功安装Manager。

已获取待创建的WS浮动IP路由的相关信息。

禁用系统创建的路由信息

步骤1 以**omm**用户登录到主管理节点。执行以下命令，禁用系统创建的路由信息。

cd ${BIGDATA_HOME}/om-server/om/sbin

./autoroute.sh disable

```
Deactivating Route.
Route operation (disable) successful.
```

步骤2 执行以下命令，查看运行结果。如下例

ip rule list

```
0:from all lookup local
32766:from all lookup main
32767:from all lookup default
```

步骤3 执行以下命令，输入**root**用户密码，切换到**root**用户下。

su - root

步骤4 分别执行以下命令，手动创建新的WS浮动IP路由信息。

ip route add *WS浮动IP网段号/WS浮动IP子网掩码* **scope link src** *WS浮动IP* **dev** *WS浮动IP对应网卡* **table om_rt**

ip route add default via *WS浮动IP网关* **dev** *WS浮动IP对应网卡* **table om_rt**

ip rule add from *WS浮动IP* **table om_rt**

例如：

ip route add 192.168.0.0/255.255.255.0 scope link src 192.168.0.117 dev eth0:ws table om_rt

ip route add default via 192.168.0.254 dev eth0:ws table om_rt

ip rule add from 192.168.0.117 table om_rt

说明

当前网络的IP地址模式为IPv6时，应执行**ip -6 route add**命令。

步骤5 分别执行以下命令，手动创建新的ntp服务路由信息。未配置外部NTP时钟源时，跳过此步骤。

ip route add default via *NtpIP网关* **dev** *本机IP对应网卡* **table ntp_rt**

ip rule add to *ntpIP* **table ntp_rt**

*本机IP对应网卡*是指可与NTP服务器所在网段互通的网卡。

例如：

ip route add default via 10.10.100.254 dev eth0 table ntp_rt

ip rule add to 10.10.100.100 table ntp_rt

步骤6 执行以下命令，查看运行结果。

如下例，如产生路由表名为“om_rt”和“ntp_rt”的路由信息，则操作成功。

ip rule list

```
0:from all lookup local
32764:from all to 10.10.100.100 lookup ntp_rt  #未配置外部NTP时钟源时无此信息
32765:from 192.168.0.117 lookup om_rt
32766:from all lookup main
32767:from all lookup default
```

----结束

启用系统创建的路由信息

步骤1 以**omm**用户登录到主管理节点。

步骤2 执行以下命令，启用系统创建的路由信息。

cd ${BIGDATA_HOME}/om-server/om/sbin

./autoroute.sh enable

```
Activating Route.
Route operation (enable) successful.
```

步骤3 执行以下命令，查看运行结果。

如下例，如产生路由表名为“ntp_rt”和“om_rt”的两条路由信息，则操作成功。

ip rule list

```
0:from all lookup local
32764:from all to 10.10.100.100 lookup ntp_rt  #未配置外部NTP时钟源时无此信息
32765:from 192.168.0.117 lookup om_rt
32766:from all lookup main
32767:from all lookup default
```

----结束

10.9.5 切换维护模式

操作场景

FusionInsight Manager支持将集群、服务、主机或者OMS配置为维护模式，进入维护模式的对象将不再上报告警，避免在升级等维护变更期间系统产生大量无意义的告警，影响运维人员对集群状态的判断。

- 集群维护模式

  集群未正式上线或暂时离线进行运维操作时（例如非滚动方式的升级），可将整个集群配置为维护模式。

- 服务维护模式

  对特定服务进行维护操作时（例如对该服务的实例进行批量重启等可能影响业务的调试操作、对该服务相关的节点进行直接上下电或修复服务等），可仅将涉及的服务配置为维护模式。

- 主机维护模式

  对主机进行维护操作时（例如节点上下电、隔离主机、重装主机、升级操作系统、替换节点等），可仅将涉及的主机配置为维护模式。

- OMS维护模式

  对OMS节点进行重启、替换、修复等操作时，可将OMS配置为维护模式。

对系统影响

设置维护模式后，非维护操作引起的告警也将被抑制无法上报，直至退出维护模式后，仍然存在的故障才能上报告警，请谨慎操作。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 配置维护模式。

根据实际操作场景，确认需要配置维护模式的对象，参考**表10-55**进行操作。

表 10-55 切换维护模式

场景	步骤
配置集群进入维护模式	1. 在管理界面主页，选择待操作集群名称后的“ ⋮ > 进入维护模式”。 2. 在弹出的窗口中单击“确定”。 集群进入维护状态后，集群名称后的状态显示为 。维护操作完成后，单击“退出维护模式”，集群将退出维护模式。
配置服务进入维护模式	1. 在管理界面选择“集群 > *待操作的集群名称* > 服务 > *服务名称*”。 2. 在服务详情页面选择“更多 > 进入维护模式”。 3. 在弹出的窗口中单击“确定”。 服务进入维护状态后，服务列表的对应服务名称后的状态显示为 。维护操作完成后，单击“退出维护模式”，服务将退出维护模式。 **说明** 配置某服务进入维护模式时，建议将依赖该服务的其他上层服务也都设置为维护模式。
配置主机进入维护模式	1. 在管理界面单击“主机”。 2. 在主机页面勾选待操作的主机，选择“更多 > 进入维护模式”。 3. 在弹出的窗口中单击“确定”。 主机进入维护状态后，主机列表的对应主机名称后的状态显示为 。维护操作完成后，单击“退出维护模式”，主机将退出维护模式。
配置OMS进入维护模式	1. 在管理界面选择“系统 > OMS > 进入维护模式”。 2. 在弹出的窗口中单击“确定”。 OMS进入维护状态后，OMS状态显示为 。维护操作完成后，单击“退出维护模式”，OMS将退出维护模式。

步骤3 查看集群维护视图。

在管理界面主页，选择待操作集群名称后的“ ··· > 维护模式视图”，在弹出的窗口中可查看当前集群内处于维护模式的服务及主机，方便查询。

维护操作完成后，可在维护模式视图中批量勾选服务与主机，然后单击“退出维护模式”，即可批量退出维护模式。

图 10-19 批量退出维护模式

----结束

10.9.6 例行维护

为了保证系统长期正常、稳定的运行，管理员或维护工程师需要定期对表10-56所示的项目进行检查，并根据检查出的异常结果排除故障。建议检查人员根据企业管理规范，记录每个任务场景的结果并签名确认。

表 10-56 项目一览表

例行维护周期	任务场景	例行维护内容
每天	检查集群服务状态	• 检查每个服务的运行状态和配置状态是否正常，是否为绿色。 • 检查每个服务中，角色实例的运行状态和配置状态是否正常，是否为绿色。 • 检查每个服务中，角色实例的主备状态是否可以正常显示。 • 检查服务与角色实例的“概览”显示结果是否正常。
	检查集群主机状态	• 检查每个主机当前的运行状态是否正常，是否为绿色。 • 检查每个主机当前的磁盘使用率、内存使用率和CPU使用率。关注当前内存与CPU使用率是否处于上升趋势。
	检查集群告警信息	检查前一天是否生成了未处理异常告警，包含已自动恢复的告警。
	检查集群审计信息	检查前一天是否有“高危”和“危险”的操作，是否已确认操作的合法性。
	检查集群备份情况	检查前一天OMS、LDAP、DBService和NameNodeOMS、LDAP和DBServiceOMS、LDAP是否自动备份。
	检查健康检查结果	在FusionInsight Manager执行健康检查，下载健康检查报告确认当前集群是否存在异常状态。建议启用自动健康检查，并及时导出最新的集群健康检查结果，根据检查结果修复不健康项目。
	检查网络通讯	检查集群网络工作状态，节点之间的网络通讯是否存在延时。
	检查存储状态	检查集群总体数据存储量是否出现了突然的增长： • 磁盘使用量是否已接近阈值，排查原因，例如是否有业务留下的垃圾数据或冷数据需要清理。 • 业务是否有增长需求，磁盘分区是否需要扩容。
	检查日志情况	• 检查是否有失败、失去响应的MapReduce、Spark任务，查看HDFS中“/tmp/logs/${username}/logs/${application id}”日志文件并排除问题。 • 检查Yarn的任务日志，查看失败、失去响应的任务日志，并删除重复数据。 • 检查Storm的worker日志。 • 备份日志到存储服务器。
每周	用户管理	检查用户密码是否将要过期并通知修改。“机机用户”用户修改密码需要重新下载keytab文件。
	分析告警	导出指定周期内产生的告警并分析。

例行维护周期	任务场景	例行维护内容
	扫描磁盘	对磁盘健康状态进行检查，建议使用专门的磁盘检查工具。
	统计存储	分批次排查集群节点磁盘数据是否均匀存储，筛选出明显数据增加或不足的硬盘，并确认硬盘是否正常。
	记录变更	安排并记录对集群配置参数和文件实施的操作，为故障分析处理场景提供依据。
每月	分析日志	• 收集集群节点服务器的硬件日志，例如BMC系统日志，并进行分析。 • 收集集群节点服务器的操作系统日志，并进行分析。 • 收集集群日志，并进行分析。
	诊断网络	对集群的网络健康状态进行分析。
	管理硬件	检查设备运行的机房环境，安排清洁设备。

10.10 日志管理

10.10.1 关于日志

日志描述

MRS集群的日志保存路径为“/var/log/Bigdata”。日志分类见下表：

表 10-57 日志分类一览表

日志类型	日志描述
安装日志	安装日志记录了Manager、集群和服务安装的程序信息，可用于定位安装出错的问题。
运行日志	运行日志记录了集群各服务运行产生的运行轨迹信息及调试信息、状态变迁、未产生影响的潜在问题和直接的错误信息。
审计日志	审计日志中记录了用户活动信息和用户操作指令信息，可用于安全事件中定位问题原因及划分事故责任。

MRS日志目录清单见下表：

表 10-58 日志目录一览表

文件目录	日志内容
/var/log/Bigdata/audit	组件审计日志。
/var/log/Bigdata/controller	日志采集脚本日志。 controller进程日志。 controller监控日志。
/var/log/Bigdata/dbservice	DBService日志。
/var/log/Bigdata/flume	Flume日志。
/var/log/Bigdata/hbase	HBase日志。
/var/log/Bigdata/hdfs	HDFS日志。
/var/log/Bigdata/hive	Hive日志。
/var/log/Bigdata/httpd	httpd日志。
/var/log/Bigdata/hue	Hue日志。
/var/log/Bigdata/kerberos	Kerberos日志。
/var/log/Bigdata/ldapclient	LDAP客户端日志。
/var/log/Bigdata/ldapserver	LDAP服务端日志。
/var/log/Bigdata/loader	Loader日志。
/var/log/Bigdata/logman	logman脚本日志管理日志。
/var/log/Bigdata/mapreduce	MapReduce日志。
/var/log/Bigdata/nodeagent	NodeAgent日志。
/var/log/Bigdata/okerberos	OMS Kerberos日志。
/var/log/Bigdata/oldapserver	OMS LDAP日志。
/var/log/Bigdata/ metric_agent	MetricAgent运行日志。
/var/log/Bigdata/omm	oms：“omm”服务端的复杂事件处理日志、告警服务日志、HA日志、认证与授权管理日志和监控服务运行日志。 oma：“omm”代理端的安装运行日志。 core：“omm”代理端与“HA”进程失去响应的dump日志。
/var/log/Bigdata/spark2x	Spark2x日志。
/var/log/Bigdata/sudo	omm执行sudo命令产生的日志。
/var/log/Bigdata/timestamp	时间同步管理日志。
/var/log/Bigdata/tomcat	Tomcat日志。

文件目录	日志内容
/var/log/Bigdata/watchdog	Watchdog日志。
/var/log/Bigdata/yarn	Yarn日志。
/var/log/Bigdata/zookeeper	ZooKeeper日志。
/var/log/Bigdata/oozie	Oozie日志。
/var/log/Bigdata/kafka	Kafka日志。
/var/log/Bigdata/storm	Storm日志。
/var/log/Bigdata/upgrade	升级OMS日志。
/var/log/Bigdata/update-service	升级服务日志。

说明

启用多实例功能后，如果系统管理员添加了多个HBase、Hive和Spark服务的实例，新增加服务实例的日志描述、日志级别和日志格式，与原服务日志相同。服务实例的日志将独立保存在名为“/var/log/Bigdata/servicenameN”的目录中，HBase和Hive服务实例的审计日志保存在名为“/var/log/Bigdata/audit/servicenameN”的目录中。以HBase1为例，对应日志分别保存在“/var/log/Bigdata/hbase1”和“/var/log/Bigdata/audit/hbase1”。

安装日志

表 10-59 安装信息一览表

安装日志	日志描述
安装配置日志	记录了安装前配置过程的信息。
安装Manager日志	记录了安装双机Manager操作的信息。
安装集群日志	记录了安装集群步骤的信息。

运行日志

运行日志记录的运行信息描述如表10-60所示。

表 10-60 运行信息一览表

运行日志	日志描述
服务安装前的准备日志	记录服务安装前的准备工作，如检测、配置和反馈操作的信息。
进程启动日志	记录进程启动过程中执行的命令信息。

运行日志	日志描述
进程启动异常日志	记录进程启动失败时产生异常的信息，如依赖服务错误、资源不足等
进程运行日志	记录进程运行轨迹信息及调试信息，如函数入口和出口打印、模块间接口消息等。
进程运行异常日志	记录导致进程运行时错误的错误信息，如输入对象为空、编解码失败等错误。
进程运行环境信息日志	记录进程运行环境的信息，如资源状态、环境变量等。
脚本日志	记录脚本执行的过程信息。
资源回收日志	记录资源回收的过程信息。
服务卸载时的清理日志	记录卸载服务时执行的步骤操作信息，如清除目录数据、执行时间等

审计日志

审计日志记录的审计信息包含Manager审计信息和组件审计信息。

表 10-61 Manager 审计信息一览表

操作类型	操作
用户管理	创建用户 修改用户 删除用户 创建组 修改组 删除组 添加角色 修改角色 删除角色 密码策略修改 修改密码 密码重置 用户登录 用户注销 屏幕解锁 下载认证凭据 用户越权操作 用户帐号解锁 用户帐号锁定 屏幕锁定 导出用户 导出用户组 导出角色

操作类型	操作
集群	启动集群 停止集群 重启集群 滚动重启集群 重启所有过期实例 保存配置 同步集群配置 定制集群监控指标 配置监控转储 保存监控阈值 下载客户端配置 北向Syslog接口配置 北向SNMP接口配置 SNMP清除告警 SNMP添加trap目标 SNMP删除trap目标 SNMP检查告警 SNMP同步告警 创建阈值模板 删除阈值模板 应用阈值模板 保存集群监控配置数据 导出配置数据 导入集群配置数据 导出安装模板 修改阈值模板 取消阈值模板应用 屏蔽告警 发送告警 修改OMS数据库密码 重置组件数据库密码 重启OMM和Controller 启动集群的健康检查 导入证书文件 配置SSO信息 删除健康检查历史报告 修改集群属性 同步维护命令

操作类型	操作
	异步维护命令 定制报表监控指标 导出报表监控数据 SNMP执行异步命令 重启WEB服务 定制静态资源池监控指标 导出静态资源池监控数据 定制主页监控指标 中止任务 还原配置 修改域和互信的配置 修改系统参数 集群进入维护模式 集群退出维护模式 OMS进入维护模式 OMS退出维护模式 批量退出维护模式 修改OMS配置 启用阈值告警 同步所有集群配置

操作类型	操作
服务	启动服务 停止服务 同步服务配置 刷新服务队列 定制服务监控指标 重启服务 滚动重启服务 导出服务监控数据 导入服务配置数据 启动服务的健康检查 服务配置 上传配置文件 下载配置文件 同步实例配置 实例入服 实例退服 启动实例 停止实例 定制实例监控指标 重启实例 滚动重启实例 导出实例监控数据 导入实例配置数据 创建实例组 修改实例组 删除实例组 移动到另一个实例组 服务进入维护模式 服务退出维护模式 修改服务显示名称 修改服务关联关系 下载监控数据 屏蔽告警 取消屏蔽告警 导出服务的报表数据 添加报表的自定义参数 修改报表的自定义参数 删除报表的自定义参数

操作类型	操作
	倒换控制节点 新增挂载表 修改挂载表
主机	设置节点机架 启动所有角色 停止所有角色 隔离主机 取消隔离主机 定制主机监控指标 导出主机监控数据 主机进入维护模式 主机退出维护模式 导出主机基本信息 导出主机分布的报表数据 导出主机趋势的报表数据 导出主机集群的报表数据 导出服务的报表数据 定制主机集群监控指标 定制主机趋势监控指标
告警	导出告警 清除告警 导出事件 批量清除告警
采集日志	采集日志文件 下载日志文件 采集服务堆栈信息 采集实例堆栈信息 准备服务堆栈信息 准备实例堆栈信息 清理服务堆栈信息 清理实例堆栈信息
审计日志	修改审计转储配置 导出审计日志

操作类型	操作
备份恢复	创建备份任务 执行备份任务 批量执行备份任务 停止备份任务 删除备份任务 修改备份任务 锁定备份任务 解锁备份任务 创建恢复任务 执行恢复任务 停止恢复任务 重试恢复任务 删除恢复任务
多租户	保存静态配置 添加租户 删除租户 关联租户服务 删除租户服务 配置资源 创建资源 删除资源 增加资源池 修改资源池 删除资源池 恢复租户数据 修改租户全局配置 修改容量调度器队列配置 修改超级调度器队列配置 修改容量调度器资源分布 清除容量调度器资源分布 修改超级调度器资源分布 清除超级调度器资源分布 添加资源目录 修改资源目录 删除资源目录 定制租户监控指标

操作类型	操作
健康检查	启动集群的健康检查 启动服务的健康检查 启动主机的健康检查 启动oms健康检查 启动系统的健康检查 更新健康检查的配置 导出健康检查报告 导出集群健康检查的结果 导出服务健康检查的结果 导出主机健康检查的结果 删除健康检查历史报告 导出健康检查历史报告 下载健康检查报告

表 10-62 组件审计信息一览表

审计日志	操作类型	操作
ClickHouse审计日志	维护管理	授权 收回权限 认证和登录信息
	业务操作	创建数据库/表 插入、删除、查询、执行数据迁移任务
DBService审计日志	维护管理	备份恢复操作
HBase审计日志	DDL（数据定义）语句	创建表 删除表 修改表 增加列族 修改列族 删除列族 启用表 禁用表 用户信息修改 修改密码 用户登录

审计日志	操作类型	操作
	DML（数据操作）语句	put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表）
	权限控制	给用户授权 取消用户授权
HDFS审计日志	权限管理	文件/文件夹访问权限 文件/文件夹owner信息
	文件操作	创建文件夹 创建文件 打开文件 追加文件内容 修改文件名称 删除文件/文件夹 设置文件时间属性 设置文件副本个数 多文件合并 文件系统检查 文件链接
Hive审计日志	元数据操作	元数据定义，如创建数据库、表等 元数据删除，如删除数据库、表等 元数据修改，如增加列、重命名表等 元数据导入/导出
	数据维护	向表中加载数据 向表中插入数据
	权限管理	创建/删除角色 授予/回收角色 授予/回收权限
Hue审计日志	服务启动	启动Hue
	用户操作	用户登录 用户退出

审计日志	操作类型	操作
	任务操作	创建任务 修改任务 删除任务 提交任务 保存任务 任务状态更新
KrbServer审计日志	维护管理	修改kerberos帐号密码 添加kerberos帐号 删除kerberos帐号 用户认证
LdapServer审计日志	维护管理	添加操作系统用户 添加组 添加用户到组 删除用户 删除组
Loader审计日志	安全管理	用户登录
	元数据管理	查询connector 查询framework 查询step
	数据源连接管理	查询数据源连接 增加数据源连接 更新数据源连接 删除数据源连接 激活数据源连接 禁用数据源连接
	作业管理	查询作业 创建作业 更新作业 删除作业 激活作业 禁用作业 查询作业所有执行记录 查询作业最近执行记录 提交作业 停止作业

审计日志	操作类型	操作
Mapreduce审计日志	程序运行	启动Container请求 停止Container请求 Container结束，状态为成功 Container结束，状态为失败 Container结束，状态为中止 提交任务 结束任务
Oozie审计日志	任务管理	提交任务 启动任务 kill任务 暂停任务 恢复任务 重新运行任务
Spark2x审计日志	元数据操作	元数据定义，如创建数据库、表等 元数据删除，如删除数据库、表等 元数据修改，如增加列、重命名表等 元数据导入/导出
	数据维护	向表中加载数据 向表中插入数据
Storm审计日志	Nimbus	提交拓扑 中止拓扑 重分配拓扑 去激活拓扑 激活拓扑
	UI	中止拓扑 重分配拓扑 去激活拓扑 激活拓扑
Yarn审计日志	任务提交	提交作业到队列相关的操作
Zookeeper审计日志	权限管理	设置ZNODE访问权限
	ZNODE操作	创建ZNODE 删除ZNODE 设置ZNODE数据

FusionInsight Manager的审计日志保存在数据库中，可通过“审计”页面查看及导出审计日志。

组件审计日志的文件信息见下表。部分组件审计日志文件保存在“/var/log/Bigdata/audit”，例如HDFS、HBase、Mapreduce、Hive、Hue、Yarn、Storm和ZooKeeper。每天凌晨3点自动将组件审计日志压缩备份到“/var/log/Bigdata/audit/bk”，最多保留最近的90个压缩备份文件，不支持修改备份时间。配置保留个数，请参见**配置审计日志本地备份数**。

其他组件审计日志文件保存在组件日志目录中。

表 10-63 组件审计日志目录

组件名称	审计日志目录
DBService	/var/log/Bigdata/audit/dbservice/dbservice_audit.log
HBase	/var/log/Bigdata/audit/hbase/hm/hbase-audit-hmaster.log /var/log/Bigdata/audit/hbase/hm/hbase-ranger-audit-hmaster.log /var/log/Bigdata/audit/hbase/rs/hbase-audit-regionserver.log /var/log/Bigdata/audit/hbase/rs/hbase-ranger-audit-regionserver.log /var/log/Bigdata/audit/hbase/rt/hbase-audit-restserver.log /var/log/Bigdata/audit/hbase/ts/hbase-audit-thriftserver.log
HDFS	/var/log/Bigdata/audit/hdfs/nn/hdfs-audit-namenode.log /var/log/Bigdata/audit/hdfs/nn/ranger-plugin-audit.log /var/log/Bigdata/audit/hdfs/dn/hdfs-audit-datanode.log /var/log/Bigdata/audit/hdfs/jn/hdfs-audit-journalnode.log /var/log/Bigdata/audit/hdfs/zkfc/hdfs-audit-zkfc.log /var/log/Bigdata/audit/hdfs/httpfs/hdfs-audit-httpfs.log /var/log/Bigdata/audit/hdfs/router/hdfs-audit-router.log
Hive	/var/log/Bigdata/audit/hive/hiveserver/hive-audit.log /var/log/Bigdata/audit/hive/hiveserver/hive-rangeraudit.log /var/log/Bigdata/audit/hive/metastore/metastore-audit.log /var/log/Bigdata/audit/hive/webhcat/webhcat-audit.log
Hue	/var/log/Bigdata/audit/hue/hue-audits.log
Kafka	/var/log/Bigdata/audit/kafka/audit.log
Loader	/var/log/Bigdata/loader/audit/default.audit
Mapreduce	/var/log/Bigdata/audit/mapreduce/jobhistory/mapred-audit-jobhistory.log
Oozie	/var/log/Bigdata/audit/oozie/oozie-audit.log

组件名称	审计日志目录
Spark2x	/var/log/Bigdata/audit/spark2x/jdbcserver/jdbcserver-audit.log /var/log/Bigdata/audit/spark2x/jdbcserver/ranger-audit.log /var/log/Bigdata/audit/spark2x/jobhistory/jobhistory-audit.log
Storm	/var/log/Bigdata/audit/storm/logviewer/audit.log /var/log/Bigdata/audit/storm/nimbus/audit.log /var/log/Bigdata/audit/storm/supervisor/audit.log /var/log/Bigdata/audit/storm/ui/audit.log
Yarn	/var/log/Bigdata/audit/yarn/rm/yarn-audit-resourcemanager.log /var/log/Bigdata/audit/yarn/rm/ranger-plugin-audit.log /var/log/Bigdata/audit/yarn/nm/yarn-audit-nodemanager.log
ZooKeeper	/var/log/Bigdata/audit/zookeeper/quorumpeer/zk-audit-quorumpeer.log

10.10.2 Manager 日志清单

日志描述

日志存储路径：Manager相关日志的默认存储路径为“/var/log/Bigdata/*Manager组件*”。

- ControllerService：/var/log/Bigdata/controller/（OMS安装、运行日志）
- Httpd：/var/log/Bigdata/httpd（httpd安装、运行日志）
- logman：/var/log/Bigdata/logman（日志打包工具日志）
- NodeAgent：/var/log/Bigdata/nodeagent（NodeAgent安装、运行日志）
- okerberos：/var/log/Bigdata/okerberos（okerberos安装、运行日志）
- oldapserver：/var/log/Bigdata/oldapserver（oldapserver安装、运行日志）
- MetricAgent：/var/log/Bigdata/metric_agent（MetricAgent运行日志）
- omm：/var/log/Bigdata/omm（omm安装、运行日志）
- timestamp：/var/log/Bigdata/timestamp（NodeAgent启动时间日志）
- tomcat：/var/log/Bigdata/tomcat（Web进程日志）
- watchdog：/var/log/Bigdata/watchdog（watchdog日志）
- upgrade：/var/log/Bigdata/upgrade（升级OMS日志）
- UpdateService：/var/log/Bigdata/update-service（升级服务日志）
- Sudo：/var/log/Bigdata/sudo（sudo脚本执行日志）
- OS：/var/log/*message文件*（OS系统日志）
- OS Performance：/var/log/osperf（OS性能统计日志）
- OS Statistics：/var/log/osinfo/statistics（OS参数配置信息日志）

日志归档规则：

Manager的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过10MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件。

表 10-64 Manager 日志列表

日志类型	日志文件名	描述
Controller运行日志	controller.log	记录组件安装、升级、配置、监控、告警和日常运维操作日志。
	controller_client.log	Rest接口运行日志。
	acs.log	Acs运行日志。
	acs_spnego.log	acs中spnego用户日志
	aos.log	Aos运行日志。
	plugin.log	Aos插件日志
	backupplugin.log	备份恢复进程运行日志
	controller_config.log	配置运行日志
	controller_nodesetup.log	Controller加载任务日志
	controller_root.log	Controller进程系统日志
	controller_trace.log	Controller与NodeAgent之间RPC通信日志
	controller_monitor.log	监控日志
	controller_fsm.log	状态机日志
	controller_alarm.log	Controller发送告警日志
	controller_backup.log	Controller备份恢复日志
	install.log，restore_package.log，installPack.log，distributeAdapterFiles.log，install_os_optimization.log	oms安装日志
	oms_ctl.log	oms启停日志
	preInstall_client.log	客户端安装前预处理日志
	installntp.log	ntp安装日志
	modify_manager_param.log	修改Manager参数日志

日志类型	日志文件名	描述
	backup.log	OMS备份脚本运行日志
	supressionAlarm.log	告警脚本运行日志
	om.log	生成om证书日志
	backupplugin_ctl.log	备份恢复插件进程启动日志
	getLogs.log	采集日志脚本运行日志
	backupAuditLogs.log	审计日志备份脚本运行日志
	certStatus.log	证书定期检查日志
	distribute.log	证书分发日志
	ficertgenetrate.log	证书替换日志，包括生成二级证书、cas证书、httpd证书的日志。
	genPwFile.log	生成证书密码文件日志
	modifyproxyconf.log	修改HTTPD代理配置的日志
	importTar.log	证书导入信任库日志
Httpd	install.log	Httpd安装日志
	access_log, error_log	Httpd运行日志
logman	logman.log	日志打包工具日志。
NodeAgent	install.log，install_os_optimization.log	NodeAgent安装日志
	installntp.log	ntp安装日志
	start_ntp.log	ntp启动日志
	ntpChecker.log	ntp检查日志
	ntpMonitor.log	ntp监控日志
	heartbeat_trace.log	NodeAgent与Controller心跳日志
	alarm.log	告警日志
	monitor.log	监控日志
	nodeagent_ctl.log，start-agent.log	NodeAgent启动日志
	agent.log	NodeAgent运行日志

日志类型	日志文件名	描述
	cert.log	证书日志
	agentplugin.log	监控agent侧插件运行日志
	omaplugin.log	OMA插件运行日志
	diskhealth.log	磁盘健康检查日志
	supressionAlarm.log	告警脚本运行日志
	updateHostFile.log	更新主机列表日志
	collectLog.log	节点日志采集脚本运行日志
	host_metric_collect.log	主机指标采集运行日志
	checkfileconfig.log	文件权限配置检查运行日志
	entropycheck.log	熵值检查运行日志
	timer.log	节点定时调度日志
	pluginmonitor.log	组件监控插件日志
	agent_alarm_py.log	NodeAgent检查文件权限发送告警日志
okerberos	addRealm.log，modifyKerberosRealm.log	切域日志
	checkservice_detail.log	Okerberos健康检查日志
	genKeytab.log	生成keytab日志
	KerberosAdmin_genConfigDetail.log	启动kadmin进程时，生成kadmin.conf的运行日志
	KerberosServer_genConfigDetail.log	启动krb5kdc进程时，生成krb5kdc.conf的运行日志
	oms-kadmind.log	kadmin进程的运行日志
	oms_kerberos_install.log，postinstall_detail.log	okerberos安装日志
	oms-krb5kdc.log	krbkdc运行日志
	start_detail.log	okerberos启动日志
	realmDataConfigProcess.log	切域失败，回滚日志

日志类型	日志文件名	描述
	stop_detail.log	okerberos停止日志
oldapserver	ldapserver_backup.log	Oldapserver备份日志
	ldapserver_chk_service.log	Oldapserver健康检查日志
	ldapserver_install.log	Oldapserver安装日志
	ldapserver_start.log	Oldapserver启动日志
	ldapserver_status.log	Oldapserver进程状态检查日志。
	ldapserver_stop.log	Oldapserver停止日志
	ldapserver_wrap.log	Oldapserver服务管理日志。
	ldapserver_uninstall.log	Oldapserver卸载日志
	restart_service.log	Oldapserver重启日志
	ldapserver_unlockUser.log	记录解锁Ldap用户和管理帐户的日志
metric_agent	gc.log	MetricAgent JAVA虚拟机gc日志
	metric_agent.log	MetricAgent运行日志
	metric_agent_qps.log	MetricAgent内部队列长度及qps信息记录日志
	metric_agent_root.log	MetricAgent所有运行日志
	start.log	MetricAgent启停信息日志
omm	omsconfig.log	OMS配置日志
	check_oms_heartbeat.log	OMS心跳运行日志
	monitor.log	OMS监控日志
	ha_monitor.log	HA_Monitor操作日志
	ha.log	HA操作日志
	fms.log	告警日志
	fms_ha.log	告警的HA监控日志
	fms_script.log	告警控制日志
	config.log	告警配置日志

日志类型	日志文件名	描述
	iam.log	IAM日志
	iam_script.log	IAM控制日志
	iam_ha.log	IAM的HA监控日志
	config.log	IAM配置日志
	operatelog.log	IAM操作日志
	heartbeatcheck_ha.log	OMS心跳的HA监控日志
	install_oms.log	OMS安装日志
	pms_ha.log	监控的HA监控日志
	pms_script.log	监控控制日志
	config.log	监控配置日志
	plugin.log	监控插件运行日志
	pms.log	监控日志
	ha.log	HA运行日志
	cep_ha.log	CEP的HA监控日志
	cep_script.log	CEP控制日志
	cep.log	CEP日志
	config.log	CEP配置日志
	omm_gaussdba.log	gaussdb的HA监控日志
	gaussdb-<SERIAL>.log	gaussdb运行日志
	gs_ctl-<DATE>.log	gaussdb控制日志的归档日志
	gs_ctl-current.log	gaussdb控制日志
	gs_guc-current.log	gaussdb操作日志
	encrypt.log	omm加密日志
	omm_agent_ctl.log	OMA控制日志
	oma_monitor.log	OMA监控日志
	install_oma.log	OMA安装日志
	config_oma.log	OMA配置日志
	omm_agent.log	OMA运行日志
	acs.log	acs资源日志。

日志类型	日志文件名	描述
	aos.log	aos资源日志
	controller.log	controller资源日志
	feed_watchdog.log	feed_watchdog资源日志
	floatip.log	floatip资源日志
	ha_ntp.log	ntp资源日志
	httpd.log	httpd资源日志
	okerberos.log	okerberos资源日志
	oldap.log	oldap资源日志
	tomcat.log	tomcat资源日志
	send_alarm.log	管理节点HA告警发送脚本运行日志
timestamp	restart_stamp	NodeAgent启动时间
tomcat	cas.log，localhost_access_cas_log.log	cas运行日志
	catalina.log，catalina.out，host-manager.log，localhost.log，manager.log	tomcat运行日志
	localhost_access_web_log.log	记录访问FusionInsight Manager系统REST接口的日志
	web.log	web进程运行日志
	northbound_ftp_sftp.log，snmp.log	北向日志
	perfStats.log	性能数据统计日志
watchdog	watchdog.log，feed_watchdog.log	watchdog.log运行日志
update-service	omm_upd_server.log	updserver的运行日志
	omm_upd_agent.log	updagent的运行日志
	update-manager.log	updmanager的运行日志
	install.log	升级服务安装日志
	uninstall.log	升级服务卸载日志

日志类型	日志文件名	描述
	catalina.<时间>.log，catalina.out，host-manager.<时间>.log，localhost.<时间>.log，manager.<时间>.log，manager_access_log.<时间>.txt，web_service_access_log.<时间>.txt，catalina.log，gc-update-service.log.0.current，update-manager.controller，update-web-service.controller，update-web-service.log，commit_rm_distributed.log，commit_rm_upload_package.log，common_omagent_operator.log，forbid_monitor.log，initialize_package_atoms.log，initialize_unzip_pack.log，omm-upd.log，register_patch_pack.log，resume_monitor.logrollback_clear_patch.log，unregister_patch_pack.log，update-rcommupd.log，update-rcupdatemanager.log，update-service.log	升级服务运行日志
upgrade	upgrade.log_<时间>	升级OMS日志
	rollback.log_<时间>	回滚OMS日志
sudo	sudo.log	sudo脚本执行日志

日志级别

Manager中提供了如表10-65所示的日志级别。日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 10-65 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO记录系统及各事件正常运行状态信息
DEBUG	DEBUG记录系统及系统的调试信息。

日志格式

Manager的日志格式如下所示：

表 10-66 日志格式

日志类型	组件	格式	示例
Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-06-30 00:37:09,067 INFO [pool-1-thread-1] Completed Discovering Node. com.xxx.hadoop.om.controller.tasks.nodesetup.DiscoverNodeTask.execute(DiscoverNodeTask.java:299)

10.10.3 配置日志级别与文件大小

操作场景

如果需要在日志中调整记录的日志级别，则管理员可以修改FusionInsight Manager的日志级别。对于某个具体的服务，除了可以修改日志级别，还可以修改日志文件大小，防止磁盘空间不足日志无法保存。

对系统的影响

保存新的配置需要重启服务，此时对应的服务不可用。

修改 FusionInsight Manager 日志级别

1. 以**omm**用户登录主管理节点。
2. 执行以下命令，切换路径。

   cd ${BIGDATA_HOME}/om-server/om/sbin
3. 执行以下命令，修改日志级别。

   ./setLogLevel.sh *日志级别参数*

   日志级别参数如下，优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少：
   - “DEFAULT”：设置后恢复到默认日志级别。
   - “FATAL”：严重错误日志级别，设置后日志只会打印输出“FATAL”信息。
   - “ERROR”：错误日志级别，设置后日志打印输出“ERROR”和“FATAL”信息。
   - “WARN”：警告日志级别，设置后日志打印输出“WARN”、“ERROR”和“FATAL”信息。
   - “INFO”（默认）：提示信息日志级别，设置后日志打印输出“INFO”、“WARN”、“ERROR”和“FATAL”信息。
   - “DEBUG”：调试日志级别，设置后日志打印输出“DEBUG”、“INFO”、“WARN”、“ERROR”和“FATAL”信息。
   - “TRACE”：跟踪日志级别，设置后日志打印输出"TRACE"、“DEBUG”、“INFO”、“WARN”、“ERROR”和“FATAL”信息。

   说明

   由于开源中定义的不同，组件的日志级别定义略有差异。
4. 验证日志级别设置已生效，请下载日志并查看。请参见**日志**。

修改服务日志级别与日志文件大小

说明

KrbServer，LdapServer以及DBService不支持修改服务日志级别与日志文件大小。

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务列表中的某个服务，切换到“配置”页签。

步骤4 选择“全部配置”，左边菜单栏中展开角色实例的菜单，单击所需修改的角色所对应的“日志”。

步骤5 搜索各项参数，获取参数说明，在参数配置页面勾选所需的日志级别或修改日志文件大小。日志文件大小需填写单位“MB”。

须知

- 系统会根据配置的日志大小自动清理日志，如果需要保存更多的信息请设置一个较大的数值。为确保日志文件的完整性，建议根据实际业务量大小，在日志文件基于规则清理前，手动将日志文件备份存储至其他文件夹中。
- 个别服务不支持通过界面修改日志级别。

步骤6 单击“保存”，在“保存配置”单击“确定”。

步骤7 验证日志级别设置已生效，请下载日志并查看。

----结束

10.10.4 配置审计日志本地备份数

操作场景

集群组件的审计日志按名称分类，保存在集群各节点“/var/log/Bigdata/audit”，OMS每天凌晨3点自动备份这些审计日志目录。

各节点审计日志目录会按<节点IP>.tar.gz的文件名压缩，所有压缩文件再按<yyyy-MM-dd_HH-mm-ss>.tar.gz的文件名格式，压缩保存在主管理节点“/var/log/Bigdata/audit/bk/”，同时备管理节点会同步保存一个相同的副本。

默认情况下，OMS备份的文件最大保留个数为90，该任务指导系统管理员配置此最大保留个数。

操作步骤

步骤1 以**omm**用户登录主管理节点。

说明

用户只需在主管理节点执行此操作，不支持在备管理节点上修改审计日志备份文件数，否则可能造成集群无法正常工作。

步骤2 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/sbin

步骤3 执行以下命令，修改审计日志备份文件数。

./modifyLogConfig.sh -m *最大保留个数*

OMS备份组件审计日志默认最大保留90个，可选值为“0”到“365”，如果设置的保留个数越大，会占用更多的磁盘空间。

显示如下结果，说明修改成功：

```
Modify log config successfully
```

----结束

10.10.5 查看角色实例日志

操作场景

FusionInsight Manager支持在线直接查看各角色实例的日志内容，

操作步骤

步骤1 登录FusionInsight Manager。

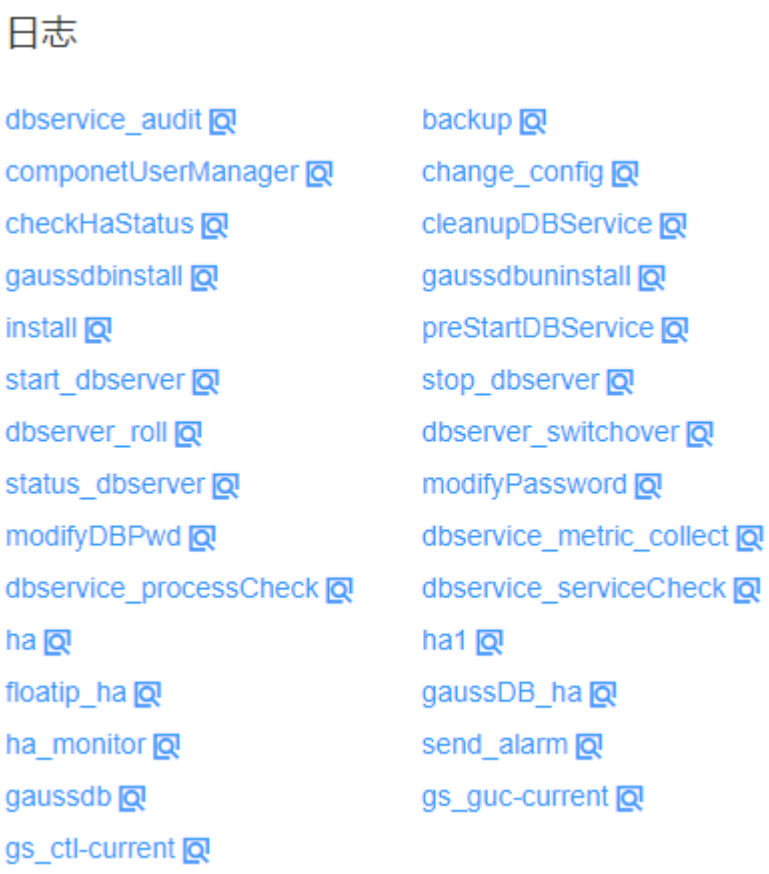
步骤2 选择“集群 > *待操作的集群名称* > 服务 > *服务名称* > 实例”，单击需要查看日志的实例名称，进入实例状态页面。

步骤3 在“日志”区域，单击要查看的日志文件名称，即可在线预览对应日志内容。

说明

- 在“主机”页面单击主机名称，在主机详情页面的“实例”区域，也可查看该主机上所有角色实例的日志文件。
- 日志内容默认最多显示100行，可单击“加载更多”按钮查看更多日志内容。单击“下载”按钮，可将该日志文件下载到本地。如需批量下载服务日志，请参考**下载日志**。

图 10-20 查看实例日志



----结束

10.11 备份恢复管理

10.11.1 备份恢复简介

概述

FusionInsight Manager提供对集群内的用户数据及系统数据的备份恢复能力，备份功能按组件提供。系统支持备份Manager的数据、组件元数据及业务数据。

备份功能支持将数据备份至本地磁盘（LocalDir）、本端HDFS（LocalHDFS）、远端HDFS（RemoteHDFS）、NAS（NFS/CIFS）、SFTP服务器（SFTP）、OBS，具体操作请参考**备份数据**。

对于支持多服务的组件，支持同服务多个实例的备份恢复功能且备份恢复操作与自身服务实例一致。

说明

MRS 3.1.0及之后版本才支持备份数据到OBS。

备份恢复任务的使用场景如下：

- 用于日常备份，确保系统及组件的数据安全。
- 当系统故障导致无法工作时，使用已备份的数据完成恢复操作。
- 当主集群完全故障，需要创建一个与主集群完全相同的镜像集群，可以使用已备份的数据完成恢复操作。

表 10-67 根据业务需要备份 Manager 配置数据

备份类型	备份内容	备份目录类型
OMS	默认备份集群管理系统中的数据库数据（不包含告警数据）以及配置数据。	• LocalDir • LocalHDFS • RemoteHDFS • NFS • CIFS • SFTP • OBS

表 10-68 根据业务需要备份组件元数据或其他数据

备份类型	备份内容	备份目录类型
DBService	备份DBService管理的组件（Loader、Hive、Spark、Oozie、Hue）的元数据。对于安装了多服务的集群，包含多个Hive和Spark服务实例的元数据。	• LocalDir • LocalHDFS • RemoteHDFS • NFS • CIFS • SFTP • OBS
Kafka	Kafka的元数据。	• LocalDir • LocalHDFS • RemoteHDFS • NFS • CIFS • OBS
NameNode	备份HDFS元数据。添加多个NameService后，支持不同NameService的备份恢复功能且备份恢复操作与默认实例“hacluster”一致。	• LocalDir • RemoteHDFS • NFS • CIFS • SFTP • OBS
Yarn	备份Yarn服务资源池相关信息。	
HBase	HBase系统表的tableinfo文件和数据文件。	

表 10-69 根据业务需要备份特定组件业务数据

备份类型	备份内容	备份目录类型
HBase	备份表级别的用户数据。对于安装了多服务的集群，支持多个HBase服务实例的备份恢复功能且备份恢复操作与HBase服务实例一致。	• RemoteHDFS • NFS • CIFS • SFTP
HDFS	备份用户业务对应的目录或文件。 **说明** 加密目录不支持备份恢复。	
Hive	备份表级别的用户数据。对于安装了多服务的集群，支持多个Hive服务实例的备份恢复功能且备份恢复操作与Hive服务实例一致。	

需要特别说明的是，部分组件不提供单独的数据备份与恢复功能：

- Kafka支持副本特性，在创建主题时可指定多个副本来备份数据。
- Mapreduce和Yarn的数据存放在HDFS上，故其依赖HDFS提供备份与恢复即可。
- ZooKeeper中存储的业务数据，其备份恢复能力由各上层组件按需独立实现。

原理

任务

在进行备份恢复之前，需要先创建备份恢复任务，并指定任务的参数，例如任务名称、备份数据源和备份文件保存的目录类型等等。通过执行备份恢复任务，用户可完成数据的备份恢复需求。在使用Manager执行恢复HDFS、HBase、Hive和NameNode数据时，无法访问集群。

每个备份任务可同时备份不同的数据源，每个数据源将生成独立的备份文件，每次备份的所有备份文件组成一个备份文件集，可用于恢复任务。备份任务支持将备份文件保存在Linux本地磁盘、本集群HDFS与备集群HDFS中。

备份任务提供全量备份或增量备份的策略，云数据备份任务不支持增量备份策略。如果备份的路径类型是NFS或CIFS，不建议使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

说明

任务运行规则：

- 某个任务已经处于执行状态，则当前任务无法重复执行，其他任务也无法启动。
- 周期任务自动执行时，距离该任务上次执行的时间间隔需要在120秒以上，否则任务推迟到下个周期启动。手动启动任务无时间间隔限制。
- 周期任务自动执行时，当前时间不得晚于任务开始时间120秒以上，否则任务推迟到下个周期启动。
- 周期任务锁定时无法自动执行，需要手动解锁。
- OMS、DBService、Kafka和NameNode备份任务开始执行前，若主管理节点“LocalBackup”分区可用空间小于20GB，则无法开始执行。

管理员在规划备份恢复任务时，请严格根据业务逻辑、数据存储结构、数据库或表关联关系，选择需要备份或者恢复的数据。系统默认创建间隔为1小时的周期备份任务“default-oms”、“default-*集群ID*”，支持全量备份OMS及集群的DBService、NameNode等元数据到本地磁盘。

快照

系统通过快照技术，快速备份数据。快照包含HBase快照、HDFS快照快照。

- HBase快照

  HBase快照是HBase表在特定时间的一个备份，该备份文件不复制业务数据，不影响RegionServer。HBase快照主要复制表的元数据，包含table descriptor，region info和HFile的引用信息。通过这些元数据信息可以恢复快照时间点之前的数据。

- HDFS快照

  HDFS快照是HDFS文件系统在特定时间点的只读备份副本，主要用于数据备份、用户误操作保护和灾难恢复的场景。

  任意HDFS目录均可以配置启用快照功能并创建对应的快照文件，为目录创建快照前系统会自动启用此目录的快照功能。创建快照不会对正常的HDFS操作有任何影响。每个HDFS目录最多可创建65536个快照。

如果一个HDFS目录已创建快照，那么在快照完全删除以前，此目录无法删除或修改名称。该目录的上级目录或子目录也无法再创建快照。

DistCp

DistCp（distributed copy）是一个用于在本集群HDFS中或不同集群HDFS间进行大量数据复制的工具。在HBase、HDFS或Hive元数据的备份恢复任务中，如果选择将数据备份在备集群HDFS中，系统将调用DistCp完成操作。主备集群请选择安装相同版本的MRS软件版本并安装集群系统。

DistCp使用Mapreduce来影响数据的分布、异常处理及恢复和报告，此工具会把指定列表中包含的多个源文件和目录输入不同的Map任务，每个Map任务将复制列表中指定文件对应分区的数据。

使用DistCp在两个集群的HDFS间进行数据复制，集群双方需要分别配置互信（同一个FusionInsight Manager管理下的集群不需要配置互信）和启用集群间拷贝功能。集群数据备份到另一个集群的HDFS时，需要安装Yarn组件，否则备份失败。

本地快速恢复

使用DistCp将本集群HBase、HDFS和Hive数据备份在备集群HDFS中以后，本集群HDFS保留了备份数据的快照。用户可以通过创建本地快速恢复任务，直接从本集群HDFS的快照文件中恢复数据。

NAS

NAS（Network Attached Storage）是一种特殊的专用数据存储服务器，包括存储器件和内嵌系统软件，可提供跨平台文件共享功能。利用NFS（支持NFSv3、NFSv4）和CIFS（支持SMBv2、SMBv3）协议，用户可以连通MRS的业务平面与NAS服务器，将数据备份至NAS或从NAS恢复数据。

说明

- 数据备份至NAS前，系统会自动将NAS共享地址挂载为本地分区。在备份结束后，系统会卸载NAS共享分区。
- 为防止备份恢复失败，数据备份及恢复期间，请勿访问NAS服务器挂载至本地的共享地址，如：“/srv/BigData/LocalBackup/nas”。
- 业务数据备份至NAS时，会使用DistCp。

规格

表 10-70 备份恢复特性规格

项目	参数
备份或恢复任务最大数量（个）	100
同一集群同时运行的任务数量（个）	1
等待运行的任务最大数量（个）	199
Linux本地磁盘最大备份文件大小（GB）	600

说明

若业务数据存储在Zookeeper中的上层组件，在备份恢复这类数据时，需确保单个备份或恢复任务的znode数量不会过大，否则会造成任务失败，并影响Zookeeper的服务性能。可通过如下方法确认单个备份或恢复任务的znode数量：

- 单个备份或恢复任务的znode数量要少于操作系统的文件句柄限制。查看句柄限制的方式如下：
  1. 使用shell命令输入：**cat /proc/sys/fs/file-max**，用于查看系统级的最大限制。
  2. 使用shell命令输入：**ulimit -n**，用于查看用户级的限制。
- 对于父目录的znode数量超过上述限制的情形，可以通过其子目录进行批量备份与恢复。使用Zookeeper提供的客户端脚本查看znode数量的方式：
  1. 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Zookeeper > 实例”，查看Zookeeper各角色的管理IP。
  2. 登录客户端所在节点，执行如下命令：

     zkCli.sh -server *ip*:*port*，其中*ip*可以为任意管理IP，*port*默认值是2181。
  3. 当看到如下输出信息时，表示已经成功连接上Zookeeper服务器。
     ```
     WatchedEvent state:SyncConnected type:None path:null
     [zk: ip:port(CONNECIED) 0]
     ```
  4. 使用getusage命令查看待备份目录的znode数量，例如：

     getusage /hbase/region，输出结果中"Node count=xxxxxx"即表示region目录下存储的znode数量。

表 10-71 “default”任务规格

项目	OMS	HBase	Kafka	DBService	NameNode
备份周期	1小时				
最大备份数	168个（7天历史数据）				24个（1天历史数据）
单个备份文件最大大小	10MB	10 MB	512MB	100MB	20GB
最大占用磁盘大小	1.64GB	1.64 GB	84GB	16.41GB	480GB
备份数据保存位置	主备管理节点“*数据存放路径*/LocalBackup/”				

说明

- 默认任务保存的备份数据，请管理员根据企业运维要求，定期转移并保存到集群外部。
- 管理员可直接创建DistCp备份任务将OMS、DBService和NameNode等的数据保存到外部集群。
- 集群数据的备份任务运行时长可根据要备份的数据量除以集群与备份设备之间的网络带宽来计算得出，在实际场景中，建议将计算得出的时常乘以1.5作为任务执行时长参考值。
- 执行数据备份任务会对集群的最大IO性能产生影响，建议备份任务运行时间与集群业务高峰错开。

10.11.2 备份数据

10.11.2.1 备份 OMS 数据

操作场景

为了确保FusionInsight Manager系统日常数据安全，或者系统管理员需要对Manager进行重大操作（如扩容、减容等）前后，需要对Manager数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

管理员可以通过FusionInsight Manager创建备份Manager任务并备份数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 设置“备份对象”为“OMS”。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-72 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。

参数名称	描述
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“OMS”。

步骤7 在“OMS”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。

- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “备份时使用集群”：填写备份目录对应的集群名称。
  - “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

  - “源集群”：选择要备份数据使用的Yarn队列所在的集群。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和源集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为“*备份任务名_任务创建时间*”，用于保存数据源的备份文件。

备份文件的名称为“*版本号_数据源_任务执行时间*.tar.gz”。

----结束

10.11.2.2 备份 DBService 数据

操作场景

为了确保DBService日常数据安全，或者系统管理员需要对DBService进行重大操作（如升级或迁移等）时，需要对DBService数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份DBService任务并备份数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-73 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。

参数名称	描述
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
备份策略	• 首次全量备份，后续增量备份 • 每次都全量备份 • 每n次进行一次全量备份 **说明** • 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 • 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“DBService”。

说明

若安装了多个DBService服务，默认备份所有DBService服务，可单击“指定服务”指定需要备份的DBService服务。

步骤7 在“DBService”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。

  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和源集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。

备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

10.11.2.3 备份 HBase 元数据

操作场景

为了确保HBase元数据（主要包括tableinfo文件和HFile）安全，防止因HBase的系统表目录或者文件损坏导致HBase服务不可用，或者系统管理员需要对HBase系统表进行重大操作（如升级或迁移等）时，需要对HBase元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份HBase任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- HBase的“fs.defaultFS”配置参数需要与Yarn、HDFS的配置保持一致。
- 如果HBase数据存储在本地HDFS，支持将HBase元数据备份到OBS。如果HBase数据存储在OBS，则不支持数据备份。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-74 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“元数据和其它数据”下的“HBase”。

说明

若安装了多个HBase服务，默认备份所有HBase服务，可单击“指定服务”指定需要备份的HBase服务。

步骤7 在“HBase”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。

  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和源集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

10.11.2.4 备份 HBase 业务数据

操作场景

为了确保HBase日常数据安全，或者系统管理员需要对HBase进行重大操作（如升级或迁移等），需要对HBase业务数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份HBase任务并备份数据。支持创建任务自动或手动备份数据。

HBase备份业务数据时，可能存在以下场景：

- 用户创建HBase表时，“KEEP_DELETED_CELLS”属性默认值为“false”，备份该HBase表时会将已经删除的数据备份，可能导致恢复后出现垃圾数据。请根据业务需要，在创建HBase表时手动修改参数值为“true”。
- 用户在HBase表写入数据时手动指定了时间戳，且时间早于上一次该HBase表的备份时间，则在增量备份任务中可能无法备份新数据。
- HBase备份功能不支持对HBase的global或者命名空间的读取、写入、执行、创建和管理权限的访问控制列表（ACL）进行备份，恢复HBase数据后需要管理员在FusionInsight Manager上重新设置角色的权限。
- 已创建的HBase备份任务，如果本次备份任务在备集群的备份数据丢失，当下次执行增量备份时备份任务将失败，需要重新创建HBase的备份任务。若下次执行全量则备份正常。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用HDFS客户端，以**hdfs**用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建HDFS快照的目录清单，确保待备份的数据文件所在HDFS路径的父目录或子目录不存在HDFS快照，否则无法创建备份任务。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- HBase的“fs.defaultFS”配置参数需要与Yarn，HDFS的配置保持一致。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-75 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“业务数据”下的“HBase > HBase”。

步骤7 在“HBase”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。

- “用户名”：填写使用SFTP协议连接服务器时的用户名。
- “密码”：填写使用SFTP协议连接服务器时的密码。
- “服务器共享路径”：SFTP服务器上的备份路径。
- “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为20。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为100。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在“备份内容”中，选择一个或多个需要备份的HBase表。

支持两种方式选择备份数据：

- 直接选择

  单击导航中某个命名空间的名称，将展开显示此命名空间中的所有表，勾选指定的表。

- 正则表达式筛选

  a. 单击“正则表达式输入”。

  b. 根据界面提示，在第一个输入框填写HBase表所在的命名空间，需要与当前存在的命名空间完全匹配。例如“default”。

  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选命名空间中所有的表，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的表，如**tb1**可输入“tb\d*”。

  d. 单击“刷新”，在“目录名称”查看筛选的表。

  e. 单击“同步”保存筛选结果。

  说明

  - 输入正则表达式时，可以使用＋和－增加或删除一条表达式。
  - 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 队列名称不正确。
- 待备份的HBase表数据文件所在HDFS路径的父目录或子目录存在HDFS快照。
- 待备份的目录或表不存在。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为“*备份任务名_数据源_任务创建时间*”，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

10.11.2.5 备份 NameNode 数据

操作场景

为了确保NameNode日常数据安全，或者系统管理员需要对NameNode进行重大操作（如升级或迁移等），需要对NameNode数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份NameNode任务。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-76 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。

参数名称	描述
备份策略	仅支持“每次都全量备份”。 **说明** • 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 • 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“NameNode”。

步骤7 在“NameNode”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“*数据存放路径*/LocalBackup/”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：备集群NameNode的业务平面IP地址。
  - “目的端路径”：备份文件存放的位置。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为“*备份任务名_任务创建时间*”，用于保存数据源的备份文件。

备份文件的名称为“*版本号_数据源_任务执行时间*.tar.gz”。

----结束

10.11.2.6 备份 HDFS 业务数据

操作场景

为了确保HDFS日常用户的业务数据安全，或者系统管理员需要对HDFS进行重大操作（如升级或迁移等），需要对HDFS数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份HDFS任务并备份数据。支持创建任务自动或手动备份数据。

说明

加密目录不支持备份恢复。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用HDFS客户端，以“hdfs”用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建HDFS快照的目录清单，确保待备份的数据文件所在HDFS路径的父目录或子目录不存在HDFS快照，否则无法创建备份任务。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-77 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。

参数名称	描述
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“HDFS”。

步骤7 在“HDFS”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。

  - “服务器IP地址”：填写NAS服务器IP地址。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。

- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在HDFS“备份内容”中，根据业务需要选择一个或多个需要备份的HDFS目录。

支持两种方式选择备份数据：

- 直接选择

  单击导航中某个目录的名称，将展开显示此目录中的所有子目录，勾选指定的目录。

- 正则表达式筛选

  a. 单击“正则表达式输入”。

  b. 根据界面提示，在第一个输入框填写目录的父目录完整路径，需要与当前存在的目录完全匹配。例如“/tmp”。

  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选父目录中所有的文件或子目录，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的文件，如**file1**可输入“file\d*”。

  d. 单击“刷新”，在“目录名称”查看筛选的目录。

  e. 单击“同步”保存筛选结果。

  说明

  - 输入正则表达式时，可以使用+和—增加或删除一条表达式。
  - 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。
  - 备份目录不可包含长期写入的文件，否则会导致备份任务失败，因此不建议对顶层目录进行操作，例如“/user”、“/tmp”、“/mr-history”。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 队列名称不正确。
- 待备份的数据文件所在HDFS路径的父目录或子目录存在HDFS快照。
- 待备份的目录或表不存在。
- NameService名称不正确。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为“*备份任务名_数据源_任务创建时间*”，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

10.11.2.7 备份 Hive 业务数据

操作场景

为了确保Hive日常用户的业务数据安全，或者系统管理员需要对Hive进行重大操作（如升级或迁移等），需要对Hive数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份Hive任务。支持创建任务自动或手动备份数据。

- Hive备份恢复功能不支持识别用户的Hive表、索引、视图等对象在业务和结构上存在的关联关系。用户在执行备份恢复任务时，需要根据业务场景管理统一的恢复点，防止影响业务正常运行。
- Hive备份恢复功能不支持Hive on RDB数据表，需要在外部数据库中单独备份恢复原始数据表。
- 已创建的Hive备份任务且包含Hive on HBase表，如果本次备份任务在备集群的备份数据丢失，当下次执行增量备份时备份任务将失败，需要重新创建Hive的备份任务。若下次执行全量则备份正常。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用HDFS客户端，以“hdfs”用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建HDFS快照的目录清单，确保待备份的数据文件所在HDFS路径的父目录或子目录不存在HDFS快照，否则无法创建备份任务。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-78 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“Hive > Hive”。

步骤7 在“Hive”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在“备份内容”中，选择一个或多个需要备份的Hive表。

支持两种方式选择备份数据：

- 直接选择

  单击导航中某个数据库的名称，将展开显示此数据库中的所有表，勾选指定的表。

- 正则表达式筛选

  a. 单击“正则表达式输入”。

  b. 根据界面提示，在第一个输入框填写Hive表所在的数据库，需要与当前存在的数据库完全匹配。例如“defalut”。

  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选数据库中所有的表，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的表，如**tb1**可输入“tb\d*”。

  d. 单击“刷新”，在“目录名称”查看筛选的表。

  e. 单击“同步”保存筛选结果。

  说明

  - 输入正则表达式时，可以使用+和—增加或删除一条表达式。
  - 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 队列名称不正确。
- 待备份的数据文件所在HDFS路径的父目录或子目录存在HDFS快照。
- 待备份的目录或表不存在。
- NameService名称不正确。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为*“备份任务名_数据源_任务创建时间”*，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

10.11.2.8 备份 Kafka 元数据

操作场景

为了确保Kafka元数据安全，或者系统管理员需要对ZooKeeper进行重大操作（如升级或迁移等）时，需要对Kafka元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份Kafka任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 10-79 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。

参数名称	描述
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“Kafka”。

说明

若安装了多个Kafka服务，默认备份所有Kafka服务，可单击“指定服务”指定需要备份的Kafka服务。

步骤7 在“Kafka”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“*数据存放路径*/LocalBackup/”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。

- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。

  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

10.11.3 恢复数据

10.11.3.1 恢复 OMS 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对FusionInsight Manager系统进行重大数据调整等操作后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，需要对Manager进行恢复数据操作。

管理员可以通过FusionInsight Manager创建恢复Manager任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的Manager数据。

对系统的影响

- 恢复过程中需要重启Controller，重启时FusionInsight Manager无法登录和操作。
- 恢复过程中需要重启所有集群，集群重启时无法访问。
- Manager数据恢复后，会丢失从备份时刻到恢复时刻之间的数据，例如系统设置、用户信息、告警信息或审计信息。可能导致无法查询到数据，或者某个用户无法访问集群。
- Manager数据恢复后，系统将强制各集群的LdapServer从OLadp同步一次数据。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查OMS资源状态是否正常，检查各集群的LdapServer实例状态是否正常。如果不正常，不能执行恢复操作。
- 检查集群主机和服务的状态是否正常。如果不正常，不能执行恢复操作。
- 检查恢复数据时集群主机拓扑结构与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份。
- 检查恢复数据时集群中已添加的服务与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份。
- 停止依赖集群运行的上层业务应用。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理” 。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 选择“运维 > 备份恢复 > 恢复管理 > 创建 ”。

步骤4 在“任务名称”填写恢复任务的名称。

步骤5 在“恢复对象”选择“OMS”。

步骤6 勾选“OMS”。

步骤7 在“OMS”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “恢复时使用集群”：表示选择恢复任务执行时使用集群的名称。
  - “源NameService名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “源集群”：选择恢复数据使用的Yarn队列所在的集群。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “SFTP”：表示备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤8 单击“确定”保存。

步骤9 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。

- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤10 以**omm**用户分别登录主、备管理节点。

步骤11 执行以下命令，重新启动OMS。

sh ${BIGDATA_HOME}/om-server/om/sbin/restart-oms.sh

提示以下信息表示命令执行成功：

```
start HA successfully.
```

执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**，查看管理节点的“HAAllResOK”是否为“Normal”，并可以重新登录FusionInsight Manager表示重启OMS成功。

步骤12 在FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > KrbServer > 更多 > 同步配置”，单击“确定”，等待KrbServer服务配置同步过程完成。

步骤13 选择“集群 > *待操作集群的名称* > 更多 > 同步配置”，单击“确定”，等待集群配置同步成功。

步骤14 选择“集群 > *待操作集群的名称* > 更多 > 重启”，输入当前登录的用户密码确认身份，单击“确定”，等待集群重启成功。

----结束

10.11.3.2 恢复 DBService 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对DBService进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对DBService进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复DBService任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的DBService数据。
- MRS集群中默认使用DBService保存Hive、Hue、Loader、Spark、Oozie的元数据。恢复DBService的数据将恢复全部相关组件的元数据。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查DBService主备实例状态是否正常。如果不正常，不能执行恢复操作。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。
  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“DBService”。

说明

若安装了多个DBService服务，请勾选需要恢复的DBservice服务名称。

步骤8 在“DBService”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。
  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。
  选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “源NameService名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。
  选择此参数值，还需要配置以下参数：

  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

10.11.3.3 恢复 HBase 元数据

操作场景

为了确保HBase元数据（主要包括tableinfo文件和HFile）安全，防止因HBase的系统表目录或者文件损坏导致HBase服务不可用，或者系统管理员需要对HBase系统表进行重大操作（如升级或迁移等）时，需要对HBase元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建恢复HBase任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的HBase数据。
- 建议一个恢复任务只恢复一个组件的元数据，避免因停止某个服务或实例影响其他组件的数据恢复。同时恢复多个组件数据，可能导致数据恢复失败。

  HBase元数据不能与NameNode元数据同时恢复，会导致数据恢复失败。

对系统的影响

- 元数据恢复前，需要停止HBase服务，在这期间所有上层应用都会受到影响，无法正常工作。
- 元数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 元数据恢复后，需要重新启动HBase的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 检查HBase元数据备份文件保存路径。
- 恢复HBase元数据需要先停止HBase服务。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择 “运维 > 备份恢复 > 备份管理” 。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“元数据和其他数据”下的“HBase”。

说明

若安装了多个HBase服务，请勾选需要恢复的HBase服务名称。

步骤8 在“HBase”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。

  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

说明

MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

10.11.3.4 恢复 HBase 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对HBase进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对HBase业务数据进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复HBase任务并恢复数据。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的HBase数据。

对系统的影响

- 恢复过程的数据还原阶段，系统会把待恢复的HBase表禁用，此时无法访问该表。还原阶段可能需要几分钟时间，此时HBase的上层应用无法正常工作。
- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动HBase的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查HBase备份文件保存路径。

- 停止HBase的上层应用。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“业务数据”下的“HBase”。

步骤8 在“HBase”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。

- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

步骤9 在“数据配置”中的“备份数据”列根据业务需要勾选一个或多个需要恢复的已备份数据，并在“目标名称空间”列，指定备份数据恢复的命名空间。

“目标名称空间”建议选择一个备份命名空间不同的位置。

步骤10 在“强制覆盖”选择“true”，表示存在同名数据表时强制恢复备份的所有数据，如果数据表中存在备份后新增加的数据，那恢复后将丢失这些数据。选择“false”表示存在同名表时不执行恢复任务。

步骤11 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的命名空间，校验失败。
- 如果不满足强制覆盖的条件，校验失败。

步骤12 单击“确定”保存。

步骤13 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤14 检查是否是在全新安装，或者重新安装HBase的环境中恢复了HBase数据。

- 是，需要管理员在FusionInsight Manager上根据原有的业务规划重新设置角色的权限。
- 否，任务结束。

----结束

10.11.3.5 恢复 NameNode 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对NameNode进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对NameNode进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复NameNode任务并恢复数据。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的NameNode数据。
- 建议一个恢复任务只恢复一个组件的元数据，避免因停止某个服务或实例影响其他组件的数据恢复。同时恢复多个组件数据，可能导致数据恢复失败。

  HBase元数据不能与NameNode元数据同时恢复，会导致数据恢复失败。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 恢复数据后需要重启NameNode，重启完成前NameNode不可访问。
- 恢复数据后可能导致元数据与业务数据无法匹配，HDFS进入安全模式且HDFS服务启动失败。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 登录FusionInsight Manager，请参见**登录管理系统**。
- 在FusionInsight Manager停止所有待恢复数据的NameNode角色实例，其他的HDFS角色实例必须保持正常运行，恢复数据后重启NameNode。NameNode角色实例重启前无法访问。
- 检查NameNode备份文件保存路径是否保存在主管理节点“*数据存放路径*/LocalBackup/”。

操作步骤

步骤1 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > NameNode”，查看待恢复数据的NameNode角色实例是否已经停止，如果NameNode角色实例未停止，请停止NameNode角色实例运行。

步骤2 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤3 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤4 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤5 单击“创建”。

步骤6 在“任务名称”填写恢复任务的名称。

步骤7 在“恢复对象”选择待操作的集群。

步骤8 在“恢复配置”，勾选“NameNode”。

步骤9 在“NameNode”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：表示备份文件在本地磁盘中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：

- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “服务器IP地址”：填写NAS服务器IP地址。
- “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
- “用户名”：填写配置CIFS协议时设置的用户名。
- “密码”：填写配置CIFS协议时设置的密码。
- “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

● “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

选择此参数值，还需要配置以下参数：

- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “服务器IP地址”：填写备份数据的服务器IP地址。
- “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
- “用户名”：填写使用SFTP协议连接服务器时的用户名。
- “密码”：填写使用SFTP协议连接服务器时的密码。
- “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

● “OBS”：表示将备份文件保存在OBS中。

选择此参数值，还需要配置以下参数：

- “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

说明

MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤10 单击“确定”保存。

步骤11 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤12 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 更多 > 重启服务”。

在弹出窗口中输入当前登录的管理员密码确认管理员身份，单击“确定”。界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

10.11.3.6 恢复 HDFS 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对HDFS进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对HDFS进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复HDFS任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的HDFS数据。
- 对于Yarn任务运行时使用的目录（例如“/tmp/logs”、“/tmp/archived”、“/tmp/hadoop-yarn/staging”），不能进行HDFS恢复操作，否则进行恢复的Distcp任务会由于文件丢失而导致恢复失败。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动HDFS的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查HDFS备份文件保存路径。
- 停止HDFS的上层应用。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“业务数据”下的“HDFS”。

步骤8 在“HDFS”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。

  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。

- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

步骤9 在“数据配置”中的“备份数据”列根据业务需要勾选一个或多个需要恢复的已备份数据，并在“目标路径”列，指定备份数据恢复后的位置。

“目标路径”建议选择一个与目的端路径不同的新路径。

步骤10 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的恢复目录，校验失败。

步骤11 单击“确定”保存。

步骤12 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

10.11.3.7 恢复 Hive 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对Hive进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对Hive进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复Hive任务并恢复数据。只支持创建任务手动恢复数据。

Hive备份恢复功能不支持识别用户的Hive表、索引、视图等对象在业务和结构上存在的关联关系。用户在执行备份恢复任务时，需要根据业务场景管理统一的恢复点，防止影响业务正常运行。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的Hive数据。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动Hive的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 规划好恢复数据保存表的数据库，数据表在HDFS的保存位置，以及访问恢复数据的用户清单。
- 检查Hive备份文件保存路径。
- 停止Hive的上层应用。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“Hive”。

步骤8 在“Hive”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。

  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的Hive备份文件集。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的Hive备份文件集。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的Hive备份文件集。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“1”。

步骤9 在“数据配置”的“数据选择”中，根据业务需要勾选一个或多个需要恢复的已备份数据，并分别在“目标数据库”和“目标路径”列，指定备份数据恢复后的数据库和文件保存位置。

配置约束：

- 支持恢复到原数据库，但数据表保存在一个与目的端路径不同的新路径。
- 如果恢复Hive的索引表，请同时选择恢复索引表对应的Hive数据表。
- 如果为了防止影响当前数据，选择了新的恢复目录，那么新目录需要手动授予HDFS权限，使对备份表拥有权限的用户可以访问此目录。
- 支持恢复到其他数据库。如果恢复到其他数据库，那么此数据库对应在HDFS中的目录，需要手动授予HDFS权限，使对备份表拥有权限的用户可以访问此目录。

步骤10 在“强制覆盖”选择“true”，表示存在同名数据表时强制恢复备份的所有数据，如果数据表中存在备份后新增加的数据，那恢复后将丢失这些数据。选择“false”表示存在同名表时不执行恢复任务。

步骤11 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的恢复目录，校验失败。
- 如果不满足强制覆盖的条件，校验失败。

步骤12 单击“确定”保存。

步骤13 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

10.11.3.8 恢复 Kafka 元数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对ZooKeeper进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，导致Kafka组件全部故障无法使用，或者迁移数据到新集群的场景中，需要对Kafka元数据进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复Kafka任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复Kafka元数据，建议手动备份最新Kafka元数据后，再执行恢复操作。否则会丢失从备份时刻到恢复时刻之间的Kafka元数据信息。

对系统的影响

- 元数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 元数据恢复后，Kafka的消费者在ZooKeeper上保存的offset信息将会回退，可能导致重复消费。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 先停止Kafka服务，待恢复完成后，再启动Kafka服务。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择 “运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“Kafka”。

说明

若安装了多个Kafka服务，请勾选需要恢复的Kafka服务名称。

步骤8 在“Kafka”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “源NameService名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “NFS”：表示备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

10.11.4 启用集群间拷贝功能

操作场景

当用户需要将保存在HDFS中的数据从当前集群备份到另外一个集群时，需要使用DistCp工具。DistCp工具依赖于集群间拷贝功能，该功能默认未启用。拷贝数据的集群双方都需要配置。

管理员可以根据以下指导，在FusionInsight Manager修改参数以启用集群间拷贝功能。启用之后即可创建将数据备份至远端HDFS（RemoteHDFS）的备份任务。

对系统的影响

启用集群间复制功能需要重启Yarn，服务重启期间无法访问。

前提条件

- 拷贝数据的集群的HDFS的参数“hadoop.rpc.protection”需使用相同的数据传输方式。默认设置为“privacy”表示加密，“authentication”表示不加密。
- 对于安全模式的集群，集群之间需要配置系统互信。

操作步骤

步骤1 登录其中一个集群的FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，单击“全部配置”。

步骤3 左边菜单栏中选择“Yarn > 集群间拷贝”。

步骤4 修改参数“dfs.namenode.rpc-address”，在“haclusterX.remotenn1”右侧填写对端集群其中一个NameNode实例的业务IP和RPC端口，在“haclusterX.remotenn2”右侧填写对端集群另外一个NameNode实例的业务IP和RPC端口。

“haclusterX.remotenn1”和“haclusterX.remotenn2”不区分主备NameNode。NameNode RPC端口默认为“8020”，不支持通过Manager修改。

修改后参数值例如：“10.1.1.1:8020”和“10.1.1.2:8020”。

说明

- 如果本集群数据要备份至多个集群的HDFS中，可以继续配置对应的NameNode RPC地址至haclusterX1、haclusterX2、haclusterX3、haclusterX4。

步骤5 单击“保存”，并在确认对话框中单击“确定”。

步骤6 重启Yarn服务。

步骤7 登录另外一个集群的FusionInsight Manager，重复**步骤2**～**步骤6**。

----结束

10.11.5 管理本地快速恢复任务

操作场景

使用DistCp备份数据时，本集群HDFS中将保存备份数据的快照信息。FusionInsight Manager支持使用本地的快照快速恢复数据，减少从备集群恢复数据使用的时间。

管理员可以通过FusionInsight Manager与本集群HDFS保存的快照信息，创建本地快速恢复任务并执行恢复任务。

操作步骤

步骤1 登录FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在备份任务列表已创建任务的“操作”列，单击“恢复”。

步骤3 确认界面是否提示“没有可快速恢复的数据，请在恢复管理界面创建恢复任务进行恢复。”。

- 是，备份任务未在主集群产生备份数据快照，任务结束。
- 否，可以创建本地快速恢复任务，执行**步骤4**。

说明

元数据不支持快速恢复。

步骤4 在“任务名称”填写本地快速恢复任务的名称。

步骤5 在“备份配置”选择数据源。

步骤6 在“可恢复点列表”选择一个包含目标备份数据的恢复点。

步骤7 在“队列名称”填写任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。

步骤8 在“数据配置”选择需要恢复的对象。

步骤9 单击“校验”，界面显示“校验恢复任务配置成功”。

步骤10 单击“确定”。

步骤11 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

任务执行完成后，“任务状态”显示为“成功”。

----结束

10.11.6 修改备份任务

操作场景

系统管理员可以通过FusionInsight Manager修改已创建的备份任务的配置参数，以适应业务需求的变化。不支持修改任何恢复任务配置参数，只能查看恢复任务的配置参数。

对系统的影响

修改备份任务后，新的参数在下一次执行任务时生效。

前提条件

- 已创建备份任务。
- 已根据业务实际需求，规划新的备份任务策略。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，单击“配置”，打开修改配置页面。

在新页面中修改任务参数，支持修改的主要参数项如下：

- 开始时间
- 周期
- 目的端NameService名称

- 目的端NameNode IP地址
- 目的端路径
- 最大备份数
- 最大恢复点个数
- 最大map数
- 单个map的最大带宽

说明

修改某个备份任务参数“目的端路径”后，第一次执行此任务默认为全量备份。

步骤3 单击“确定”保存。

----结束

10.11.7 查看备份恢复任务

操作场景

系统管理员可以通过FusionInsight Manager查看已创建的备份恢复任务，以及任务的运行情况。

前提条件

登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复”。

步骤2 单击“备份管理”或“恢复管理”。

步骤3 在任务列表中，查看“任务状态”与“任务进度”列获取上一次任务运行的结果。绿色表示运行成功，红色表示运行失败。

步骤4 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”或单击“查询历史”，打开备份恢复任务运行记录。

在弹出的窗口中，在指定一次执行记录前单击，打开此次任务运行的日志信息。

----结束

相关任务

- 启动备份恢复任务

  在任务列表指定任务的“操作”列，选择“更多 > 即时备份”或单击“执行”，启动处于准备或失败状态的备份恢复任务。已成功执行过的恢复任务不能重新运行。

- 停止备份恢复任务

  在任务列表指定任务的“操作”列，选择“更多 > 停止”或单击“停止”，停止处于运行状态的备份恢复任务。停止成功后，该任务的“任务状态”变为“已停止”。

- 删除备份恢复任务

  在任务列表指定任务的“操作”列，选择“更多 > 删除”或单击“删除”，删除备份恢复任务。删除任务后备份的数据默认会保留。

- 挂起备份任务

  在任务列表指定任务的“操作”列，选择“更多 > 挂起”，挂起备份任务。仅支持周期备份的任务，挂起后周期备份任务不再自动执行。挂起正在执行的备份任务时，该任务会停止运行。需要解锁时，选择“更多 > 重新执行”。

10.12 安全管理

10.12.1 安全概述

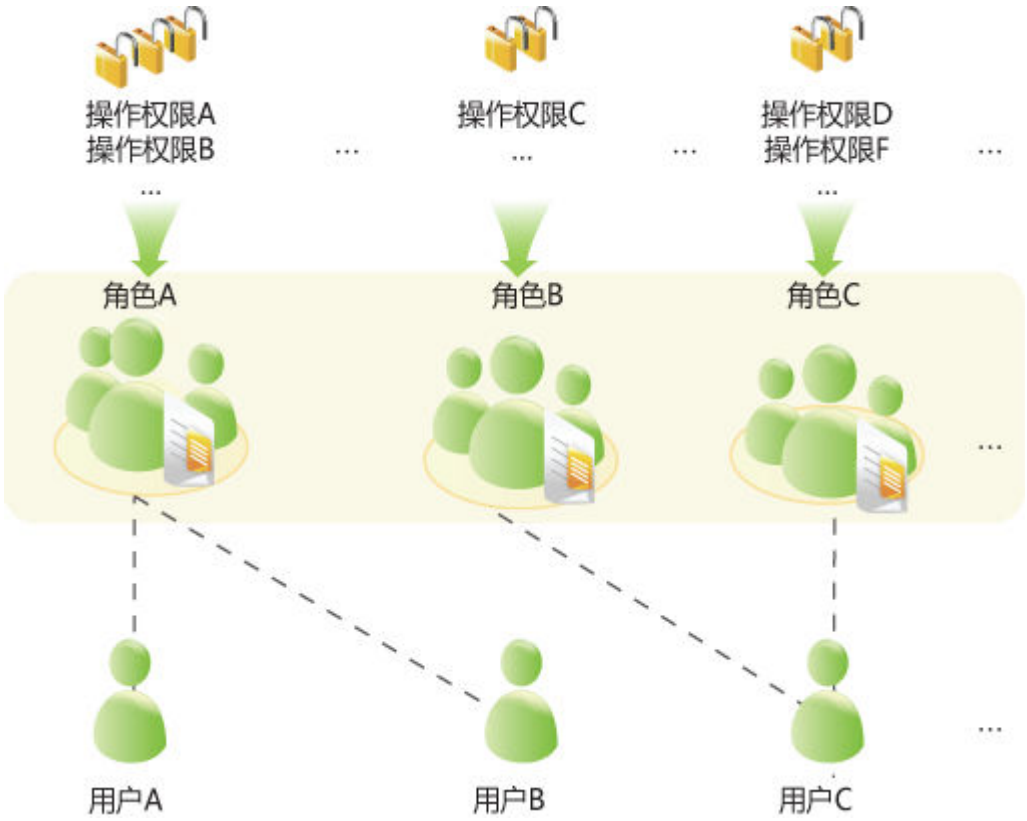
10.12.1.1 权限模型

基于角色的权限控制

FusionInsight通过采用RBAC（role-based access control，基于角色的权限控制）方式对大数据系统进行权限管理，将系统中各组件零散的权限管理功能集中呈现和管理，对普通用户屏蔽掉了内部的权限管理细节，对管理员简化了权限管理的操作方法，提升权限管理的易用性和用户体验。

FusionInsight权限模型由“用户－用户组－角色－权限”四类对象构成。

图 10-21 权限模型



- **权限**

  由组件侧定义，允许访问组件某个资源的能力。不同组件针对自己的资源，有不同的权限。

  例如：

  – HDFS针对文件资源权限，有读、写、执行等权限。

  – HBase针对表资源权限，有创建、读、写等权限。

- **角色**

  组件权限的一个集合，一个角色可以包含多个组件的多个权限，不同的角色也可以拥有同一个组件的同一个资源的权限。

- **用户组**

  用户的集合，当用户组关联某个或者多个角色后，该用户组内的用户就将拥有这些角色所定义的组件权限。

  不同用户组可以关联同一个角色，一个用户组也可以不关联任何角色，该用户组原则上将不具有任何组件资源的权限。

  说明

  部分组件针对特定的默认用户组，系统默认赋予了部分权限。

- **用户**

  系统的访问者，每个用户的权限由该用户关联的用户组和角色所对应的权限构成，用户需要加入用户组或者关联角色来获得对应的权限。

基于策略的权限控制

Ranger组件通过PBAC（policy-based access control，基于策略的权限控制）方式进行权限管理，可对HDFS、Hive、HBase等组件进行更加细粒度的数据访问控制。

说明

组件同时只支持一种权限控制机制，当组件启用Ranger权限控制策略后，通过FusionInsight Manager创建的角色中关于该组件的权限将失效（HDFS与Yarn的组件ACL规则仍将生效），用户需通过Ranger管理界面添加策略进行资源的赋权。

Ranger的权限模型由多条权限策略组成，权限策略主要由以下几方面组成：

- 资源

  组件所提供的可由用户访问的对象，例如HDFS的文件或文件夹、Yarn中的队列、Hive中的数据库/表/列等。

- 用户

  系统的访问者，每个用户的权限由该用户关联的策略来获得。LDAP中的用户、用户组、角色信息会周期性的同步至Ranger。

- 权限

  策略中针对资源可配置各种访问条件，例如文件的读写，具体可以配置允许条件、拒绝条件以及例外条件等。

10.12.1.2 权限机制

FusionInsight采用LDAP存储用户和用户组的数据；角色的定义信息保存在关系数据库中，角色和权限的对应关系则保存在组件侧。

FusionInsight使用Kerberos进行统一认证。

用户权限校验流程大致如下：

1. 客户端（用户终端或FusionInsight组件服务）调用FusionInsight认证接口。
2. FusionInsight使用登录用户名和密码，到Kerberos进行认证。
3. 如果认证成功，客户端会发起访问服务端（FusionInsight组件服务）的请求。
4. 服务端会根据登录的用户，找到其属于的用户组和角色。
5. 服务端获得用户组拥有的所有权限和角色拥有的所有权限的并集。
6. 服务端判断客户端是否有权限访问其请求的资源。

示例场景（RBAC）：

HDFS中有三个文件fileA、fileB、fileC。

- 定义角色roleA对fileA有读和写权限，角色roleB对fileB有读权限。
- 定义groupA属于roleA；groupB属于roleB。
- 定义userA属于groupA和roleB，userB属于GroupB。

当userA登录成功并访问HDFS时：

1. HDFS获得useA属于的所有角色（roleB）。
2. HDFS同时还会获得userA属于的所有用户组所属于的角色（roleA）。
3. 此时，userA拥有roleA和roleB对应权限的并集。
4. 因此对于fileA，则userA有读写权限；对fileB，有读权限；对于fileC，无任何权限。

同理userB登录后：

1. userB只拥有roleB对应的权限。
2. 对于fileA，则userB无权限；对fileB，有读权限；对于fileC，无任何权限。

10.12.1.3 认证策略

大数据平台用户需要对用户进行身份认证，防止不合法用户访问集群。安全模式或者普通模式的集群均提供认证能力。

安全模式

安全模式的集群统一使用Kerberos认证协议进行安全认证。Kerberos协议支持客户端与服务端进行相互认证，提高了安全性，可有效消除使用网络发送用户凭据进行模拟认证的安全风险。集群中由KrbServer服务提供Kerberos认证支持。

Kerberos用户对象

Kerberos协议中，每个用户对象即一个principal。一个完整的用户对象包含两个部分信息：用户名和域名。在运维管理或应用开发的场景中，需要在客户端认证用户身份后才能连接到集群服务端。系统操作运维与业务场景中主要使用的用户分为“人机”用户和“机机”用户。二者主要区别在于“机机”用户密码由系统随机生成。

Kerberos认证

Kerberos认证支持两种方式：密码认证及keytab认证。认证有效时间默认为24小时。

- 密码认证：通过输入用户正确的密码完成身份认证。主要在运维管理场景中使用“人机”用户进行认证，命令为**kinit** *用户名*。

- keytab认证：keytab文件包含了用户principal和用户凭据的加密信息。使用keytab文件认证时，系统自动使用加密的凭据信息进行认证无需输入用户密码。主要在组件应用开发场景中使用“机机”用户进行认证。keytab文件也支持在**kinit**命令中使用。

普通模式

普通模式的集群不同组件使用原生开源的认证机制，不支持**kinit**认证命令。FusionInsight Manager（含DBService、KrbServer和LdapServer）使用的认证方式为用户名密码方式。组件使用的认证机制如表10-80所示。

表 10-80 组件认证方式一览表

服务	认证方式
ClickHouse	simple认证
Flume	无认证
HBase	• WebUI：无认证 • 客户端：simple认证
HDFS	• WebUI：无认证 • 客户端：simple认证
Hive	simple认证
Hue	用户名密码认证
Kafka	无认证
Loader	• WebUI：用户名密码认证 • 客户端：无认证
Mapreduce	• WebUI：无认证 • 客户端：无认证
Oozie	• WebUI：用户名密码认证 • 客户端：simple认证
Spark2x	• WebUI：无认证 • 客户端：simple认证
Storm	无认证
Yarn	• WebUI：无认证 • 客户端：simple认证
ZooKeeper	simple认证

认证方式解释如下：

- “simple认证”：在客户端连接服务端的过程中，默认以客户端执行用户（例如操作系统用户“root”或“omm”）自动进行认证，管理员或业务用户不显式感知认证，不需要**kinit**完成认证过程。
- “用户名密码认证”：使用集群中“人机”用户的用户名与密码进行认证。
- “无认证”：默认任意的用户都可以访问服务端。

10.12.1.4 鉴权策略

安全模式

大数据平台用户完成身份认证后，系统还需要根据实际权限管理配置，选择是否对用户进行鉴权，确保系统用户拥有资源的有限或全部权限。如果系统用户权限不足，需要由系统管理员为用户授予各个组件对应的权限后，才能访问资源。安全模式或者普通模式集群均提供鉴权能力，组件的具体权限项在两种模式中相同。

新安装的安全模式集群默认即安装了Ranger服务并启用了Ranger鉴权，用户可以通过组件的权限插件对组件资源的访问设置细粒度的安全访问策略。若不需使用Ranger进行鉴权，管理员可在服务页面手动停用Ranger鉴权，停用Ranger鉴权后，访问组件资源的时系统将继续基于FusionInsight Manager的角色模型进行权限控制。

安全模式集群中，支持使用Ranger鉴权的组件包括：HDFS、Yarn、Kafka、Hive、HBase、Storm、Spark2x、Impala。

从历史版本升级的集群，用户访问组件资源时默认不使用Ranger鉴权，管理员可在安装了Ranger服务后手动启用Ranger鉴权。

安全版本的集群所有组件默认统一对及访问进行鉴权，不支持关闭鉴权功能。

普通模式

普通模式的集群不同组件使用各自原生开源的鉴权行为，详细鉴权机制如**表10-81**所示。

在安装了Ranger服务的普通模式集群中，Ranger可以支持基于OS用户进行组件资源的权限控制，支持启用Ranger鉴权的组件包括：HBase、HDFS、Hive、Spark2x、Yarn。

表 10-81 普通模式组件鉴权一览表

服务	是否鉴权	是否支持开关鉴权
ClickHouse	鉴权	不支持修改
Flume	无鉴权	不支持修改
HBase	无鉴权	支持修改
HDFS	鉴权	支持修改
Hive	无鉴权	不支持修改
Hue	无鉴权	不支持修改
Kafka	无鉴权	不支持修改

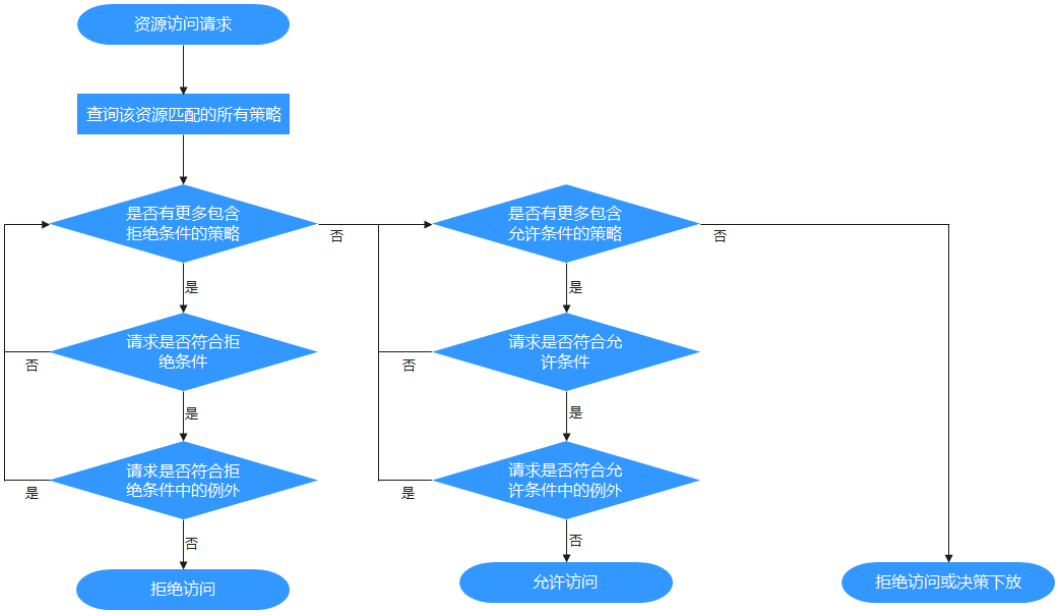
服务	是否鉴权	是否支持开关鉴权
Loader	无鉴权	不支持修改
Mapreduce	无鉴权	不支持修改
Oozie	鉴权	不支持修改
Spark2x	无鉴权	不支持修改
Storm	无鉴权	不支持修改
Yarn	无鉴权	支持修改
ZooKeeper	鉴权	支持修改

Ranger 权限策略条件判断优先级

配置资源的权限策略时，可配置针对该资源的允许条件（Allow Conditions）、允许例外条件（Exclude from Allow Conditions）、拒绝条件（Deny Conditions）以及拒绝例外条件（Exclude from Deny Conditions），以满足不同场景下的例外需求。

不同条件的优先级由高到低为：拒绝例外条件 > 拒绝条件 > 允许例外条件 > 允许条件。

系统判断流程可参考下图所示，如果组件资源请求未匹配到Ranger中的权限策略，系统默认将拒绝访问。但是对于HDFS和Yarn，系统会将决策下放给组件自身的访问控制层继续进行判断。



例如要将一个文件夹FileA的读写权限授权给用户组groupA，但是该用户组内某个用户UserA除外，这时可以增加一个允许条件及一个例外条件即可实现。

10.12.1.5 用户帐号一览表

用户分类

MRS集群提供以下3类用户，请系统管理员定期修改密码，不建议使用默认密码。

说明

本章节介绍MRS集群内的相关默认用户信息。

用户类型	使用说明
系统用户	• 通过FusionInsight Manager创建，是系统操作运维与业务场景中主要使用的用户，包含两种类型： – “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。创建此类型用户时需要参考**创建用户**设置“密码”和“确认密码”。 – “机机”用户：用于系统应用开发的场景。 • 用于OMS系统进程运行的用户。
系统内部用户	集群提供的用于Kerberos认证、进程通信、保存用户组信息和关联用户权限的内部用户。系统内部用户不建议在操作与维护的场景下使用。请通过**admin**用户操作，或联系系统管理员根据业务需要创建新用户。
数据库用户	• 用于OMS数据库管理和数据访问的用户。 • 用于业务组件（Hue、Hive、Loader、Oozie、Ranger和DBService）数据库的用户。

系统用户

说明

- 需要使用操作系统中**root**用户，所有节点**root**用户需设置为统一的密码。
- 需要使用操作系统中**ldap**用户，此帐号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作系统管理员维护。

类别	用户名称	初始密码	描述	密码修改方法
系统管理员	admin	用户创建集群时自定义	FusionInsight Manager的管理员。 **说明** **admin**用户默认不具备其他组件的管理权限，例如访问组件原生界面时，需要使用具备对应组件管理权限的用户才可以访问到完整内容。	请参见**修改admin密码**。

类别	用户名称	初始密码	描述	密码修改方法
节点操作系统用户	ommdba	随机密码	创建系统数据库的用户。在管理节点生成，属于操作系统用户，无需设置为统一的密码。该用户不能用于远程登录。	请参见**修改操作系统用户密码**。
	omm	Bigdata123@	系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。	

系统内部用户

类别	默认用户	初始密码	描述	密码修改方法
Kerberos管理员	kadmin/admin	Admin@123	用于增加、删除、修改及查询Kerberos上的用户帐号。	请参见**修改Kerberos管理员密码**。
OMS Kerberos管理员	kadmin/admin	Admin@123	用于增加、删除、修改及查询OMS Kerberos上的用户帐号。	请参见**修改OMS Kerberos管理员密码**。
LDAP管理员	cn=root,dc=hadoop,dc=com	LdapChangeMe@123	用于增加、删除、修改及查询LDAP用户帐号信息。	请参见**修改LDAP管理员和LDAP用户密码（含OMS LDAP）**。
OMS LDAP管理员	cn=root,dc=hadoop,dc=com	LdapChangeMe@123	用于增加、删除、修改及查询OMS LDAP用户帐号信息。	
LDAP用户	cn=pg_search_dn,ou=Users,dc=hadoop,dc=com	系统随机生成	用于查询LDAP中存储的用户和用户组信息。	
OMS LDAP用户	cn=pg_search_dn,ou=Users,dc=hadoop,dc=com	系统随机生成	用于查询OMS LDAP中存储的用户和用户组信息。	
LDAP管理帐户	cn=krbkdc,ou=Users,dc=hadoop,dc=com	LdapChangeMe@123	用于查询Kerberos组件认证帐户信息。	请参见**修改LDAP管理帐户密码**。

类别	默认用户	初始密码	描述	密码修改方法
	cn=krbadmin,ou=Users,dc=hadoop,dc=com	LdapChangeMe@123	用于增加、删除、修改及查询Kerberos组件认证帐户信息。	
组件运行用户	hdfs	Hdfs@123	HDFS系统管理员，用户权限： 1. 文件系统操作权限： • 查看、修改、创建文件 • 查看、创建目录 • 查看、修改文件属组 • 查看、设置用户磁盘配额 2. HDFS管理操作权限： • 查看webUI页面状态 • 查看、设置HDFS主备状态 • 进入、退出HDFS安全模式 • 检查HDFS文件系统 3. 登录FTP服务	请参见**修改组件运行用户密码**。

类别	默认用户	初始密码	描述	密码修改方法
	hbase	Hbase@123	HBase，HBase1~4系统管理员，用户权限： ● 集群管理权限：表的Enable、Disable操作，触发MajorCompact，ACL操作 ● 授权或回收权限，集群关闭等操作相关的权限 ● 表管理权限：建表、修改表、删除表等操作权限 ● 数据管理权限：表级别、列族级别以及列级别的数据读写权限 ● 登录HMaster WebUI界面 ● 登录FTP服务	
	mapred	Mapred@123	MapReduce系统管理员，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 登录FTP服务 ● 登录Yarn WebUI界面	
	zookeeper	ZooKeeper@123	ZooKeeper系统管理员，用户权限： ● 对Zookeeper上所有节点的增删改查权限 ● 对Zookeeper上所有节点的配额修改查询权限	

类别	默认用户	初始密码	描述	密码修改方法
	rangeradmin	Rangeradmin@123	Ranger的系统管理权限，用户权限。 ● Ranger Web UI的管理权限 ● 使用Ranger鉴权的各组件管理权限	
	rangerauditor	Rangerauditor@123	Ranger系统的默认审计用户。	
	hive	Hive@123	Hive系统管理员，用户权限： 1. Hive管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hive1	Hive1@123	Hive1系统管理员，用户权限： 1. Hive1管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hive2	Hive2@123	Hive2系统管理员，用户权限： 1. Hive2管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hive3	Hive3@123	Hive3系统管理员，用户权限： 1. Hive3管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hive4	Hive4@123	Hive4系统管理员，用户权限： 1. Hive4管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	
	kafka	Kafka@123	Kafka的系统管理员，用户权限: ● Topic的创建、删除、生产、消费、配置修改。 ● Cluster的元数据控制、配置修改、副本迁移、leader选举、acl管理。 ● ConsumerGroup Offset的提交、查询、删除。 ● DelegationToken的查询。 ● Transaction的查询、提交。	

类别	默认用户	初始密码	描述	密码修改方法
	storm	Admin@123	storm的系统管理员。 用户权限：storm任务提交。	
	rangerusersync	系统随机生成	用于同步用户及用户组的内部用户。	
	rangertagsync	系统随机生成	用于同步标签的内部用户。	
	oms/manager	系统随机生成	用于Controller和NodeAgent认证的用户，拥有“supergroup”组权限。	
	backup/manager	系统随机生成	用于运行备份恢复任务的用户，拥有“supergroup”、“wheel”和“ficommon”组权限。配置跨系统互信后拥有访问互信系统HDFS、HBase、Hive、ZooKeeper数据的权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hdfs/hadoop.<*系统域名*>	系统随机生成	HDFS系统启动用户，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看WebUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统 3. 登录FTP服务	
	mapred/hadoop.<*系统域名*>	系统随机生成	MapReduce系统启动用户，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 登录FTP服务 ● 登录Yarn WebUI界面	
	mr_zk/hadoop.<*系统域名*>	系统随机生成	用于MapReduce访问ZooKeeper。	
	hbase/hadoop.<*系统域名*>	系统随机生成	HBase系统启动过程用于内部组件之间认证的用户。	
	hbase/zkclient.<*系统域名*>	系统随机生成	安全集群下，HBase做ZooKeeper认证时使用的用户。	

类别	默认用户	初始密码	描述	密码修改方法
	thrift/hadoop.<*系统域名*>	系统随机生成	ThriftServer系统启动用户。	
	thrift/*<hostname>*	系统随机生成	ThriftServer系统访问HBase的用户，拥有HBase所有NameSpace和表的读、写、执行、创建和管理的权限。*<hostname>*表示集群中安装ThriftServer节点的主机名。	
	hive/hadoop.<*系统域名*>	系统随机生成	Hive系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	

类别	默认用户	初始密码	描述	密码修改方法
	hive1/hadoop.<*系统域名*>	系统随机生成	Hive1系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive1管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	

类别	默认用户	初始密码	描述	密码修改方法
	hive2/hadoop.<*系统域名*>	系统随机生成	Hive2系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive2管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	

类别	默认用户	初始密码	描述	密码修改方法
	hive3/hadoop.<*系统域名*>	系统随机生成	Hive3系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive3管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	

类别	默认用户	初始密码	描述	密码修改方法
	hive4/hadoop.<*系统域名*>	系统随机生成	Hive4系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive4管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	
	loader/hadoop.<*系统域名*>	系统随机生成	Loader系统启动与Kerberos认证用户。	
	HTTP/<*hostname*>	系统随机生成	用于连接各组件的HTTP接口，<*hostname*>表示集群中节点主机名。	
	hue	系统随机生成	Hue系统启动与Kerberos认证用户，并用于访问HDFS和Hive。	
	flume	系统随机生成	Flume系统启动用户，用于访问HDFS和Kafka，对HDFS目录“/flume”有读写权限。	

类别	默认用户	初始密码	描述	密码修改方法
	flume_server	系统随机生成	Flume系统启动用户，用于访问HDFS和Kafka，对HDFS目录“/flume”有读写权限。	
	spark2x/hadoop.<*系统域名*>	系统随机生成	Spark2x系统管理员用户，用户权限： 1、Spark2x服务启动用户 2、提交Spark2x任务的权限	
	spark_zk/hadoop.<*系统域名*>	系统随机生成	用于Spark2x访问ZooKeeper。	
	spark2x1/hadoop.<*系统域名*>	系统随机生成	Spark2x1系统管理员用户，用户权限： 1. Spark2x1服务启动用户 2. 提交Spark2x任务的权限	
	spark2x2/hadoop.<*系统域名*>	系统随机生成	Spark2x2系统管理员用户，用户权限： 1. Spark2x2服务启动用户 2. 提交Spark2x任务的权限	
	spark2x3/hadoop.<*系统域名*>	系统随机生成	Spark2x3系统管理员用户，用户权限： 1. Spark2x3服务启动用户 2. 提交Spark2x任务的权限	
	spark2x4/hadoop.<*系统域名*>	系统随机生成	Spark2x4系统管理员用户，用户权限： 1. Spark2x4服务启动用户 2. 提交Spark2x任务的权限	
	zookeeper/hadoop.<*系统域名*>	系统随机生成	ZooKeeper系统启动用户。	

类别	默认用户	初始密码	描述	密码修改方法
	zkcli/hadoop.<*系统域名*>	系统随机生成	登录Zookeeper服务器用户。	
	oozie	系统随机生成	Oozie系统启动与Kerberos认证用户。	
	kafka/hadoop.<*系统域名*>	系统随机生成	用于Kafka安全认证。	
	storm/hadoop.<*系统域名*>	系统随机生成	Storm系统启动用户。	
	storm_zk/hadoop.<*系统域名*>	系统随机生成	用于Worker进程访问ZooKeeper。	
	flink/hadoop.<*系统域名*>	系统随机生成	Flink服务的内部用户。	
	check_ker_M	系统随机生成	系统内部测试Kerberos服务功能是否正常的用户。	
	tez	系统随机生成	TezUI系统启动与Kerberos认证用户，并用于访问Yarn。	
	K/M	系统随机生成	Kerberos内部功能用户，不能删除，不支持密码修改，未安装Kerberos服务的节点无法使用内部帐户。	无
	kadmin/changepw	系统随机生成		
	kadmin/history	系统随机生成		
	krbtgt/<*系统域名*>	系统随机生成		
LDAP用户	admin	无	FusionInsight Manager的管理员。 主组为compcommon，不具备组权限，具备Manager_administrator角色的权限。	LDAP用户不支持登录与认证，无密码修改方法。
	backup		主组为compcommon	

类别	默认用户	初始密码	描述	密码修改方法
	backup/manager		主组为compcommon	
	oms		主组为compcommon	
	oms/manager		主组为compcommon	
	clientregister		主组为compcommon	
	zookeeper		主组为hadoop	
	zookeeper/hadoop.<*系统域名*>		主组为hadoop	
	zkcli		主组为hadoop	
	zkcli/hadoop.<*系统域名*>		主组为hadoop	
	flume		主组为hadoop	
	flume_server		主组为hadoop	
	hdfs		主组为hadoop	
	hdfs/hadoop.<*系统域名*>		主组为hadoop	
	mapred		主组为hadoop	
	mapred/hadoop.<*系统域名*>		主组为hadoop	
	mr_zk		主组为hadoop	
	mr_zk/hadoop.<*系统域名*>		主组为hadoop	
	hue		主组为supergroup	
	hive		主组为hive	
	hive/hadoop.<*系统域名*>		主组为hive	
	hive1		主组为hive1	

类别	默认用户	初始密码	描述	密码修改方法
	hive1/hadoop.<*系统域名*>		主组为hive1	
	hive2		主组为hive2	
	hive2/hadoop.<*系统域名*>		主组为hive2	
	hive3		主组为hive3	
	hive3/hadoop.<*系统域名*>		主组为hive3	
	hive4		主组为hive4	
	hive4/hadoop.<*系统域名*>		主组为hive4	
	hbase		主组为hadoop	
	hbase/hadoop.<*系统域名*>		主组为hadoop	
	thrift		主组为hadoop	
	thrift/hadoop.<*系统域名*>		主组为hadoop	
	oozie		主组为hadoop	
	hbase/zkclient.<*系统域名*>		主组为hadoop	
	loader		主组为hadoop	
	loader/hadoop.<*系统域名*>		主组为hadoop	
	spark2x		主组为hadoop	
	spark2x/hadoop.<*系统域名*>		主组为hadoop	
	spark_zk		主组为hadoop	
	spark2x1		主组为hadoop	

类别	默认用户	初始密码	描述	密码修改方法
	spark2x1/hadoop.<*系统域名*>		主组为hadoop	
	spark2x2		主组为hadoop	
	spark2x2/hadoop.<*系统域名*>		主组为hadoop	
	spark2x3		主组为hadoop	
	spark2x3/hadoop.<*系统域名*>		主组为hadoop	
	spark2x4		主组为hadoop	
	spark2x4/hadoop.<*系统域名*>		主组为hadoop	
	kafka		主组为kafkaadmin	
	kafka/hadoop.<*系统域名*>		主组为kafkaadmin	
	storm		主组为stormadmin	
	storm/hadoop.<*系统域名*>		主组为stormadmin	
	storm_zk		主组为storm	
	storm_zk/hadoop.<*系统域名*>		主组为storm	
	kms/hadoop		主组为kmsadmin	
	knox		主组是compcommon	
	executor		主组是compcommon	

说明

用户可登录FusionInsight Manager后，选择“系统 > 权限 > 域和互信”，查看“本端域”参数，即为当前系统域名。上表中系统内部用户的用户名所包含的系统域名所有字母为小写。

例如“本端域”参数为“9427068F-6EFA-4833-B43E-60CB641E5B6C.COM”，则HDFS默认启动用户为“hdfs/hadoop.9427068f-6efa-4833-b43e-60cb641e5b6c.com”。

数据库用户

系统数据库用户包含OMS数据库用户、DBService数据库用户。

类别	默认用户	初始密码	描述	密码修改方法
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作。	请参见**修改OMS数据库管理员密码**。
	omm	ChangeMe@123456	OMS数据库数据访问用户。	请参见**修改OMS数据库访问用户密码**。
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户。	请参见**修改组件数据库用户密码**。
	hive	HiveUser@	Hive连接DBService数据库hivemeta的用户。	
	hive1	HiveUser@	Hive1连接DBService数据库hivemeta1的用户。	
	hive2	HiveUser@	Hive2连接DBService数据库hivemeta2的用户。	
	hive3	HiveUser@	Hive3连接DBService数据库hivemeta3的用户。	
	hive4	HiveUser@	Hive4连接DBService数据库hivemeta4的用户。	
	hive*N*-*N*	HiveUser@	安装多服务时，Hive-*N*连接DBService数据库hive*N*meta的用户。 例如Hive-1服务连接DBService数据库hive1meta的用户为hive11。	

类别	默认用户	初始密码	描述	密码修改方法
	hue	HueUser@123	Hue连接DBService数据库hue的用户。	
	sqoop	SqoopUser@	Loader连接DBService数据库sqoop的用户。	
	sqoop*N*	SqoopUser@	安装多服务时，Loader-*N*连接DBService数据库sqoop*N*的用户。 例如Loader-1服务连接DBService数据库sqoop1的用户为sqoop1。	
	oozie	OozieUser@	Oozie连接DBService数据库oozie的用户。	
	oozie*N*	OozieUser@	安装多服务时，Oozie-*N*连接DBService数据库oozie*N*的用户。 例如Oozie-1服务连接DBService数据库oozie1的用户为oozie1。	
	rangeradmin	Admin12!	Ranger连接DBservice数据库ranger的用户。	

10.12.1.6 默认权限信息一览

角色

默认角色	描述
Manager_administrator	Manager管理员，具有Manager所有权限。 可创建一级租户，可创建、修改新的用户组，指定用户权限，以满足不同用户对系统的管理需求。
Manager_operator	Manager操作员，具有**主页、集群、主机、运维**页签所有权限。
Manager_auditor	Manager审计员，具有**审计**页签的所有权限。 可查看和管理Manager系统审计日志的权限。

默认角色	描述
Manager_viewer	Manager查看员，具有**主页、集群、主机、告警与事件、系统>权限**相关信息的查看权限。
Manager_tenant	Manager租户管理员。 可为当前用户所属于的非叶子租户创建子租户并管理。具有“**运维 > 告警**”页面下“**告警**”、“**事件**”的查看权限。
System_administrator	系统管理员，具有Manager的管理员权限及所有组件服务管理员的权限。
default	为集群default租户创建的默认角色。拥有Yarn组件default队列的管理权限。非首个安装集群的default租户默认角色为“c<*集群ID*>_default”。
Manager_administrator_180	FusionInsight Manager系统管理员组。系统内部角色，仅限组件间内部使用。
Manager_auditor_181	FusionInsight Manager系统审计员组。系统内部角色，仅限组件间内部使用。
Manager_operator_182	FusionInsight Manager系统操作员组。系统内部角色，仅限组件间内部使用。
Manager_viewer_183	FusionInsight Manager系统查看员组。系统内部角色，仅限组件间内部使用。
System_administrator_186	系统管理员组。系统内部角色，仅限组件间内部使用。
Manager_tenant_187	租户系统用户组。系统内部角色，仅限组件间内部使用。
default_1000	为租户创建的用户组。系统内部角色，仅限组件间内部使用。

用户组

类型	默认用户组	描述
集群默认用户组	hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。
	hadoopmanager	将用户加入此用户组，可获得HDFS和Yarn的组件运维管理员权限。对HDFS来说，运维管理员可以访问NameNode WebUI，还能进行手动主备倒换等操作。对Yarn来说，运维管理员可以执行Yarn集群的管理操作，例如访问ResourceManager WebUI，管理NodeManager节点，刷新队列，设置NodeLabel等，但不能提交任务。
	hive	普通用户组。Hive用户必须属于该用户组。

类型	默认用户组	描述
	hive1	普通用户组。Hive1用户必须属于该用户组。
	hive2	普通用户组。Hive2用户必须属于该用户组。
	hive3	普通用户组。Hive3用户必须属于该用户组。
	hive4	普通用户组。Hive4用户必须属于该用户组。
	kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限，才能访问对应Topic。
	kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
	kafkasuperuser	Kafka的Topic读写用户组。添加入本组的用户，拥有所有Topic的读写权限。
	storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。
	stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。
	supergroup	这个用户组内的用户具有HBase，HDFS和Yarn的管理员权限，并且可以使用Hive。
	yarnviewgroup	Yarn任务只读用户组。将用户加入此用户组，可获得Yarn和Mapreduce界面上任务的只读权限。
	check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
	compcommon	系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
操作系统默认用户组	wheel	系统内部运行用户“omm”的主组。
	ficommon	系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

说明

如果当前集群不是在FusionInsight Manager内第一次安装的集群，集群内除Manager以外其他组件对应的默认用户组名称为“c*<集群ID>_默认用户组名*”，例如“c2_hadoop”。

用户

请参见**用户帐号一览表**。

服务相关用户安全参数

- **HDFS**

  参数“dfs.permissions.superusergroup”表示HDFS最高权限管理员组，默认值为“supergroup”。

- **Spark2x以及对应多实例**

  参数“spark.admin.acls”表示Spark2x的管理员列表，列表中成员有权限管理所有Spark任务，若用户未加入此列表则无法管理所有Spark任务。默认值为“admin”。

10.12.1.7 FusionInsight Manager 安全功能

通过FusionInsight Manager的以下模块，可以方便的完成用户权限数据的查看和设置。

- 用户管理：提供用户的增、删、改、查基本功能，提供用户绑定用户组和角色的功能。

  具体请参见**用户管理**。

- 用户组管理：提供用户组的增、删、改、查基本功能，提供用户组绑定角色的功能。

  具体请参见**用户组管理**。

- 角色管理：提供角色的增、删、改、查基本功能，提供角色绑定某个或者多个组件的资源访问权限的功能。

  具体请参见**角色管理**。

- 租户管理：提供租户的增、删、改、查基本功能以及租户与组件资源的绑定关系。FusionInsight为了便于管理，为每个租户都会默认产生一个角色。如果定义租户拥有某些资源的权限，则租户对应的角色就拥有这些资源的权限。

  具体请参见**租户资源**。

10.12.2 帐户管理

10.12.2.1 帐户安全设置

10.12.2.1.1 解锁 LDAP 用户和管理帐户

操作场景

管理员在LDAP用户和管理帐户被锁定时，需要在管理节点解锁集群LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”以及LDAP管理帐户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”和“cn=krbadmin,ou=Users,dc=hadoop,dc=com”。

说明

Ldap用户或管理帐户连续使用错误密码操作Ldap次数大于5次时，会造成LDAP用户或管理帐户被锁定。用户被锁定之后，5分钟后会自动解锁。

操作步骤

步骤1 以**omm**用户登录主管理节点。

步骤2 执行以下命令，切换到目录：

cd ${BIGDATA_HOME}/om-server/om/ldapserver/ldapserver/local/script

步骤3 执行以下命令，解锁LDAP用户或管理帐户：

./ldapserver_unlockUsers.sh *USER_NAME*

其中，*USER_NAME*表示将要解锁的用户名称。

例如，解锁LDAP管理帐户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”的方法如下：

./ldapserver_unlockUsers.sh krbkdc

运行脚本之后，在ROOT_DN_PASSWORD之后输入krbkdc用户密码，显示如下结果，说明解锁成功：

```
Unlock user krbkdc successfully.
```

----结束

10.12.2.1.2 解锁系统内部用户

操作场景

若服务出现异常状态，有可能是系统内部用户被锁定，请及时解锁，否则会影响集群正常运行。系统内部用户列表请参见**用户帐号一览表**。系统内部用户无法使用FusionInsight Manager解锁。

前提条件

根据**用户帐号一览表**获取LDAP管理员“cn=root,dc=hadoop,dc=com”的默认密码。

操作步骤

步骤1 使用以下方法确认系统内部用户是否被锁定：

1. 查询oldap端口：
   a. 登录FusionInsight Manager，选择“系统 > OMS > oldap > 修改配置”。
   b. “Ldap服务监听端口”参数值即为oldap端口。
2. 查询域名方法：
   a. 登录FusionInsight Manager，选择“系统 > 权限 > 域和互信”。
   b. “本端域”参数即为域名。
      例如当前系统域名为“9427068F-6EFA-4833-B43E-60CB641E5B6C.COM”。
3. 在集群内节点上以**omm**用户执行以下命令查询密码认证失败次数：
   ldapsearch -H ldaps://*OMS浮动IP地址*:*OLdap端口* **-LLL -x -D cn=root,dc=hadoop,dc=com -b krbPrincipalName=***系统内部用户名*@*当前域名*,cn=*当前域名*,cn=krbcontainer,dc=hadoop,dc=com -w** *LDAP管理员密码* **-e ppolicy | grep krbLoginFailedCount**

例如，查看oms/manager用户认证失败次数：

ldapsearch -H ldaps://10.5.146.118:21750 -LLL -x -D cn=root,dc=hadoop,dc=com -b krbPrincipalName=oms/manager@9427068F-6EFA-4833-B43E-60CB641E5B6C.COM,cn=9427068F-6EFA-4833-B43E-60CB641E5B6C.COM,cn=krbcontainer,dc=hadoop,dc=com -w *cn=root,dc=hadoop,dc=com用户密码* **-e ppolicy | grep krbLoginFailedCount**

```
krbLoginFailedCount: 5
```

4. 登录FusionInsight Manager，选择“系统 > 权限 > 安全策略 > 密码策略”。
5. 查看“密码连续错误次数”参数值，若小于等于“krbLoginFailedCount”参数值，则用户已被锁定。

说明

查看运行日志，也可以确认系统内部用户是否被锁定。

步骤2 以**omm**用户登录主管理节点，执行以下命令解锁。

sh ${BIGDATA_HOME}/om-server/om/share/om/acs/config/unlockuser.sh --userName *系统内部用户名*

例如，**sh ${BIGDATA_HOME}/om-server/om/share/om/acs/config/unlockuser.sh --userName oms/manager**

----结束

10.12.2.1.3 修改集群组件鉴权配置开关

操作场景

集群部署为安全模式或者普通模式时，HDFS和ZooKeeper默认会对访问服务的用户进行鉴权，没有权限的用户无法访问HDFS和ZooKeeper中的资源。集群部署为普通模式时，HBase和Yarn默认不会对访问用户进行鉴权，所有用户可以访问HBase和Yarn中的资源。

管理员可以根据业务实际需要，在普通模式集群中配置开启HBase和Yarn鉴权，或关闭HDFS和ZooKeeper鉴权。

对系统的影响

修改开关后服务的配置将过期，需要重启对应的服务使配置生效。

开启 HBase 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“hbase.coprocessor.region.classes”、“hbase.coprocessor.master.classes”和“hbase.coprocessor.regionserver.classes”。

将协处理器参数“org.apache.hadoop.hbase.security.access.AccessController”添加到以上参数原有参数值末尾，使用英文逗号与原有协处理器分隔。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

关闭 HBase 鉴权

说明

关闭HBase鉴权后，原有的权限数据会继续保留。如果需要删除权限信息，请在关闭鉴权后，进入hbase shell删除表hbase:acl。

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“hbase.coprocessor.region.classes”、“hbase.coprocessor.master.classes”和“hbase.coprocessor.regionserver.classes”。

将协处理器参数“org.apache.hadoop.hbase.security.access.AccessController”去除。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

关闭 HDFS 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“dfs.namenode.acls.enabled”和“dfs.permissions.enabled”。

- “dfs.namenode.acls.enabled”表示是否启用HDFS ACL，默认为“true”启用ACL，请修改为“false”。
- “dfs.permissions.enabled”表示是否为HDFS启用权限检查，默认为“true”启用权限检查，请修改为“false”。修改后HDFS中的目录和文件的属主、属组以及权限信息保持不变。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

开启 Yarn 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“yarn.acl.enable”。

“yarn.acl.enable”表示是否为Yarn启用权限检查。

- 普通模式下默认为“false”不启用权限检查，如果要启用，请修改为“true”。
- 安全模式下默认为“true”，表示开启鉴权。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

关闭 ZooKeeper 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“skipACL”。

“skipACL”表示是否跳过ZooKeeper权限检查，默认为“no”启用权限检查，请修改为“yes”。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

10.12.2.1.4 使用普通模式集群用户在非集群节点登录

操作场景

集群安装为普通模式时，各组件客户端不支持安全认证且无法使用**kinit**命令，所以集群外的节点默认无法使用集群中的用户，可能导致在这些节点访问某个组件服务端时用户鉴权失败。

如果需要在集群外节点以组件用户身份访问集群资源，管理员需为集群外节点设置同名用户可通过SSH协议登录节点的功能，并以登录操作系统用户身份连接集群各组件服务端。

前提条件

- 集群外的节点需要与集群的业务平面是连通的。
- 集群的KrbServer服务运行状态正常。
- 获取集群外的节点**root**用户密码。
- 集群已规划并添加“人机”用户，并获取认证凭据文件。请参见**创建用户**和**导出认证凭据文件**。

操作步骤

步骤1 以**root**用户登录到需要添加用户的节点。

步骤2 执行以下命令：

rpm -qa | grep pam和**rpm -qa| grep krb5-client**

界面一共显示以下rpm包：

```
pam_krb5-32bit-2.3.1-47.12.1
pam-modules-32bit-11-1.22.1
yast2-pam-2.17.3-0.5.211
pam-32bit-1.1.5-0.10.17
pam_mount-32bit-0.47-13.16.1
pam-config-0.79-2.5.58
pam_krb5-2.3.1-47.12.1
pam-doc-1.1.5-0.10.17
pam-modules-11-1.22.1
pam_mount-0.47-13.16.1
pam_ldap-184-147.20
pam-1.1.5-0.10.17
krb5-client-1.6.3
```

步骤3 检查操作系统实际是否已安装清单中的rpm包?

- 是，执行**步骤5**。
- 否，执行**步骤4**。

步骤4 从操作系统镜像中获取缺少的rpm包，并上传文件到当前目录，然后执行以下命令安装rpm包：

rpm -ivh *.rpm

说明

安装的RPM包可能带来安全风险，请用户对操作系统进行加固时考虑安装这些RPM包所带来的风险。

安装完成后执行**步骤5**。

步骤5 执行以下命令，配置pam使用Kerberos认证。

pam-config --add --krb5

说明

如果需要在非集群节点取消Kerberos认证与系统用户登录，以“root”用户执行**pam-config --delete --krb5**命令。

步骤6 解压认证凭据文件得到“krb5.conf”，并使用WinSCP将此配置文件上传到集群外节点的“/etc”目录，执行以下命令设置权限使其他用户可以访问，例如“604”：

chmod 604 /etc/krb5.conf

步骤7 以**root**用户继续在连接会话中执行以下命令为“人机”用户添加对应的操作系统用户，并指定用户主组为“root”。

此操作系统用户密码与在Manager创建“人机”用户时设置的初始密码相同。

useradd *用户名* **-m -d /home/admin_test -g root -s /bin/bash**

例如，“人机”用户名为“admin_test”，执行以下命令：

useradd admin_test -m -d /home/admin_test -g root -s /bin/bash

说明

第一次使用新添加的操作系统用户通过SSH协议登录节点时，首次输入用户密码系统提示密码过期，第二次输入用户密码后系统提示修改密码。请输入一个同时满足节点操作系统及集群密码复杂度的新密码。

----结束

10.12.2.2 修改系统用户密码

10.12.2.2.1 修改 admin 密码

操作场景

“admin”是FusionInsight Manager的系统管理员帐号，建议用户通过FusionInsight Manager定期修改密码，提高系统安全性。

操作步骤

步骤1 登录FusionInsight Manager。

需使用“admin”登录。

步骤2 移动鼠标到界面右上角的“Hello, admin”。

在弹出菜单中单击“修改密码”。

步骤3 分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

默认密码复杂度要求：

- 密码字符长度为8～64位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为~`!?,.;-_'(){}[]/<>@#$%^&*+|\=。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**密码策略配置**中“重复使用规则”的值。

----结束

10.12.2.2.2 修改操作系统用户密码

操作场景

安装FusionInsight Manager时系统自动在集群每个节点上创建用户“omm”和“ommdba”，建议管理员定期修改集群节点操作系统用户“omm”、“ommdba”的登录密码，以提升系统运维安全性。

各节点“omm”、“ommdba”无需设置为统一的密码。

前提条件

- 获取待修改密码“omm”、“ommdba”用户对应节点的IP地址。
- 修改**omm**和**ommdba**用户需要获取**root**用户密码。

修改操作系统用户密码

步骤1 以**root**登录待修改密码节点。

步骤2 执行如下命令，修改用户密码。

passwd *ommdba*

Red Hat系统显示：

```
Changing password for user ommdba.
New password:
```

步骤3 输入用户的新密码。操作系统的密码修改策略由用户实际使用的操作系统类型决定。

```
Retype New Password:
Password changed.
```

----结束

10.12.2.3 修改系统内部用户密码

10.12.2.3.1 修改 Kerberos 管理员密码

操作场景

管理员应定期修改Kerberos管理员“kadmin”的密码，以提升系统运维安全性。

修改此用户密码将同步修改OMS Kerberos管理员密码。

前提条件

已在集群内的任一节点安装了客户端，并获取此节点IP地址。

操作步骤

步骤1 以**root**用户通过节点IP地址登录安装了客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient”。

cd /opt/hadoopclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，修改kadmin/admin密码。此操作对所有服务器生效。

kpasswd kadmin/admin

默认密码复杂度要求：

- 密码字符长度最小为8位。

- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为~`!?,.;-_'(){}[]/<>@#$%^&*+|\=。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码，例如Admin@12345。
- 不可与最近N次使用过的密码相同，N为**密码策略配置**中“重复使用规则”的值。

----结束

10.12.2.3.2 修改 OMS Kerberos 管理员密码

操作场景

管理员应定期修改OMS Kerberos管理员“kadmin”的密码，以提升系统运维安全性。

修改此用户密码将同步修改Kerberos管理员密码。

操作步骤

步骤1 以**omm**用户登录任意管理节点。

步骤2 执行以下命令，切换到目录。

cd ${BIGDATA_HOME}/om-server/om/meta-0.0.1-SNAPSHOT/kerberos/scripts

步骤3 执行以下命令，配置环境变量。

source component_env

步骤4 执行以下命令，修改kadmin/admin密码。此操作对所有服务器生效。

kpasswd kadmin/admin

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!?,.;-_'(){}[]/<>@#$%^&*+|\=中的4种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码，例如Admin@12345。
- 不可与最近N次使用过的密码相同，N为**密码策略配置**中“重复使用规则”的值。

----结束

10.12.2.3.3 修改 LDAP 管理员和 LDAP 用户密码（含 OMS LDAP）

操作场景

建议管理员定期修改集群的LDAP管理员用户“cn=root,dc=hadoop,dc=com”和LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码，以提升系统运维安全性。

修改上述用户密码将同步修改OMS LDAP管理员或用户密码。

说明

旧版本集群升级到新版本后，LDAP管理员密码将继承旧集群的密码策略，为保证系统安全，建议集群升级后及时修改密码。

对系统的影响

- 修改LdapServer服务的用户密码为高危操作，需要重启KrbServer和LdapServer服务。重启KrbServer可能会导致集群中的节点短时间内出现执行**id**命令查询不到用户的现象，请谨慎执行。
- 修改LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码后，可能会导致该用户在组件LDAP上被锁定。因此，建议修改密码后对该用户进行解锁，解锁方法请参见**解锁LDAP用户和管理帐户**章节。

前提条件

修改LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码前需先确认该用户没有被锁定，在集群主管理节点上执行如下命令：

说明

oldap端口查询方法：

1. 登录FusionInsight Manager，选择“系统 > OMS > oldap > 修改配置”；
2. “Ldap服务监听端口”参数值即为oldap端口。

ldapsearch -H ldaps://*OMS浮动地址:OLdap端口* -LLL -x -D cn=pg_search_dn,ou=Users,dc=hadoop,dc=com -W -b cn=pg_search_dn,ou=Users,dc=hadoop,dc=com -e ppolicy

输入LDAP用户pg_search_dn的密码，出现如下提示表示该用户被锁定，则需要解锁用户，具体请参见**解锁LDAP用户和管理帐户**。

说明

LDAP用户pg_search_dn的密码为系统随机生成，具体可在主节点的“/etc/sssd/sssd.conf”或“/etc/ldap.conf”文件中获取。

```
ldap_bind: Invalid credentials (49); Account locked
```

操作步骤

步骤1 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > LdapServer”。

步骤2 选择 “更多 > 修改数据库密码”，在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤3 在“修改密码”对话框的“用户信息”中选择需要修改密码的用户名。

步骤4 在“旧密码”输入旧密码，“新密码”和“确认密码”输入新密码。

默认密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符`~!@#$%^&*()-_=+|[{}];,<.>/?中的3种类型字符。

- 不可和用户名相同或用户名的倒序字符相同。
- 不可与当前密码相同。

步骤5 勾选“我已阅读此信息并了解其影响”，单击“确定”确认修改并重启服务。

----结束

10.12.2.3.4 修改 LDAP 管理帐户密码

操作场景

建议管理员定期修改集群LDAP管理帐户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”和“cn=krbadmin,ou=Users,dc=hadoop,dc=com”的密码，以提升系统运维安全性。

对系统的影响

- 修改密码后需要重启KrbServer服务。
- 修改密码后需要确认LDAP管理帐户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”和“cn=krbadmin,ou=Users,dc=hadoop,dc=com”是否被锁定，在集群主管理节点上执行如果下命令查看krbkdc是否被锁定（krbadmin用户方法类似）：

  说明

  oldap端口查询方法：

  1. 登录FusionInsight Manager，选择“系统 > OMS > oldap > 修改配置”；
  2. “Ldap服务监听端口”参数值即为oldap端口。

  ldapsearch -H ldaps://*OMS_FLOAT_ IP地址:OLdap端口* -LLL -x -D cn=krbkdc,ou=Users,dc=hadoop,dc=com -W -b cn=krbkdc,ou=Users,dc=hadoop,dc=com -e ppolicy

  输入LDAP管理帐户krbkdc的密码，出现如下提示表示该用户被锁定，则需要解锁用户，具体请参见**解锁LDAP用户和管理帐户**。

  ```
  ldap_bind: Invalid credentials (49); Account locked
  ```

前提条件

已确认主管理节点IP地址。

操作步骤

步骤1 以**omm**用户通过管理节点IP登录主管理节点。

步骤2 执行以下命令，切换到目录。

cd ${BIGDATA_HOME}/om-server/om/meta-0.0.1-SNAPSHOT/kerberos/scripts

步骤3 执行以下命令，修改LDAP管理帐户密码。

./okerberos_modpwd.sh

输入旧密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为16～32位。

- 至少需要包含大写字母、小写字母、数字、特殊字符`~!@#$%^&*()-_=+|[{}];,<.>/?中的3种类型字符。
- 不可与当前密码相同。

显示如下结果，说明修改成功：

```
Modify kerberos server password successfully.
```

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > KrbServer > 更多 > 重启服务”。

验证用户身份后不勾选“同时重启上层服务”，单击“确定”重启KrbServer服务。

----结束

10.12.2.3.5 修改组件运行用户密码

操作场景

建议管理员定期修改集群内组件运行用户的密码，以提升系统运维安全性。

组件运行用户，根据初始密码是否是系统随机生成，可分为两类：

- 密码随机生成的，用户类型为“机机”用户。
- 密码不是随机生成的，用户类型为“人机”用户。

对系统的影响

初始密码为系统随机生成的组件运行用户，在修改密码后需要重启集群，重启期间会造成业务暂时中断。

前提条件

已在集群内的任一节点安装了客户端，并获取此节点IP地址。

操作步骤

步骤1 以客户端安装用户，登录安装了客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，输入kadmin/admin用户密码后进入kadmin控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“Admin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤5 执行以下命令，修改系统内部组件运行用户密码。此操作对所有服务器生效。

cpw *系统内部用户名*

例如：**cpw oms/manager**

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为~`!?,.;-_'(){}[]/<>@#$%^&*+|\=。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码，例如Admin@12345。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。此策略只影响“人机”用户。

说明

执行如下命令，可以查看用户的信息。

getprinc *系统内部用户名*

例如：**getprinc oms/manager**

步骤6 确认修改密码的用户，用户类型是哪种?

- 用户类型为“机机”用户，执行**步骤7**。
- 用户类型为“人机”用户，密码修改完成，任务结束。

步骤7 登录FusionInsight Manager。

步骤8 选择“集群 > *待操作的集群名称* > 更多 > 重启”。

步骤9 在弹出窗口中，输入当前登录的用户密码确认身份，单击“确定”。

步骤10 在确认重启的对话框中，单击“确定”。

步骤11 等待界面提示重启成功。

----结束

10.12.2.4 修改默认数据库用户密码

10.12.2.4.1 修改 OMS 数据库管理员密码

操作场景

建议管理员定期修改OMS数据库管理员的密码，以提升系统运维安全性。

操作步骤

步骤1 以**root**用户登录主管理节点。

说明

ommdba用户密码不支持在备管理节点修改，否则集群无法正常工作。只需在主管理节点执行修改操作，无需在备管理节点操作。

步骤2 执行以下命令，切换用户。

su - omm

步骤3 执行以下命令，切换目录。

cd $OMS_RUN_PATH/tools

步骤4 执行以下命令，修改**ommdba**用户密码。

mod_db_passwd ommdba

步骤5 输入**ommdba**的原密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

显示如下结果，说明修改成功：

```
Congratulations, update [ommdba] password successfully.
```

----结束

10.12.2.4.2 修改 OMS 数据库访问用户密码

操作场景

建议管理员定期修改OMS数据库访问用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启OMS服务，服务在重启时系统无法访问。

操作步骤

步骤1 在FusionInsight Manager选择“系统 > OMS > gaussDB > 修改密码”。

步骤2 在omm用户所在行，单击“操作”列下的“修改密码”。

步骤3 在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤4 根据界面信息，输入新旧密码。

密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

步骤5 单击“确定”，等待界面提示操作成功。

步骤6 在**omm**用户所在行，单击“操作”列下的“重启OMS服务”。

步骤7 在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤8 在确定重启的对话框中，单击“确定”，重新启动OMS服务。

----结束

10.12.2.4.3 修改组件数据库用户密码

操作场景

建议管理员定期修改组件数据库用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启服务，服务在重启时无法访问。

操作步骤

步骤1 在FusionInsight Manager选择“集群 > *待操作的集群名称* > 服务”。

步骤2 确定修改哪个组件数据库用户密码。

修改DBService数据库**omm**用户密码，参考**修改DBService数据库omm用户密码**章节进行操作，修改其他组件数据库用户密码，需要先停止服务再执行**步骤3**。

步骤3 单击待修改数据库用户密码的服务，选择“更多 > 修改数据库密码”，在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤4 根据界面信息，输入新旧密码。

密码复杂度要求：

- 组件数据库用户密码字符长度为8～32。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

步骤5 勾选“我已阅读此信息并了解其影响”，单击“确定”。

步骤6 密码修改完成后，选择“更多 > 重启服务”，在弹出窗口中输入当前登录的用户密码，单击“确定”，勾选“同时重启上层服务。”，单击“确定”开始重启服务。

----结束

10.12.2.4.4 修改 DBService 数据库 omm 用户密码

步骤1 以**root**用户登录DBService主节点。

说明

DBService数据库omm用户密码不支持在DBService备节点修改。只需在DBService主节点执行修改操作，无需在备管理节点操作。

步骤2 执行以下命令，切换用户。

su - omm

步骤3 执行以下命令，切换目录。

source $DBSERVER_HOME/.dbservice_profile

cd ${DBSERVICE_SOFTWARE_DIR}/sbin/

步骤4 执行以下命令，修改**omm**用户密码。

sh modifyDBPwd.sh

步骤5 输入**omm**的原密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

显示如下结果，说明修改成功：

```
Successful to modify password.
```

----结束

10.12.3 安全加固

10.12.3.1 加固策略

加固 Tomcat

在FusionInsight Manager软件安装及使用过程中，针对Tomcat基于开源做了如下功能增强：

- 升级Tomcat版本为官方稳定版本。
- 设置应用程序webapplications之下的目录权限为500，对webapplications之下的部分目录支持写权限。
- 系统软件安装完成后自动清除Tomcat安装包。
- webapplications下针对工程禁用自动部署功能，只部署了web、cas和client-registry三个工程。
- 禁用部分未使用的http方法，防止被他人利用攻击。
- 更改Tomcat服务器默认shutdown端口号和命令，避免被黑客捕获利用关闭服务器，降低对服务器和应用的威胁。
- 出于安全考虑，更改“maxHttpHeaderSize”的取值，给服务器管理员更大的可控性，以控制客户端不正常的请求行为。
- 安装Tomcat后，修改Tomcat版本描述文件。
- 为了避免暴露Tomcat自身的信息，更改Connector的Server属性值，使攻击者不易获知服务器的相关信息。
- 控制Tomcat自身配置文件、可执行文件、日志目录、临时目录等文件和目录的权限。

- 关闭会话facade回收重用功能，避免请求泄漏风险。
- CookieProcessor使用LegacyCookieProcessor，避免cookie中的敏感数据泄漏。

加固 LDAP

在安装完集群后，针对LDAP做了如下功能增强：

- LDAP配置文件中管理员密码使用SHA加密，当升级openldap版本为2.4.39或更高时，主备LDAP节点服务自动采用SASL External机制进行数据同步，避免密码信息被非法获取。
- 集群中的LDAP服务默认支持SSLv3协议，可安全使用。当升级openldap版本为2.4.39或更高时，LDAP将自动使用TLS1.0以上的协议通讯，避免未知的安全风险。

加固 JDK

- 如果客户端程序使用了AES256加密算法，则需要对JDK进行安全加固，具体操作如下：

  获取与JDK版本对应的JCE（Java Cryptography Extension）文件。JCE文件解压后包含“local_policy.jar”和“US_export_policy.jar”。拷贝此jar包到如下路径并替换文件：

  - Linux：“JDK安装目录/jre/lib/security”
  - Windows：“JDK安装目录\jre\lib\security”

  说明

  请访问Open JDK开源社区获取JCE文件。

- 如果客户端程序需要支持SM4加密算法，则需要更新jar包：

  在“客户端安装目录/JDK/jdk/jre/lib/ext/”目录下获取“SMS4JA.jar”，并拷贝到如下目录：

  - Linux：“JDK安装目录/jre/lib/ext/”
  - Windows：“JDK安装目录\jre\lib\ext\”

10.12.3.2 配置受信任 IP 访问 LDAP

操作场景

默认情况下，部署在OMS和集群中的LDAP服务允许任意IP访问。如果需要只允许受信任的IP地址访问LDAP服务，可以配置iptables过滤列表的INPUT策略。

对系统的影响

配置受信任IP访问LDAP以后，未配置的IP无法访问LDAP。扩容前，新增加的IP需要配置为受信任的IP。

前提条件

- 根据安装规划，收集集群内全部节点的管理平面IP、业务平面IP和所有浮动IP。
- 获取集群内节点的root用户和密码。

操作步骤

配置OMS LDAP信任的IP地址

步骤1 确定管理节点IP地址，请参见**登录管理节点**。

步骤2 登录FusionInsight Manager，请参见**登录管理系统**。

步骤3 选择“系统 > OMS”，在“服务”选择“oldap > 修改配置”，查看OMS LDAP端口号，即“Ldap服务监听端口”参数值。默认为“21750”。

步骤4 以**root**用户通过主管理节点的IP地址登录主管理节点。

步骤5 执行以下命令，查看iptables过滤列表中INPUT策略。

iptables -L

例如未配置任何规则时，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
```

步骤6 执行以下命令，将集群使用的所有IP地址配置为受信任的IP。每个IP需要添加一次。

iptables -A INPUT -s *受信任IP地址* **-p tcp --dport** *端口号* **-j ACCEPT**

例如，将10.0.0.1配置为受信任的IP，可以访问端口21750，执行：

iptables -A INPUT -s 10.0.0.1 -p tcp --dport 21750 -j ACCEPT

步骤7 执行以下命令，将全部IP地址配置为不受信任的IP。已配置为信任IP不受此规则影响。

iptables -A INPUT -p tcp --dport *端口号* **-j DROP**

例如，配置全部IP不能访问端口21750，执行：

iptables -A INPUT -p tcp --dport 21750 -j DROP

步骤8 执行以下命令，查看iptables过滤列表中修改后INPUT策略。

iptables -L

例如配置一个受信任IP后，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
ACCEPT     tcp  --  10.0.0.1             anywhere             tcp dpt:21750
DROP       tcp  --  anywhere             anywhere             tcp dpt:21750
```

步骤9 执行以下命令，查看iptables过滤列表中存在的规则及相对应的编号。

iptables -L -n --line-number

```
Chain INPUT (policy ACCEPT)
num target     prot opt source               destination
1   DROP       tcp  --  0.0.0.0/0            0.0.0.0/0           tcp dpt:21750
```

步骤10 根据实际需求，可执行以下命令，删除iptables过滤列表中的规则。

iptables -D INPUT *待删除的编号*

例如，删除编号为1的规则，执行：

iptables -D INPUT 1

步骤11 以**root**用户通过备管理节点的IP地址登录备管理节点，并重复**步骤5**到**步骤10**。

配置集群LDAP信任的IP地址

步骤12 登录FusionInsight Manager。

步骤13 选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 实例”，查看LDAP服务对应的节点。

步骤14 切换到“配置”，查看集群LDAP端口号，即“LDAP_SERVER_PORT”参数值。默认为“21780”。

步骤15 以**root**用户通过LDAP服务的IP地址登录LDAP节点。

步骤16 执行以下命令，查看iptables过滤列表中INPUT策略。

iptables -L

例如未配置任何规则时，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
```

步骤17 执行以下命令，将集群使用的所有IP地址配置为受信任的IP。每个IP需要添加一次。

iptables -A INPUT -s *受信任IP地址* **-p tcp --dport** *端口号* **-j ACCEPT**

例如，将10.0.0.1配置为受信任的IP，可以访问端口21780，执行：

iptables -A INPUT -s 10.0.0.1 -p tcp --dport 21780 -j ACCEPT

步骤18 执行以下命令，将全部IP地址配置为不受信任的IP。已配置为信任IP不受此规则影响。

iptables -A INPUT -p tcp --dport *端口号* **-j DROP**

例如，配置全部IP不能访问端口21780，执行：

iptables -A INPUT -p tcp --dport 21780 -j DROP

步骤19 执行以下命令，查看iptables过滤列表中修改后INPUT策略。

iptables -L

例如配置一个受信任IP后，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
ACCEPT     tcp  --  10.0.0.1             anywhere            tcp dpt:21780
DROP       tcp  --  anywhere             anywhere            tcp dpt:21780
```

步骤20 执行以下命令，查看iptables过滤列表中存在的规则及相对应的编号。

iptables -L -n --line-number

```
Chain INPUT (policy ACCEPT)
num target     prot opt source               destination
1   DROP       tcp  --  0.0.0.0/0            0.0.0.0/0           tcp dpt:21780
```

步骤21 根据实际需求，可执行以下命令，删除iptables过滤列表中的规则。

iptables -D INPUT *待删除的编号*

例如，删除编号为1的规则，执行：

iptables -D INPUT 1

步骤22 以**root**用户通过另一个LDAP服务的IP地址登录LDAP节点，并重复**步骤16**到**步骤21**。

----结束

10.12.3.3 加密 HFile 和 WAL 内容

加密 HFile 和 WAL 内容

须知

- 设置HFile和WAL为SMS4加密或AES加密方式对系统的影响较大，一旦操作失误会导致数据丢失。不推荐使用此功能。
- 使用Bulkload批量导入的数据不支持加密。

缺省情况下，HBase中的HFile和WAL（Write ahead log）内容是不加密的。如果用户需要对其进行加密，可通过如下操作进行配置。

步骤1 在任一安装HBase服务节点，使用**omm**用户执行如下命令创建密钥。

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh *<path>/hbase.jks <type> <length> <alias>*

- */<path>/hbase.jks*表示生成的jks文件存储路径。
- <type>表示加密的类型，支持SMS4或AES。
- <length>表示密钥的长度，SMS4支持16位长度，AES支持128位长度。
- *<alias>*为密钥文件的别名，第一次生成时请使用缺省值“omm”。

例如，生成SMS4加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks SMS4 16 omm

生成AES加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks AES 128 omm

说明

- 集群的操作用户需要有*<path>/hbase.jks*目录的“rw”权限，且要求目录已存在。
- 运行命令后需要再输入4遍相同的*<password>*。其中**步骤3**中进行加密的密码与此步骤的密码相同。

步骤2 将生成的密钥文件分发到集群中所有节点的相同目录下，并为**omm**用户配置该文件的读写权限。

说明

- 请管理员根据企业安全要求，选择安全的操作步骤分发密钥。
- 如果在使用过程中，有节点出现密钥文件丢失的情况，请按照此步骤从其他节点拷贝到该节点。

步骤3 在FusionInsight Manager界面中，设置“hbase.crypto.keyprovider.parameters.encryptedtext”参数的值为密文密码，设置“hbase.crypto.keyprovider.parameters.uri”参数的值为密钥路径和名称。

- “hbase.crypto.keyprovider.parameters.uri”格式为：**jceks://***<key_Path_Name>*。

  *<key_Path_Name>*填写密钥的存储路径，例如“/home/hbase/conf/hbase.jks”则对应参数值为“jceks:///home/hbase/conf/hbase.jks”。

- “hbase.crypto.keyprovider.parameters.encryptedtext”格式为：*<encrypted_password>*。

  *<encrypted_password>*填写创建密钥时的密文密码，参数值显示为密文。使用**omm**用户在安装HBase服务的节点，执行如下命令获取对应加密后的密码：

  sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh

  说明

  运行命令后需要输入<password>。该密码与**步骤1**中手动输入的密码相同。

步骤4 在FusionInsight Manager界面中，设置“hbase.crypto.key.algorithm”参数值为“SMS4”或“AES”，使HFile的内容采用SMS4或AES的方式加密。

步骤5 在FusionInsight Manager界面中，设置“hbase.crypto.wal.algorithm”参数值为“SMS4”或“AES”，使WAL的内容采用SMS4或AES的方式加密。

步骤6 在FusionInsight Manager界面中，将“hbase.regionserver.wal.encryption”参数值修改为“true”。

步骤7 保存设置，并重启HBase服务使其生效。

步骤8 在创建HBase表时，需要通过设置加密方式，<type>表示加密的类型。

- 通过命令行创建表时，直接设置加密方式为SMS4或AES。

  create '*<table name>*', {*NAME => 'd'*, ***ENCRYPTION => '****<type>****'***}

- 使用代码创建表时，在代码中添加如下信息设置加密方式为SMS4或AES。

```
public void testCreateTable()
{
    String tableName = "user";
    Configuration conf = getConfiguration();
    HTableDescriptor htd = new HTableDescriptor(TableName.valueOf(tableName));

    HColumnDescriptor hcd = new HColumnDescriptor("info");
    //设置加密方式为SMS4或AES。
    hcd.setEncryptionType("<type>");
    htd.addFamily(hcd);

    HBaseAdmin admin = null;
    try
    {
        admin = new HBaseAdmin(conf);

        if(!admin.tableExists(tableName))
        {
            admin.createTable(htd);
        }
    }
    catch (IOException e)
    {
        e.printStackTrace();
    }
    finally
```

```
    {
        if(admin != null)
        {
            try
            {
                admin.close();
            }
            catch (IOException e)
            {
                e.printStackTrace();
            }
        }
    }
}
```

步骤9 如果用户已按照**步骤1**到**步骤7**配置了SMS4或AES加密，但是在执行**步骤8**创建表时，未设置对应的加密参数，使得插入的数据未经加密。

此时，您可以执行如下步骤对之前插入的数据进行加密。

1. 针对表执行flush动作，将内存中的数据导入到HFile中。

   flush *'<table_name>'*

2. 执行以下步骤修改表属性。

   disable *'<table_name>'*

   alter *'<table_name>'*,**NAME=>** *'<column_name>'*,**ENCRYPTION =>** *'<type>'*

   enable *'<table_name>'*

3. 插入一条新的数据，然后flush表。

   说明

   必须要插入一条新的数据，HFile才会生成新的HFile，使得之前插入的未加密数据进行重写加密。

   put *'<table_name>'*, **'id2','f1:c1','value22222222222222222222222222222222'**

   flush *'<table_name>'*

4. 执行如下步骤重写HFile。

   major_compact' *<table_name>'*

须知

执行此步骤时，HBase表是被禁用的，不能提供表服务。请谨慎使用。

----结束

修改密钥文件

须知

修改密钥文件的操作对系统影响较大，一旦操作失误会导致数据丢失。不推荐使用此功能。

在**加密HFile和WAL内容**操作中需要生成对应的密钥文件并设置密码，为确保系统安全，在运行一段时间后，用户可修改密钥，使用新的密钥文件对HFile和WAL内容进行加密。

步骤1 使用**omm**用户执行如下命令生成新的密钥文件。

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh *<path>/hbase.jks <type> <length> <alias-new>*

- *<path>/hbase.jks*:表示生成的hbase.jks文件的存储路径。该路径和文件名称需与**加密HFile和WAL内容**章节生成的密钥文件相同。
- *<alias-new>*：表示密钥文件的别名，请使用与旧密钥文件不同的名字。
- *<type>*表示加密的类型，支持SMS4或AES。
- *<length>*表示密钥的长度，SMS4支持16位长度，AES支持128位长度。

例如，生成SMS4加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks SMS4 16 omm_new

生成AES加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks AES 128 omm_new

说明

- 集群的操作用户需要有*<path>/hbase.jks*目录的“rw”权限，且要求目录已存在。
- 运行命令后需要再输入3遍相同的*<password>*，该密码表示密钥文件的密码，请直接使用旧文件的密码，不会产生安全风险。

步骤2 将生成的密钥文件分发到集群中所有节点的相同目录下，并为**omm**用户配置该文件的读写权限。

说明

请管理员根据企业安全要求，选择安全的操作步骤分发密钥。

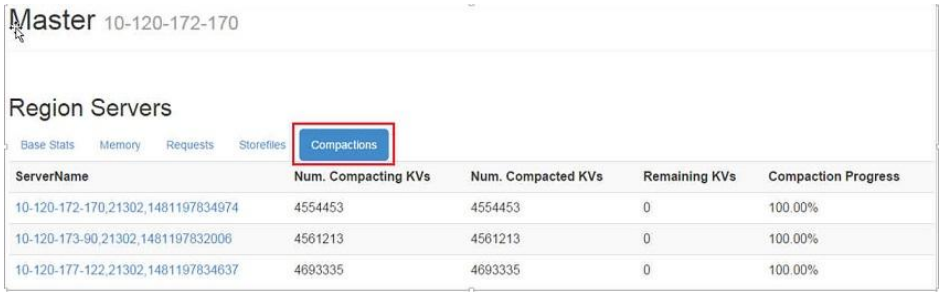
步骤3 在FusionInsight Manager的HBase服务配置界面中增加自定义配置项，设置“hbase.crypto.master.key.name”为“omm_new”，设置“hbase.crypto.master.alternate.key.name”为“omm”，然后保存配置。

步骤4 重启HBase服务，使配置生效。

步骤5 在HBase shell中执行**major compact**命令，生成基于新的加密算法的HFile文件。

major_compact *'<table_name>'*

步骤6 从HMaster的网页中可以查看到major compact进度。



步骤7 所有的“Compaction Progress”都为100%且“Remaining KVs”都为0时，使用**omm**用户执行如下命令销毁旧的密钥文件：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh *<path>/hbase.jks <alias-old>*

- *<path>/hbase.jks*:表示生成的“hbase.jks”文件的存储路径。该路径和文件名称需与**加密HFile和WAL内容**章节生成的密钥文件相同。
- *<alias-old>*：表示要删除的旧密钥文件的别名。

例如：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks omm

说明

集群的操作用户需要有*<path>/hbase.jks*目录的“rw”权限，且要求目录已存在。

步骤8 再执行**步骤2**，重新分发更新后的密钥文件。

步骤9 从FusionInsight Manager中删除**步骤3**中新增HBase自定义配置项“hbase.crypto.master.alternate.key.name”。

步骤10 再执行**步骤4**使配置生效。

----结束

10.12.3.4 安全配置

设置安全通道加密

默认情况下，组件间的通道是不加密的。您可以配置如下参数，设置安全通道是加密的。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > *服务名* > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

说明

配置后需要重启对应服务。

表 10-82 参数说明

配置参数	说明	缺省值
hbase.rpc.protection	设置HBase通道是否加密，包含HBase客户端访问HBase服务端的RPC（remote procedure call）通道，HMaster和RegionServer间的RPC通道。设置为“privacy”表示通道加密，认证、完整性和隐私性功能都全部开启，设置为“integrity”表示不加密，只开启认证和完整性功能，设置为“authentication”表示不加密，仅要求认证报文，不要求完整性和隐私性。 **说明** privacy会对传输内容进行加密，包括用户token等敏感信息，以确保传输信息的安全，但是该方式对性能影响很大，对比另外两种方式，会带来约60%的读写性能下降。请根据企业安全要求修改配置，且客户端与服务端中该配置项需使用相同设置。	-
dfs.encrypt.data.transfer	设置客户端访问HDFS的通道和HDFS数据传输通道是否加密。HDFS数据传输通道包括DataNode间的数据传输通道，客户端访问DataNode的DT（Data Transfer）通道。设置为“true”表示加密，默认不加密。	“false”
dfs.encrypt.data.transfer.algorithm	设置客户端访问HDFS的通道和HDFS数据传输通道是否加密。只有在dfs.encrypt.data.transfer配置项设置为true，此参数才会生效。 缺省值为“3des”，表示采用3DES算法进行加密。此处的值还可以设置为“rc4”，避免出现安全隐患，不推荐设置为该值。	“3des”

配置参数	说明	缺省值
hadoop.rpc.protection	设置Hadoop中各模块的RPC通道是否加密。包括： • 客户端访问HDFS的RPC通道。 • HDFS中各模块间的RPC通道，如DataNode与NameNode间。 • 客户端访问Yarn的RPC通道。 • NodeManager和ResourceManager间的RPC通道。 • Spark访问Yarn，Spark访问HDFS的RPC通道。 • MapReduce访问Yarn，Mapreduce访问HDFS的RPC通道。 • HBase访问HDFS的RPC通道。 默认设置为“privacy”表示加密，“authentication”表示不加密。 **说明** 您可以在HDFS组件的配置界面中设置该参数的值，设置后全局生效，即Hadoop中各模块的RPC通道是否加密全部生效。	• 安全模式：privacy • 普通模式：authentication

Web 最大并发连接数限制

为了保护Web服务器的可靠性，当访问的用户连接数达到一定数量之后，对新增用户的连接进行限制。防止大量同时登录和访问，导致服务不可用，同时避免DDOS攻击。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > *服务名* > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

表 10-83 参数说明

配置参数	说明	缺省值
hadoop.http.server.MaxRequests	设置各组件Web的最大并发连接数限制。 相关组件为HDFS和YARN。	2000
spark.connection.maxRequest	JobHistory允许的最大请求连接数。	5000

10.12.3.5 配置 HBase 允许修改操作的 IP 地址白名单

当HBase集群开启Replication功能时，为了保护主备集群的HBase数据一致性，对备集群HBase增加了数据修改操作的保护。当备集群HBase接收到数据修改操作的RPC请求时，首先检查发出该请求的用户的权限，只有HBase管理用户才有修改权限；其次检查发出该请求的IP的有效性，备集群只接收来自IP白名单中的机器发起的修改请求。IP白名单通过配置项“hbase.replication.allowedIPs”配置。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > *服务名* > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

表 10-84 参数说明

配置参数	说明	默认值
hbase.replication.allowedIPs	仅允许指定IP地址的复制请求。支持逗号分隔型regex模式。以下模式均支持： ● Regex模式 例如: 10.18.40.* , 10.18.* , 10.18.40.11 ● Range模式（只能指定八位字节的最后一个的范围） 例如: 10.18.40.[10-20] 参数值默认为空，为空时IP白名单为备集群RegionServer的IP，表示只接受来自备集群RegionServer的修改请求。	N/A

10.12.3.6 更新集群密钥

操作场景

在安装集群时，系统将自动生成加密密钥key值以对集群的部分安全信息（例如所有数据库用户密码、密钥文件访问密码等）进行加密存储。在集群安装成功后，如果原始密钥不慎意外泄露或者需要使用新的密钥，系统管理员可以通过以下操作手动更改密钥值。

对系统的影响

- 更新集群密钥后，集群中新增加一个随机生成的新密钥，用于加密解密新保存的数据。旧的密钥不会删除，用于解密旧的加密数据。在修改安全信息后，例如修改数据库用户密码，新密码将使用新的密钥加密。
- 更新集群密钥需要停止集群，集群停止时无法访问。

前提条件

- 已确认主备管理节点IP。请参见**登录管理节点**。
- 停止依赖集群运行的上层业务应用。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 停止”，输入当前登录的用户密码确认身份。

在确认停止的对话框单击“确定”，等待界面提示停止成功。

步骤3 以**omm**用户登录主管理节点。

步骤4 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤5 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/tools

步骤6 执行以下命令，更新集群密钥。

sh updateRootKey.sh

根据界面提示，输入**y**：

```
The root key update is a critical operation.
Do you want to continue?(y/n):
```

界面提示以下信息表示更新密钥成功：

```
Step 4-1: The key save path is obtained successfully.
...
Step 4-4: The root key is sent successfully.
```

步骤7 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 启动”。

在弹出窗口中单击“确定”，等待界面提示启动成功。

----结束

10.12.3.7 加固 LDAP

配置 LDAP 防火墙策略

在双平面组网的集群中，由于LDAP部署在业务平面中，为保证LDAP数据安全，建议通过配置整个集群对外的防火墙策略，关闭LDAP相关端口。

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 配置”。

步骤3 查看“LDAP_SERVER_PORT”参数值，即为LdapServer的服务端口。

步骤4 根据客户的实际防火墙环境，配置整个集群对外的防火墙策略，将该端口关闭，以保证数据安全。

----结束

开启 LDAP 审计日志输出

用户可以通过设置LDAP服务的审计日志输出级别，将审计内容输出至系统日志信息中（如“/var/log/messages”），用于查看用户的活动信息及操作指令信息。

说明

LDAP的审计日志开启后，会产生大量日志信息，严重影响集群性能，请谨慎开启。

步骤1 登录任一LdapServer节点。

步骤2 执行以下命令，编辑“slapd.conf.consumer”文件，将“loglevel”的值设置为“256”（loglevel定义可以在OS上使用**man slapd.conf**命令查看）。

cd ${BIGDATA_HOME}/FusionInsight_BASE_8.1.0.1/install/FusionInsight-ldapserver-2.7.0/ldapserver/local/template

vi slapd.conf.consumer

```
...
pidfile         [PID_FILE_SLAPD_PID]
argsfile        [PID_FILE_SLAPD_ARGS]
loglevel    256
...
```

步骤3 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 更多 > 重启服务”，验证当前用户身份后重启服务。

----结束

10.12.3.8 配置 Kafka 数据传输加密

操作场景

Kafka客户端和Broker之间的数据传输默认采用明文传输，客户端可能部署在不受信任的网络中，传输的数据可能遭到泄漏和篡改。

操作步骤

默认情况下，组件间的通道是不加密的。用户可以配置如下参数，设置安全通道为加密的。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

说明

配置后应重启对应服务使参数生效。

Kafka服务端的传输加密相关配置参数如表10-85所示。

表 10-85 Kafka 服务端传输加密参数

配置项	描述	默认值
ssl.mode.enable	是否开启SSL对应服务。如果设置为“true”，那么Broker启动过程中会启动SSL的相关服务。	false
security.inter.broker.protocol	Broker间通信协议。支持PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL这四种协议类型。	SASL_PLAINTEXT

“ssl.mode.enable”配置为“true”后，Broker会开启SSL、SASL_SSL两种协议的服务，然后服务端或者客户端才能配置相关的SSL协议，进行传输加密通信。

10.12.3.9 配置 HDFS 数据传输加密

设置 HDFS 安全通道加密

默认情况下，组件间的通道是不加密的。您可以配置如下参数，设置安全通道为加密的。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

说明

配置后应重启对应服务使参数生效。

表 10-86 参数说明

配置项	描述	默认值
hadoop.rpc.protection	**须知** ● 设置后需要重启服务生效，且不支持滚动重启。 ● 设置后需要重新下载客户端配置，否则HDFS无法提供读写服务。 设置Hadoop中各模块的RPC通道是否加密。通道包括： ● 客户端访问HDFS的RPC通道。 ● HDFS中各模块间的RPC通道，如DataNode与NameNode间。 ● 客户端访问Yarn的RPC通道 ● NodeManager和ResourceManager间的RPC通道。 ● Spark访问Yarn，Spark访问HDFS的RPC通道。 ● Mapreduce访问Yarn，MapReduce访问HDFS的RPC通道。 ● HBase访问HDFS的RPC通道。 **说明** 设置后全局生效，即Hadoop中各模块的RPC通道的加密属性全部生效。	● 安全模式：privacy ● 普通模式：authentication **说明** ● “authentication”：只进行认证，不加密。 ● “integrity”：进行认证和一致性校验。 ● “privacy”：进行认证、一致性校验、加密。

配置项	描述	默认值
dfs.encrypt.data.transfer	设置客户端访问HDFS的通道和HDFS数据传输通道是否加密。HDFS数据传输通道包括DataNode间的数据传输通道，客户端访问DataNode的DT（Data Transfer）通道。设置为“true”表示加密，默认不加密。 **说明** • 仅当hadoop.rpc.protection设置为privacy时使用。 • 业务数据传输量较大时，默认启用加密对性能影响严重，使用时请注意。 • 如果互信集群的一端集群配置了数据传输加密，则对端集群也需配置同样的数据传输加密。	false
dfs.encrypt.data.transfer.algorithm	设置客户端访问HDFS的通道和HDFS数据传输通道的加密算法。只有在dfs.encrypt.data.transfer配置项设置为“true”，此参数才会生效。 **说明** 缺省值为“3des”，表示采用3DES算法进行加密。此处的值还可以设置为“rc4”，避免出现安全隐患，不推荐设置为该值。	3des
dfs.encrypt.data.transfer.cipher.suites	可以设置为空或“AES/CTR/NoPadding”，用于指定数据加密的密码套件。如果不指定此参数，则使用“dfs.encrypt.data.transfer.algorithm”参数指定的加密算法进行数据加密。默认值为“AES/CTR/NoPadding”。	AES/CTR/NoPadding

10.12.3.10 配置 Controller 与 Agent 间通信加密

操作场景

安装集群后Controller和Agent之间需要进行数据通信，在通信的过程中采用了Kerberos认证，出于对集群性能的考虑，通信过程默认不加密，对于一些安全要求较高用户可以采用以下方式进行加密。

对系统的影响

- 执行加密操作时，会自动重启Controller和所有Agent，重启期间会造成FusionInsight Manager暂时中断。
- 大集群下会导致管理节点性能有所下降，建议集群不超过200节点时开启该功能。

前提条件

已确认主备管理节点IP。

操作步骤

步骤1 以**omm**用户登录到主管理节点。

步骤2 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤3 执行以下命令，切换目录。

cd ${CONTROLLER_HOME}/sbin

步骤4 执行以下命令启用通信加密：

./enableRPCEncrypt.sh -t

执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**，查看主管理节点Controller的“ResHAStatus”是否为“Normal”，并可以重新登录FusionInsight Manager表示更改成功。

步骤5 如果需要关闭加密模式，执行以下命令：

./enableRPCEncrypt.sh -f

执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**，查看主管理节点Controller的“ResHAStatus”是否为“Normal”，并可以重新登录FusionInsight Manager表示更改成功。

----结束

10.12.3.11 更新 omm 用户 ssh 密钥

操作场景

在安装集群时，系统将自动为**omm**用户生成ssh认证私钥和公钥，用来建立节点间的互信。在集群安装成功后，如果原始私钥不慎意外泄露或者需要使用新的密钥时，系统管理员可以通过以下操作手动更改密钥值。

前提条件

- 已停止集群。
- 修改时禁止同时进行其他管理类操作。

操作步骤

步骤1 以**omm**用户登录到需要替换ssh密钥的节点。

如果该节点是Manager管理节点，务必在主管理节点上执行相关操作。

步骤2 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤3 执行以下命令，为节点生成新的密钥：

- 如果当前节点是Manager管理节点，执行以下命令：

  sh ${CONTROLLER_HOME}/sbin/update-ssh-key.sh

- 如果当前节点是非管理节点，执行以下命令：

  sh ${NODE_AGENT_HOME}/bin/update-ssh-key.sh

执行上述命令时界面提示“Succeed to update ssh private key.”信息，表示ssh密钥生成成功。

步骤4 执行以下命令将该节点的公钥拷贝到主管理节点：

scp ${HOME}/.ssh/id_rsa.pub** *oms_ip*:**${HOME}/.ssh/id_rsa.pub_bak

oms_ip：表示主管理节点IP。

根据提示输入**omm**用户密码完成文件拷贝。

步骤5 以**omm**用户登录到主管理节点。

步骤6 执行以下命令，防止超时退出：

TMOUT=0

步骤7 执行以下命令，切换目录：

cd ${HOME}/.ssh

步骤8 执行以下命令添加新的公钥信息：

cat id_rsa.pub_bak >> authorized_keys

步骤9 执行以下命令移动临时公钥文件到其他目录，例如，移动到“/tmp”目录。

mv -f id_rsa.pub_bak /tmp

步骤10 拷贝主管理节点的authorized_keys文件到集群内其他节点：

scp authorized_keys *node_ip*:**/${HOME}/.ssh/authorized_keys**

node_ip：集群内其他节点IP，不支持多个IP。

步骤11 执行以下命令无需输入密码确认私钥替换完成：

ssh *node_ip*

node_ip：集群内其他节点IP，不支持多个IP。

步骤12 登录FusionInsight Manager，在“主页”中单击待操作集群名称后的“ ··· > 启动”，启动集群。

----结束

10.12.4 安全维护

10.12.4.1 帐户维护建议

建议系统管理员对帐户例行检查，检查的内容包括：

- 操作系统、FusionInsight Manager以及各组件的帐户是否有必要，临时帐户是否已删除。
- 各类帐户的权限是否合理。不同的管理员拥有不同的权限。
- 对各类帐户的登录、操作记录进行检查和审计。

10.12.4.2 密码维护建议

用户身份验证是应用系统的门户。用户的帐户和密码的复杂性、有效期等需根据客户的安全要求进行配置。

对密码的维护建议如下：

1. 专人保管操作系统密码。
2. 密码需要满足一定的强度要求，例如密码最少字符数、混合大小写等。
3. 密码传递时注意加密，尽量避免通过邮件传递密码。
4. 密码需要加密存储。
5. 系统移交时提醒企业用户更改密码。
6. 定期修改密码。

10.12.4.3 日志维护建议

利用日志记录来帮助发现非法操作、非法登录用户等异常情况。系统对于重要业务的操作需要记录日志。通过日志文件来定位异常。

定期检查日志

定期查看系统日志，若发现有非法操作、非法登录用户等异常情况，应根据异常情况进行相应的处理。

定期备份日志

FusionInsight Manager和集群提供的审计日志记录了用户活动信息和操作信息，可通过FusionInsight Manager导出审计日志。当系统中的审计日志过多时，可通过配置转储参数，将审计日志转储到指定服务器，避免引起集群节点磁盘空间不足。

维护责任人

网络监控工程师、系统维护工程师

10.12.5 安全声明

JDK 使用声明

MRS是一个大数据集群，为用户提供分布式的数据分析计算能力。本产品自带的JDK为OpenJDK，主要使用场景如下：

- 平台服务运行及维护使用。

- Linux客户端运行时使用（主要为业务提交、应用运维等）。

JDK 风险说明

系统对自带的JDK进行了权限控制，只有属于FusionInsight平台相关群组的用户才有权限访问，且平台部署在客户内网，安全风险较低。

JDK 加固

JDK加固相关操作请参考**加固策略**的“加固JDK”部分。

Hue 组件包含公网 IP 的说明

Hue组件使用的ipadrress，requests，Django等第三方包的测试用例及其注释包含的公网IP，组件在提供服务时不涉及这些IP，Hue组件的配置文件中不涉及公网IP。

10.13 告警参考（适用于 MRS 3.x 版本）

10.13.1 ALM-12001 审计日志转储失败

告警解释

根据本地历史数据备份策略，集群的审计日志需要转储到第三方服务器上。系统每天零晨3点开始周期性检测转储服务器，如果转储服务器满足配置条件，审计日志可以成功转储。审计日志转储失败，系统产生此告警。如果第三方服务器的转储目录磁盘空间不足，或者用户修改了转储服务器的用户名、密码或转储目录，将会导致审计日志转储失败。

告警属性

告警ID	告警级别	是否自动清除
12001	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

系统本地最多只能保存50个转储文件，如果该故障持续存在于转储服务器，本地审计日志可能丢失。

可能原因

- 网络连接异常。
- 转储服务器的用户名、密码或转储目录不满足配置条件。
- 转储目录的磁盘空间不足。

处理步骤

检查网络连接是否正常。

步骤1 在FusionInsight Manager界面，选择“审计 > 配置”，进入审计日志转储配置页面。

步骤2 查看转储配置页面中当前的SFTP IP值是否合法有效。

以**root**用户登录到任一管理节点，执行**ping**命令检查SFTP服务器和集群之间的网络连接是否正常。

- 是，执行**步骤5**。
- 否，执行**步骤3**。

步骤3 修复网络连接，然后重新配置SFTP服务端密码，单击“确定”，重新下发一次配置。

步骤4 2分钟后，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤5**。

检查用户名、密码和转储目录是否正确。

步骤5 查看转储配置页面中当前的第三方服务器用户名、密码和转储目录是否正确。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 修改用户名、密码和转储目录，单击“确定”，重新下发一次配置。

步骤7 2分钟后，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤8**。

检查转储目录的磁盘空间是否足够。

步骤8 根据转储配置页面中当前的转储目录，以**root**用户登录到第三方服务器，使用**df**命令检查第三方服务器的转储目录的磁盘空间是否大于100MB。

- 是，执行**步骤11**。
- 否，执行**步骤9**。

步骤9 扩大第三方服务器的磁盘空间，然后重新配置SFTP服务端密码，单击“确定”，重新下发一次配置。

步骤10 2分钟后，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤11**。

重新设置转储规则。

步骤11 在FusionInsight Manager界面，选择“审计 > 配置”。

步骤12 重新设置转储规则，填入正确的参数，单击“确定”。

步骤13 2分钟后，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤14**。

收集故障信息。

步骤14 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤15 在“服务”中勾选“OmmServer”，单击“确定”。

步骤16 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤17 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.2 ALM-12004 OLdap 资源异常

告警解释

系统按60秒周期检测Ldap资源，当连续6次监控到Manager中的Ldap资源异常时，系统产生此告警。

当Manager中的Ldap资源恢复，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12004	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Ldap资源异常，Manager和组件WebUI认证服务不可用，无法对Web上层服务提供安全认证和用户管理功能，可能引起无法登录Manager和组件的WebUI。

可能原因

Manager中LdapServer进程故障。

处理步骤

检查Manager中LdapServer进程是否正常。

步骤1 以**omm**用户登录集群中的Manager所在节点主机。

可以通过登录FusionInsight Manager浮动IP节点，执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**命令来查看当前Manager的双机信息。

步骤2 执行**ps -ef | grep slapd**，查询配置文件位于“${BIGDATA_HOME}/om-server/om/”路径下面的LdapServer资源进程是否正常。

说明

判断资源正常有两个标识：

1. 执行完**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**命令后查看到oldap的“ResHAStatus”为“Normal”。
2. 执行**ps -ef | grep slapd**，可以查看到有端口为21750的slapd进程。
   - 是，执行**步骤3**。
   - 否，执行**步骤4**。

步骤3 执行**kill -2** *ldap进程pid*，等待20s以后，HA会自动启动OLdap进程。观察当前OLdap资源状态是否正常。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“OmsLdapServer”和“OmmServer”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.3 ALM-12005 OKerberos 资源异常

告警解释

告警模块对Manager中的Kerberos资源的状态按80秒周期进行监控，当连续6次监控到Kerberos资源异常时，系统产生此告警。

当Kerberos资源恢复时，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Manager中的Kerberos资源异常，组件WebUI认证服务不可用，无法对Web上层服务提供安全认证功能，可能引起无法登录FusionInsight Manager和组件的WebUI。

可能原因

Okerberos依赖的OLdap资源异常。

处理步骤

检查Manager中的OKerberos依赖的OLdap资源是否异常。

步骤1 以**omm**用户登录到集群中Manager所在节点主机。

通过登录FusionInsight Manager浮动IP节点，执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**脚本来查看当前Manager的双机信息。

步骤2 执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的OLdap资源状态是否正常（单机模式下面，OLdap资源为Active_normal状态；双机模式下，OLdap资源在主节点为Active_normal状态，在备节点为Standby_normal状态。）。

- 是，执行**步骤4**。
- 否，执行**步骤3**。

步骤3 参考**ALM-12004 OLdap资源异常**的处理步骤进行处理，OLdap资源状态恢复后，观察当前OKerberos资源状态是否恢复正常。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“OmsKerberos”和“OmmServer”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.4 ALM-12006 节点故障

告警解释

Controller按30秒周期检测NodeAgent心跳。当Controller未接收到某一个NodeAgent的心跳，则尝试重启该NodeAgent进程，如果连续三次重启失败，产生该告警。

当Controller可以正常接收时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

节点业务无法提供。

可能原因

网络断连、硬件故障或操作系统执行命令缓慢。

处理步骤

检查网络是否断连、硬件是否故障或者操作系统执行命令缓慢。

步骤1 在FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，单击主机名，查看该告警的主机地址。

步骤2 以**root**用户登录主管理节点。

步骤3 执行**ping** *故障主机IP地址*命令检查故障节点是否可达。

- 是，执行**步骤12**。
- 否，执行**步骤4**。

步骤4 联系网络管理员查看是否为网络故障。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 修复网络故障，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤6**。

步骤6 联系系统管理员查看是否节点硬件故障（CPU或者内存等）。

- 是，执行**步骤7**。
- 否，执行**步骤12**。

步骤7 维修或者更换故障部件，并重启节点。查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤8**。

步骤8 当集群中上报大量的节点故障时，可能是浮动IP资源异常导致controller无法检测agent心跳。

登录任一管理节点，查看“/var/log/Bigdata/omm/oms/ha/scriptlog/floatip.log”，查看故障出现前后1-2分钟的日志是否完整。

例如：完整日志为如下格式：

```
2017-12-09 04:10:51,000 INFO (floatip) Read from ${BIGDATA_HOME}/om-server_8.1.0.1/om/etc/om/
routeSetConf.ini,value is : yes
2017-12-09 04:10:51,000 INFO (floatip) check wsNetExport : eth0 is up.
2017-12-09 04:10:51,000 INFO (floatip) check omNetExport : eth0 is up.
2017-12-09 04:10:51,000 INFO (floatip) check wsInterface : eRth0:oms, wsFloatIp: XXX.XXX.XXX.XXX.
2017-12-09 04:10:51,000 INFO (floatip) check omInterface : eth0:oms, omFloatIp: XXX.XXX.XXX.XXX.
2017-12-09 04:10:51,000 INFO (floatip) check  wsFloatIp : XXX.XXX.XXX.XXX is reachable.
2017-12-09 04:10:52,000 INFO (floatip) check  omFloatIp : XXX.XXX.XXX.XXX is reachable.
```

- 是，执行**步骤12**。
- 否，执行**步骤9**。

步骤9 查看检测完wsNetExport后是否打印omNetExport的检测日志或两条日志打印间隔时间超过10s或更长。

- 是，执行**步骤10**。
- 否，执行**步骤12**。

步骤10 查看操作系统的“/var/log/message”，查看故障出现时间段是否有sssd频繁重启或者nscd异常信息（Red hat操作系统确认sssd信息，SUSE操作系统确认nscd信息）。

sssd重启样例

```
Feb  7 11:38:16 10-132-190-105 sssd[pam]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd[nss]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd[nss]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd[be[default]]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd: Starting up
Feb  7 11:38:16 10-132-190-105 sssd[be[default]]: Starting up
Feb  7 11:38:16 10-132-190-105 sssd[nss]: Starting up
Feb  7 11:38:16 10-132-190-105 sssd[pam]: Starting up
```

nscd异常信息样例

```
Feb 11 11:44:42 10-120-205-33 nscd: nss_ldap: failed to bind to LDAP server ldaps://10.120.205.55:21780:
Can't contact LDAP server
Feb 11 11:44:43 10-120-205-33 ntpq: nss_ldap: failed to bind to LDAP server ldaps://10.120.205.55:21780:
Can't contact LDAP server
Feb 11 11:44:44 10-120-205-33 ntpq: nss_ldap: failed to bind to LDAP server ldaps://10.120.205.92:21780:
Can't contact LDAP server
```

- 是，执行**步骤11**。
- 否，执行**步骤12**。

步骤11 排查Ldapserver节点是否故障，例如业务IP不可达、网络延时过长等；若故障为阶段性，则需在故障时排查，并尝试执行**top**命令查看是否存在异常软件。

收集故障信息。

步骤12 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤13 在“服务”中勾选如下节点信息，单击“确定”。

- NodeAgent
- Controller
- OS

步骤14 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤15 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.5 ALM-12007 进程故障

告警解释

进程健康检查模块按5秒周期检测进程状态。当进程健康检查模块连续三次检测到进程连接状态为故障时，产生该告警。

当进程连接正常时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

该进程提供的业务不可用。

可能原因

- 实例进程异常。
- 磁盘空间不足。

说明

如果同一时间段，存在大量的进程故障告警，则可能存在安装目录文件被误删除或者权限被修改。

处理步骤

检查实例进程是否异常。

步骤1 打开FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，单击主机名查看该告警的主机名称与服务名称。

步骤2 在“告警”页面，查看是否有**ALM-12006 节点故障**告警产生。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 按**ALM-12006 节点故障**提供的步骤处理该告警。

步骤4 以**root**用户登录该告警的主机地址。查看告警角色所在安装目录用户、用户组、权限等是否正常。正常用户、用户组、权限为“omm：ficommon 750”。

例如：NameNode的安装目录为“*${BIGDATA_HOME}*/FusionInsight_Current/*1_8_NameNode*/etc”。

- 是，执行**步骤6**。
- 否，执行**步骤5**。

步骤5 执行如下命令将文件夹权限修改为“750”，并将“用户:属组”修改为“omm:ficommon”。

chmod 750 *<folder_name>*

chown omm:ficommon *<folder_name>*

步骤6 等待5分钟，查看告警列表中，“ALM-12007 进程故障”告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤7**。

检查磁盘空间是否不足。

步骤7 在FusionInsight Manager的告警列表中，查看是否有“ALM-12017 磁盘容量不足”告警产生。

- 是，执行**步骤8**。
- 否，执行**步骤11**。

步骤8 按**ALM-12017 磁盘容量不足**提供的步骤处理该故障。

步骤9 等待5分钟，查看告警列表中，“ALM-12017 磁盘容量不足”告警是否已清除。

- 是，执行**步骤10**。
- 否，执行**步骤11**。

步骤10 等待5分钟，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤12 根据在**步骤1**获取的服务名称，在“服务”中勾选对应的组件及“NodeAgent”，单击“确定”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.6 ALM-12010 Manager 主备节点间心跳中断

告警解释

当主Manager节点在7秒内没有收到备Manager节点的心跳信号时，产生该告警。

当主Manager节点收到备Manager节点的心跳信号后，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12010	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

如果主Manager进程异常，主备倒换无法进行，影响业务。

可能原因

- 主备Manager节点间链路异常。
- 节点名配置错误。
- 防火墙禁用端口。

处理步骤

检查主备Manager服务器间的网络是否正常。

步骤1 在FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，查看该告警的备Manager服务器（即Peer Manager）IP地址。

步骤2 以**root**用户登录主Manager服务器。

步骤3 执行**ping** *备Manager心跳IP地址*命令检查备Manager服务器是否可达。

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 联系网络管理员查看是否为网络故障。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 修复网络故障，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查节点名配置是否正确。

步骤6 进入软件安装目录。

cd /opt

步骤7 查找主备节点的配置文件目录。

find -name hacom_local.xml

步骤8 进入workspace目录。

cd ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/local/hacom/conf/

步骤9 使用**vim**命令打开hacom_local.xml，查看local、peer节点配置是否正确，local配置主节点，peer配置备节点。

- 是，执行**步骤12**。
- 否，执行**步骤10**。

步骤10 修改hacom_local.xml中主备节点的配置，修改完成后，按**Esc**回到命令模式，输入命令**:wq**保存退出。

步骤11 查看此告警信息是否自动清除。

- 是，处理完毕。
- 否，执行**步骤12**。

检查是否防火墙禁用端口。

步骤12 执行命令**lsof -i :20012**查询主备节点的心跳端口是否打开，有查询结果说明端口已经开放，否则说明端口被防火墙禁用。

- 是，执行**步骤13**。
- 否，执行**步骤16**。

步骤13 执行命令**iptables -P INPUT ACCEPT**，防止与服务器断开。

步骤14 清除防火墙。

iptables -F

步骤15 查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤16**。

收集故障信息。

步骤16 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤17 在“服务”中勾选如下节点信息，单击“确定”。

- OmmServer
- Controller
- NodeAgent

步骤18 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤19 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.7 ALM-12011 Manager 主备节点同步数据异常

告警解释

系统按60秒周期检测Manager主备节点同步数据情况，当备Manager无法与主Manager同步文件时，产生该告警。

当备Manager与主Manager正常同步文件时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12011	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

备Manager的配置文件没有更新。主备倒换之后，一些配置可能会丢失。Manager及部分组件可能无法正常运行。

可能原因

- 主备Manager节点间链路中断，/srv/BigData/LocalBackup目录存储空间已满。
- 同步文件不存在，或者文件权限有误。

处理步骤

检查主备Manager服务器间的网络是否正常。

步骤1 在FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，获取该告警的备Manager（即Peer Manager）IP地址。

步骤2 以**root**用户登录主Manager服务器。

步骤3 执行**ping** *备Manager IP地址*命令检查备Manager服务器是否可达。

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 联系网络管理员查看是否为网络故障。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 修复网络故障，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查/srv/BigData/LocalBackup目录存储空间是否已满。

步骤6 执行以下命令检查“/srv/BigData/LocalBackup”目录存储空间是否已满：

df -hl /srv/BigData/LocalBackup

- 是，执行**步骤7**。
- 否，执行**步骤10**。

步骤7 执行以下命令清理不需要的备份文件：

rm -rf *待清理的目录路径*

例如：

rm -rf /srv/BigData/LocalBackup/0/default-oms_20191211143443

步骤8 在FusionInsight Manager界面，选择“运维 > 备份恢复 > 备份管理”。

在待操作备份任务右侧“操作”栏下，单击“配置”，修改“最大备份数”减少备份文件集数量。

步骤9 等待大约1分钟，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤10**。

检查同步文件是否存在，文件权限是否异常。

步骤10 执行以下命令查找同步文件是否存在。

find /srv/BigData/ -name "sed*"

find /opt -name "sed*"

- 是，执行**步骤11**。
- 否，执行**步骤12**。

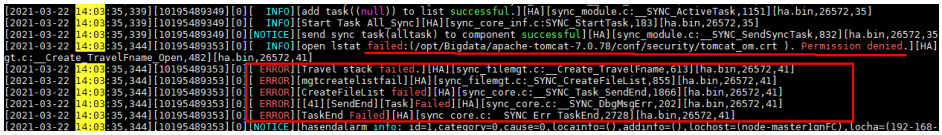
步骤11 执行以下命令，查看**步骤10**查找出的同步文件信息及权限。

ll *待查找文件路径*

- 如果文件大小为0，且权限栏全为“-”，则为垃圾文件，请执行以下命令删除。
  rm -rf *待删除文件*
  等待几分钟观察告警是否清除，如果未清除则执行**步骤12**。
- 如果文件大小不为0，则执行**步骤12**。

步骤12 查看发生告警时间段的日志文件。

1. 执行以下命令，进入当前集群的HA运行日志文件路径。
   cd /var/log/Bigdata/omm/oms/ha/runlog/
2. 解压并查看发生告警时间段的日志文件。
   例如，待查看文件名称为“ha.log.2021-03-22_12-00-07.gz”，则执行以下命令：
   gunzip *ha.log.2021-03-22_12-00-07.gz*
   vi *ha.log.2021-03-22_12-00-07*
   查看日志中，告警时间点前后是否有报错信息。
   - 是，根据相关报错信息进行处理。然后执行**步骤13**。
     例如，查询出报错信息如下，表示目录权限不足，则请修改对应目录权限与正常节点保持一致。

   - 否，执行**步骤14**。

步骤13 等待大约10分钟，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤14**。

收集故障信息。

步骤14 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤15 在“服务”中勾选如下节点信息，单击“确定”。

- OmmServer
- Controller
- NodeAgent

步骤16 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤17 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.8 ALM-12014 设备分区丢失

告警解释

系统按60秒周期进行扫描，如果检测到挂载服务目录的设备分区丢失（如由于设备拔出、设备离线、删除分区等原因）时，产生此告警。

此告警需要手动恢复。

告警属性

告警ID	告警级别	是否自动清除
12014	重要	否

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
挂载目录名	产生告警的挂载目录名。
设备分区名	产生告警的设备分区名。

对系统的影响

造成服务数据无法写入，业务系统运行不正常。

可能原因

- 硬盘被拔出。
- 硬盘离线、硬盘坏道等故障。

处理步骤

步骤1 打开FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的。

步骤2 从“定位信息”中获取“主机名”、“设备分区名”和“挂载目录名”。

步骤3 确认“主机名”节点的“设备分区名”对应的磁盘是否在对应服务器的插槽上。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 联系硬件工程师将故障磁盘在线拔出。

步骤5 以**root**用户登录发生告警的“主机名”节点，检查“/etc/fstab”文件中是否包含“挂载目录名”的行。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤6 执行**vi /etc/fstab**命令编辑文件，将包含“挂载目录名”的行删除。

步骤7 联系硬件工程师插入全新磁盘，具体操作请参考对应型号的硬件产品文档，如果原来故障的磁盘是RAID，那么请按照对应RAID卡的配置方法配置RAID。

步骤8 等待20～30分钟后执行**mount**命令（具体时间依赖磁盘的大小），检查磁盘是否已经挂载在目录“挂载目录名”上。

- 是，手动清除该告警，操作结束。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选“OmmServer”，单击“确定”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统不会自动清除此告警，需手工清除。

参考信息

无。

10.13.9 ALM-12015 设备分区文件系统只读

告警解释

系统按60秒周期进行扫描，如果检测到挂载服务目录的设备分区变为只读模式（如设备有坏扇区、文件系统存在故障等原因），则触发此告警。

系统如果检测到挂载服务目录的设备分区的只读模式消失（比如文件系统修复为读写模式、设备拔出、设备被重新格式化等原因），则告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12015	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
挂载目录名	产生告警的挂载目录名。
设备分区名	产生告警的设备分区名。

对系统的影响

造成服务数据无法写入，业务系统运行不正常。

可能原因

硬盘存在坏道等故障。

处理步骤

步骤1 打开FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的。

步骤2 从“定位信息”中获取“主机名”和“设备分区名”，其中“主机名”为故障告警的节点，“设备分区名”为故障磁盘的分区。

步骤3 联系硬件工程师确认为磁盘硬件故障之后，将服务器上故障磁盘在线拔出。

步骤4 拔出磁盘后系统会上报“ALM-12014 分区丢失”告警，参考**ALM-12014 设备分区丢失**进行处理，处理完成后，本告警即可自动消除。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.10 ALM-12016 CPU 使用率超过阈值

告警解释

系统每30秒周期性检测CPU使用率，并把实际CPU使用率和阈值相比较。CPU使用率默认提供一个阈值范围。当检测到CPU使用率连续多次（可配置，默认值为10）超出阈值范围时产生该告警。

平滑次数为1，CPU使用率小于或等于阈值时，告警恢复；平滑次数大于1，CPU使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12016	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。

参数名称	参数含义
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务进程响应缓慢或不可用。

可能原因

- 告警阈值配置或者平滑次数配置不合理。
- CPU配置无法满足业务需求，CPU使用率达到上限。

处理步骤

检查告警阈值配置或者平滑次数配置是否合理。

步骤1 基于实际CPU使用情况，修改告警阈值和平滑次数配置项。

登录FusionInsight Manager，根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > CPU > 主机CPU使用率”中更改告警的平滑次数，如图10-22所示。

说明

该选项的含义为告警检查阶段，“平滑次数”为连续检查多少次超过阈值，则发送告警。

图 10-22 设置告警平滑次数

在“主机CPU使用率”界面单击“操作”列的“修改”，更改告警阈值，如图10-23所示。

图 10-23 设置告警阈值

步骤2 等待2分钟，查看告警是否自动恢复。

- 是，处理完毕。
- 否，执行**步骤3**。

检查CPU使用率是否达到上限。

步骤3 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的节点地址。

步骤4 进入“主机”界面，单击告警的所在节点。

步骤5 在界面观察“主机CPU使用率”实时数据5分钟左右，若CPU使用率多次超过设置的阈值，请联系系统管理员提升CPU。

步骤6 检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“OmmServer”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.11 ALM-12017 磁盘容量不足

告警解释

系统每30秒周期性检测磁盘使用率，并把磁盘使用率和阈值相比较。磁盘使用率有一个默认阈值，当检测到磁盘使用率超过阈值时产生该告警。

平滑次数为1，主机磁盘某一分区使用率小于或等于阈值时，告警恢复；平滑次数大于1，主机磁盘某一分区使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12017	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
设备分区名	产生告警的磁盘分区。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务进程不可用。

可能原因

- 告警阈值配置不合理。
- 磁盘配置无法满足业务需求，磁盘使用率达到上限。

处理步骤

检查阈值设置是否合理。

步骤1 在FusionInsight Manager选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 磁盘 > 磁盘使用率”中查看该告警阈值是否不合理（默认90%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 磁盘 > 磁盘使用率”中单击“操作”列的“修改”更改告警阈值。如**图10-24**所示

图 10-24 设置告警阈值

步骤3 等待2分钟，查看告警是否消失。

- 是，处理完毕。
- 否，执行**步骤4**。

检查磁盘使用率是否达到上限

步骤4 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称和磁盘分区信息。

步骤5 以**root**用户登录告警所在节点。

步骤6 执行**df -lmPT | awk '$2 != "iso9660"' | grep '^/dev/' | awk '{"readlink -m "$1 | getline real }{$1=real; print $0}' | sort -u -k 1,1**命令，查看系统磁盘分区的使用信息。并通过**步骤4**中获取到的磁盘分区名称，查看该磁盘是否挂载在如下几个目录下：“/”、“/opt”、“/tmp”、“/var”、“/var/log”、“/srv/BigData”（可自定义）。

- 是，说明该磁盘为系统盘，执行**步骤10**。
- 否，说明该磁盘为非系统盘，执行**步骤7**。

步骤7 执行**df -lmPT | awk '$2 != "iso9660"' | grep '^/dev/' | awk '{"readlink -m "$1 | getline real }{$1=real; print $0}' | sort -u -k 1,1**命令，查看系统磁盘分区的使用信息。并通过**步骤4**中获取到的磁盘分区名称，判断该磁盘属于哪一个角色。

步骤8 查看磁盘所属服务。

MRS，是否为HDFS、Yarn、Kafka、Supervisor其中之一。

- 是，进行容量调整。执行**步骤9**。
- 否，执行**步骤12**。

步骤9 等待2分钟，查看告警是否消失。

- 是，处理完毕。
- 否，执行**步骤12**。

步骤10 执行命令**find / -xdev -size +500M -exec ls -l {} \;**，查看该节点上超过500MB的文件，查看该磁盘中，是否有误写入的大文件存在。

- 是，执行**步骤11**。
- 否，执行**步骤12**。

步骤11 处理该误写入的文件，并等待2分钟，查看告警是否清除。

- 是，执行完毕。
- 否，执行**步骤12**。

步骤12 联系系统管理员，对磁盘进行扩容。

步骤13 等待2分钟，查看告警是否消失。

- 是，处理完毕。
- 否，执行**步骤14**。

收集故障信息。

步骤14 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤15 在“服务”中勾选“OMS”，单击“确定”。

步骤16 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤17 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.12 ALM-12018 内存使用率超过阈值

告警解释

系统每30秒周期性检测内存使用率，并把实际内存使用率和阈值相比较。内存使用率默认提供一个阈值范围。当检测到内存使用率超出阈值范围时产生该告警。

平滑次数为1，主机内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，主机内存使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12018	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务进程响应缓慢或不可用。

可能原因

- 内存配置无法满足业务需求。内存使用率达到上限。
- SUSE 12.X操作系统中，系统自带的**free**命令版本较低，计算出的内存使用率不能如实反映真实的使用情况。

处理步骤

SUSE 12.X下处理方法。

步骤1 以**root**用户登录集群任意节点，执行**cat /etc/*-release**命令查看当前操作系统是否为SUSE 12.X。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 执行**cat /proc/meminfo | grep Mem**命令，查看当前操作系统内存实际使用情况。

```
MemTotal: 263576192 kB
MemFree: 198283116 kB
MemAvailable: 227641452 kB
```

步骤3 计算内存实际使用率，内存使用率 = 1 - (MemAvailable/MemTotal)。

- 若内存实际使用率低于90%，手动关闭监控转告警开关。
- 若内存实际使用率高于90%，则执行**步骤4**。

对系统进行扩容。

步骤4 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址。

步骤5 以**root**用户登录告警所在主机。

步骤6 若内存使用率超过阈值，对内存进行扩容。

步骤7 执行命令**free -m | grep Mem\: | awk '{printf("%s,", ($3-$6-$7) * 100 / $2)}'**，查看系统当前内存使用率。

步骤8 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选“OmmServer”，单击“确定”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.13 ALM-12027 主机 PID 使用率超过阈值

告警解释

系统每30秒周期性检测PID使用率，并把实际PID使用率和阈值进行比较，PID使用率默认提供一个阈值。当检测到PID使用率超出阈值时产生该告警。

平滑次数为1，主机PID使用率小于或等于阈值时，告警恢复；平滑次数大于1，主机PID使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12027	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

无法分配PID给新的业务进程，业务进程不可用。

可能原因

节点同时运行的进程过多，需要扩展pid_max值。

处理步骤

扩展pid_max值。

步骤1 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的，获取告警所在主机IP地址。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**cat /proc/sys/kernel/pid_max**，查看系统当前运行的PID最大值pid_max。

步骤4 若PID使用率超过阈值，将pid_max值增大一倍，执行命令**echo** *新pid_max值* **> /proc/sys/kernel/pid_max**。

示例：**echo 65536 > /proc/sys/kernel/pid_max**

步骤5 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选所有服务，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.14 ALM-12028 主机 D 状态进程数超过阈值

告警解释

系统每30秒周期性检测主机中omm用户D状态进程数，并把实际进程数和阈值相比较。主机D状态进程数默认提供一个阈值范围。当检测到进程数超出阈值范围时产生该告警。

平滑次数为1，主机中omm用户D状态进程数小于或等于阈值时，告警恢复；平滑次数大于1，主机中omm用户D状态进程数小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12028	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

占用系统资源，业务进程响应变慢。

可能原因

主机中正在等待的IO(磁盘IO、网络IO等)在较长时间内未得到响应，进程处于D状态。

处理步骤

查看D状态进程。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址。

步骤2 以**root**用户登录产生告警主机，执行**su - omm**命令，切换到**omm**用户。

步骤3 执行如下命令查看omm用户D状态进程号。

ps -elf | grep -v "\[thread_checkio\]" | awk 'NR!=1 {print $2, $3, $4}' | grep omm | awk -F' ' '{print $1, $3}' | grep D | awk '{print $2}'

步骤4 查看结果是否为空。

- 是，业务进程正常，执行**步骤6**。
- 否，执行**步骤5**。

步骤5 切换到**root**用户，执行**reboot**命令，重启产生告警主机（重启主机有风险，请确保重启后业务进程正常）。

步骤6 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“OMS”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.15 ALM-12033 慢盘故障

告警解释

系统每3秒执行一次**iostat**命令，监控磁盘I/O的系统指标，如果在300s内，svctm大于100ms且大于svctm_average值的1.5倍，则被认为是一个慢周期。若300s内慢周期的数量大于50%，则认为磁盘有问题，系统上报告警。

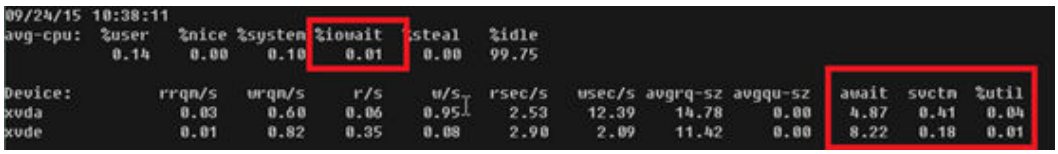
说明

svctm_average的值为当前节点中所有磁盘svctm的均值。

更换磁盘后，告警自动恢复。

当前慢盘故障告警的检查原理为：

在Linux平台上判断IO是否存在问题，输入命令**iostat -x -t 1**，观察下几个值（如图所示红色框中的部分）：



- %iowait：该值表示CPU等待IO的时间占整个CPU周期的百分比，如果该值超过50%，或者明显大于%system、%user以及%idle，这表示IO可能存在问题。
- await：该值表示该磁盘IO等待时间+IO服务时间的值，该值一般不超过20，其它DataNode数据盘可以稍高，但是不超过40。
- svctm：该值表示该磁盘IO服务时间。
- %util：该值表示磁盘繁忙程度，一般该值超过80%表示该磁盘可能处于繁忙状态。

如果%util大于10，并且svctm大于100，则记录，如果六十次里面有三十次都满足该条件，则发送慢盘故障。

告警属性

告警ID	告警级别	是否自动清除
12033	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
磁盘名	产生告警的磁盘名。

对系统的影响

磁盘慢盘故障，导致业务性能下降，阻塞业务的处理能力，严重时可能会导致服务不可用。

可能原因

磁盘老化或者磁盘坏道。

处理步骤

检查磁盘状态。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤2 查看该告警的详细信息，查看定位信息中“主机名”字段和“磁盘名”字段的值，获取该告警产生的故障磁盘信息。

步骤3 确认上报告警的节点是否为虚拟化环境。

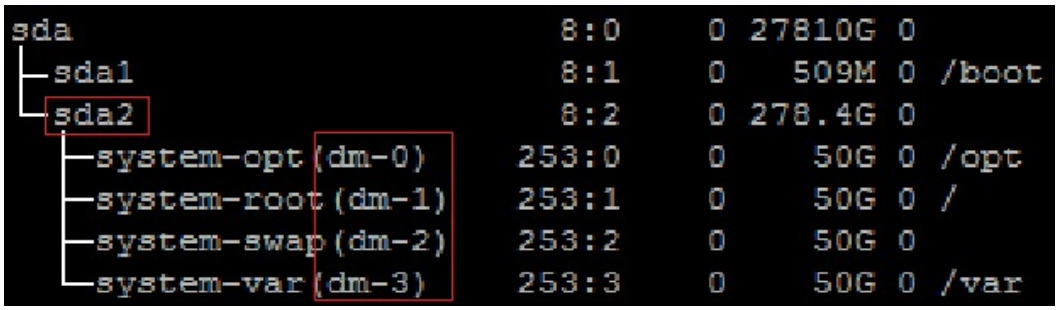
- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 请检查虚拟化环境提供的存储性能是否满足硬件要求，检查完毕之后执行**步骤5**。

步骤5 以**root**用户登录告警节点，执行**df -h**命令，查看输出内容是否包含“磁盘名”字段的值。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

步骤6 执行**lsblk**命令，是否可以查到“磁盘名”字段值与磁盘设备的映射关系。



- 是，执行**步骤7**。.
- 否，执行**步骤22**。

步骤7 以**root**用户登录上报告警的节点，执行**lsscsi | grep "/dev/sd[x]"**命令查看磁盘的设备信息，判断磁盘是否建立了RAID。

说明

其中**/dev/sd[x]**为**步骤2**中获取到的上报告警的磁盘名称。

例如执行：

lsscsi | grep "/dev/sda"

如果命令执行结果第三列显示ATA、SATA或者SAS，说明磁盘没有建立RAID；显示其他信息，则该磁盘可能建立了RAID。

- 是，执行**步骤12**。
- 否，执行**步骤8**。

步骤8 执行**smartctl -i /dev/sd[x]**命令检查硬件是否支持smart。

例如执行：

smartctl -i /dev/sda

如果命令执行结果中包含“SMART support is: Enabled”，表示磁盘支持smart；执行结果中包含“Device does not support SMART”或者其他，表示磁盘不支持smart。

- 是，执行**步骤9**。
- 否，执行**步骤17**。

步骤9 执行**smartctl -H --all /dev/sd[x]**命令查看smart的基本信息，判断磁盘是否正常。

例如执行：

smartctl -H --all /dev/sda

查看命令执行结果的“SMART overall-health self-assessment test result”内容，如果是“FAILED”，表示磁盘故障，需要更换；如果为“PASSED”，需要进一步看“Reallocated_Sector_Ct”或者“Elements in grown defect list”项的计数，如果大于100，则认为磁盘故障，需要更换。

- 是，执行**步骤10**。
- 否，执行**步骤18**。

步骤10 执行**smartctl -l error -H /dev/sd[x]**命令查看磁盘的GLIST列表，进一步继续判断磁盘是否正常。

例如执行：

smartctl -l error -H /dev/sda

查看命令执行结果的“Command/Featrue_name”列，如果出现“READ SECTOR(S)”或者“WRITE SECTOR(S)”表示磁盘有坏道；如果出现其他错误，表示磁盘电路板有问题。这两种错误均表示磁盘不正常，需要更换。

如果显示“No Errors Logged”，则表示没有错误日志，则可以触发磁盘smart自检。

- 是，执行**步骤11**。
- 否，执行**步骤18**。

步骤11 执行**smartctl -t long /dev/sd[x]**命令触发磁盘smart自检。命令执行后，会提示自检完成的时间，在等待自检完成后，重新执行**步骤9**和**步骤10**，检查磁盘是否正常。

例如执行：

smartctl -t long /dev/sda

- 是，执行**步骤17**。
- 否，执行**步骤18**。

步骤12 执行**smartctl -d [sat|scsi]+megaraid,[DID] -H --all /dev/sd[x]**命令检查硬件是否支持smart。

说明

- [sat|scsi]表示磁盘类型，需要尝试以上两种类型。
- [DID]表示槽位信息，需要尝试0～15。

例如依次执行：

smartctl -d sat+megaraid,0 -H --all /dev/sda

smartctl -d sat+megaraid,1 -H --all /dev/sda

smartctl -d sat+megaraid,2 -H --all /dev/sda

...

依次尝试不同磁盘类型和槽位信息的命令组合，如果执行结果中显示“SMART support is: Enabled”，表示磁盘支持smart，记录命令执行成功时磁盘类型和槽位信息组合参数；如果尝试完以上所有的命令组合，执行结果都未显示“SMART support is: Enabled”，表示磁盘不支持smart。

- 是，执行**步骤13**。
- 否，执行**步骤16**。

步骤13 执行**步骤12**中记录的**smartctl -d [sat|scsi]+megaraid,[DID] -H --all /dev/sd[x]**命令查看smart的基本信息，判断磁盘是否正常。

例如执行：

smartctl -d sat+megaraid,2 -H --all /dev/sda

查看命令执行结果的“SMART overall-health self-assessment test result”内容，如果是“FAILED”，表示磁盘故障，需要更换；如果为“PASSED”，需要进一步看“Reallocated_Sector_Ct”或者“Elements in grown defect list”项的计数，如果大于100，则认为磁盘故障，需要更换。

- 是，执行**步骤14**。
- 否，执行**步骤18**。

步骤14 执行**smartctl -d [sat|scsi]+megaraid,[DID] -l error -H /dev/sd[x]**命令查看硬盘的GLIST列表，进一步判断硬盘是否正常。

例如执行：

smartctl -d sat+megaraid,2 -l error -H /dev/sda

查看命令执行结果的“Command/Featrue_name”列，如果出现“READ SECTOR(S)”或者“WRITE SECTOR(S)”表示磁盘有坏道；如果出现其他错误，表示磁盘电路板有问题。这两种错误均表示磁盘不正常，需要更换。

如果显示“No Errors Logged”，则表示没有错误日志，则可以触发磁盘smart自检。

- 是，执行**步骤15**。
- 否，执行**步骤18**。

步骤15 执行**smartctl -d [sat|scsi]+megaraid,[DID] -t long /dev/sd[x]**命令触发磁盘smart自检。命令执行后，会提示自检完成的时间，在等待自检完成后，重新执行**步骤13**和**步骤14**，检查磁盘是否正常。

例如执行：

smartctl -d sat+megaraid,2 -t long /dev/sda

- 是，执行**步骤17**。
- 否，执行**步骤18**。

步骤16 磁盘不支持smart，通常是因为配置的RAID卡不支持，此时需要使用对应RAID卡厂商的检查工具进行处理，然后执行**步骤17**。

例如LSI一般是MegaCli工具。

步骤17 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”，单击该告警操作列的“清除”，并继续观察该告警，查看同一块磁盘的告警是否会继续上报。

如果当前磁盘出现三次以上该告警，建议用户更换磁盘。

- 是，执行**步骤18**。
- 否，操作结束。

更换磁盘。

步骤18 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤19 查看该告警的详细信息，查看定位信息中对应的“主机名”字段和“磁盘名”字段的值，获取该告警上报的故障磁盘信息。

步骤20 更换硬盘。

步骤21 检查告警是否清除。

- 是，操作结束。
- 否，执行**步骤22**。

收集故障信息。

步骤22 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤23 在“服务”中勾选“OMS”，单击“确定”。

步骤24 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤25 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.16 ALM-12034 周期备份任务失败

告警解释

系统每60分钟执行周期备份任务，如果周期备份任务执行失败，则上报该告警，如果下次备份执行成功，则恢复告警。

告警属性

告警ID	告警级别	是否自动清除
12034	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
任务名	任务名称。

对系统的影响

周期备份任务失败，可能会导致长时间没有可用的备份包，在系统出现异常时，无法恢复。

可能原因

该告警产生原因依赖于该任务的详细情况，直接获取日志和任务详情来处理该告警。

处理步骤

查看磁盘空间是否不足。

步骤1 在FusionInsight Manager管理界面，选择“运维 > 告警 > 告警”。

步骤2 在告警列表中单击该告警的，从“定位信息”处获得“任务名”。

步骤3 选择“运维 >备份恢复 > 备份管理”。

步骤4 根据“任务名”查找对应备份任务，单击“操作”栏下的“更多”按钮，在弹出的窗口中单击“查询历史”按钮，查看备份任务的详细信息。

步骤5 在弹出的日志详情窗口中，单击，查看是否有“Failed to backup xx due to insufficient disk space, move the data in the /srv/BigData/LocalBackup directory to other directories.”的信息。

- 是，执行**步骤6**。
- 否，执行**步骤13**。

步骤6 单击“备份路径”下的“查看”，获取备份路径。

步骤7 以**root**用户登录节点，执行以下命令查看节点挂载详情：

df -h

步骤8 在挂载详情中查看备份路径挂载点的剩余空间是否小于20GB。

- 是，执行**步骤9**。
- 否，执行**步骤13**。

步骤9 查看备份目录下是否有很多备份包。

- 是，执行**步骤10**。
- 否，执行**步骤13**。

步骤10 将备份包移出备份目录，或者直接删除备份包，直到备份目录挂载节点剩余空间大于20GB。

步骤11 再一次启动该备份任务，查看备份任务是否执行成功。

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 等待2分钟，检查告警是否消除。

- 是，结束操作。
- 否，执行**步骤13**。

收集故障信息

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选“Controller”，单击“确定”。

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.17 ALM-12035 恢复任务失败后数据状态未知

告警解释

执行恢复任务失败后，系统按60分钟周期自动回滚，如果回滚失败，可能会导致数据丢失等问题，如果该情况出现，则上报告警，如果下一次该任务恢复成功，则恢复告警。

告警属性

告警ID	告警级别	是否自动清除
12035	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
任务名	任务名称。

对系统的影响

执行恢复任务失败后，系统会自动回滚，如果回滚失败，可能会导致数据丢失，数据状态未知等问题，有可能会影响业务功能。

可能原因

该告警产生原因可能是执行恢复任务前组件状态不满足要求或执行恢复任务中某个步骤出错，执行恢复任务中出错依赖于该任务的详细情况，可以获取日志和任务详情来处理该告警。

处理步骤

查看组件状态

步骤1 在FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务”，查看组件当前的运行状态是否满足要求（OMS、DBService要求状态正常，其他组件要求停止服务）：

- 是，执行**步骤9**。
- 否，执行**步骤2**。

步骤2 恢复组件状态至要求状态，再一次启动该恢复任务。

步骤3 登录FusionInsight Manager管理界面，选择“运维 > 告警 > 告警”。

步骤4 在告警列表中单击该告警所在行的，从“定位信息”处获得任务名。

步骤5 选择“运维 > 备份恢复 > 恢复管理”。

步骤6 根据“任务名”查找对应恢复任务，查看恢复任务的详细信息。

步骤7 启动该恢复任务，查看恢复任务是否执行成功。

- 是，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 等待2分钟，检查告警是否消除。

- 是，结束操作。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选“Controller”，单击“确定”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.18 ALM-12038 监控指标转储失败

告警解释

用户在FusionInsight Manager界面配置监控指标转储后，系统按转储时间间隔（默认60秒）周期性检测监控指标转储结果，转储失败时产生该告警。

转储成功后，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12038	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

监控指标转储失败会影响上层管理系统无法获取到FusionInsight Manager系统的监控指标。

可能原因

- 无法连接服务器。
- 无法访问服务器上保存路径。
- 上传监控指标文件失败。

处理步骤

查看服务器连接是否正常。

步骤1 查看FusionInsight Manager系统与服务器网络连接是否正常。

- 是，执行**步骤3**。
- 否，执行**步骤2**。

步骤2 联系网络管理员恢复网络连接，然后检查告警是否恢复。

- 是，执行完毕。
- 否，执行**步骤3**。

步骤3 选择“系统 > 对接 > 监控数据上传”，查看监控数据上传页面配置的FTP用户名、密码、端口、转储模式、公钥是否与服务器端配置一致。

- 是，执行**步骤5**。
- 否，执行**步骤4**。

步骤4 填入正确的配置信息，然后单击“确定”，检查告警是否恢复。

- 是，执行完毕。

- 否，执行**步骤5**。

查看服务器端保存路径权限是否正常。

步骤5 选择“系统 > 对接 > 监控数据上传”，查看“FTP用户名”、“保存路径”和“转储模式”配置项。

- 是FTP模式，执行**步骤6**。
- 是SFTP模式，执行**步骤7**。

步骤6 以FTP方式登录服务器，在默认目录下查看相对路径“保存路径”是否有“FTP用户名”的读写权限。

- 是，执行**步骤9**。
- 否，执行**步骤8**。

步骤7 以SFTP方式登录服务器，查看绝对路径“保存路径”是否有“FTP用户名”的读写权限。

- 是，执行**步骤9**。
- 否，执行**步骤8**。

步骤8 增加读写权限，然后检查告警是否恢复。

- 是，执行完毕。
- 否，执行**步骤9**。

查看服务器端保存路径是否有足够磁盘空间。

步骤9 登录服务器端，查看当前保存路径下是否有足够磁盘空间。

- 是，执行**步骤11**。
- 否，执行**步骤10**。

步骤10 删除多余文件，或在监控指标转储配置页面更改保存目录。然后检查告警是否恢复。

- 是，执行完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选“OMS”，单击“确定”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后60分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.19 ALM-12039 OMS 数据库主备不同步

告警解释

OMS数据库主备不同步，系统每10秒检查一次主备数据同步状态，如果连续30次查不到同步状态，或者同步状态异常，产生告警。

当主备数据同步状态正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12039	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

OMS数据库主备不同步，如果此时主实例异常，会出现数据丢失或者数据异常的情况。

可能原因

- 主备节点网络不稳定。
- 备OMS数据库异常。
- 备节点磁盘空间满。

处理步骤

检查主备节点网络是否正常。

步骤1 在FusionInsight Manager界面上选择“运维 > 告警 > 告警”，在告警列表中，单击此告警所在行的，查看该告警的OMS数据库备节点IP地址。

步骤2 以**root**用户登录主OMS数据库节点。

步骤3 执行**ping** *备OMS数据库心跳IP地址*命令检查备OMS数据库节点是否可达。

- 是，执行**步骤6**。

- 否，执行**步骤4**。

步骤4 联系网络管理员查看是否为网络故障。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 修复网络故障，然后查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查备OMS数据库状态是否正常。

步骤6 以**root**用户登录备OMS数据库节点。

步骤7 执行**su - omm**命令切换到**omm**用户。

步骤8 进入“${BIGDATA_HOME}/om-server/om/sbin/”目录，然后执行**./status-oms.sh**命令检查备OMS数据库资源状态是否正常，查看回显中，“ResName”为“gaussDB”的一行，是否显示如下信息：

例如：

```
10_10_10_231 gaussDB Standby_normal Normal Active_standby
```

- 是，执行**步骤9**。
- 否，执行**步骤16**。

检查备节点磁盘是否已满。

步骤9 以**root**用户登录备OMS数据库节点。

步骤10 执行**su - omm**命令切换到**omm**用户。

步骤11 执行**echo ${BIGDATA_DATA_HOME}/dbdata_om**命令获取OMS数据库的数据目录。

步骤12 执行**df -h**命令，查看系统磁盘分区的使用信息。

步骤13 查看OMS数据库数据目录挂载磁盘是否已满。

- 是，执行**步骤14**。
- 否，执行**步骤16**。

步骤14 进行磁盘扩容。

步骤15 磁盘扩容后，等待2分钟检查告警是否清除。

- 是，操作结束。
- 否，执行**步骤16**。

收集故障信息。

步骤16 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤17 在“服务”中勾选“OmmServer”，单击“确定”。

步骤18 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤19 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.20 ALM-12040 系统熵值不足

告警解释

每天零点系统检查熵值，每次检查都连续检查五次，首先检查是否启用并正确配置了rng-tools工具或者haveged工具，如果没有配置，则继续检查当前熵值，如果五次均小于100，则上报故障告警。

当检查到真随机数方式已经配置或者伪随机数方式中配置了随机数参数或者没有配置但是五次检查中，至少有一次熵值大于等于100，则告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12040	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

影响系统正常运行。

可能原因

haveged服务或者rngd服务异常。

处理步骤

检查并手动配置系统熵值。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤2 查看该“定位信息”中对应的“主机名”字段值，获取该告警产生的主机名。

步骤3 以**root**用户登录告警所在节点。

步骤4 执行**/bin/rpm -qa | grep -w "haveged"**命令查看haveged安装情况，观察命令返回结果是否为空。

- 是，执行**步骤7**。
- 否，执行**步骤5**。

步骤5 执行**/sbin/service haveged status |grep "running"**，查看返回结果。

- 如果执行成功，表示haveged服务安装并正常配置运行，执行**步骤10**。
- 如果执行不成功，表示haveged服务没有正常运行。执行**步骤7**。

步骤6 执行**/bin/rpm -qa | grep -w "rng-tools"**命令，查看rng-tools安装情况，观察命令返回结果是否为空。

- 是，执行**步骤8**。
- 否，执行**步骤7**。

步骤7 执行**ps -ef | grep -v "grep" | grep rngd | tr -d " " | grep "\-o/dev/random" | grep "\-r/dev/urandom"**，查看返回结果。

- 如果执行成功，表示rngd服务安装并正常配置运行，执行**步骤10**。
- 如果执行不成功，表示rngd服务并没有正常运行，执行**步骤8**

步骤8 手动配置系统熵值设置，设置方法参见**参考信息**。

步骤9 等待第二天零点，系统下一次熵值检查，查看告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤10**。

收集故障信息。

步骤10 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

手动检查操作系统熵值

以**root**用户登录节点，执行**cat /proc/sys/kernel/random/entropy_avail**命令，检查操作系统熵值是否满足集群的安装要求（不低于500）。如果低于500，可使用以下两种方式之一进行配置：

- 使用“haveged”工具（真随机数方式）：请联系OS供应商安装并启动该工具。
- 使用“rng-tools”工具（伪随机数方式）：请联系OS供应商安装该工具，并根据操作系统类型进行配置：
  - Red Hat和CentOS下：执行以下命令进行配置：

    echo 'EXTRAOPTIONS="-r /dev/urandom -o /dev/random -t 1 -i"' >> /etc/sysconfig/rngd

    service rngd start

    chkconfig rngd on
  - SUSE下：执行以下命令进行配置：

    rngd -r /dev/urandom -o /dev/random

    echo "rngd -r /dev/urandom -o /dev/random" >> /etc/rc.d/after.local

10.13.21 ALM-12041 关键文件权限异常

告警解释

系统每隔5分钟检查一次系统中关键目录或者文件权限、用户、用户组是否正常，如果不正常，则上报故障告警。

当检查到权限等均正常，则告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12041	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
路径名	异常的文件路径或者名称。

对系统的影响

导致系统功能不可用。

可能原因

用户手动修改了文件权限、用户和用户组等信息或者系统异常下电等原因导致文件权限异常或文件丢失。

处理步骤

检查异常文件是否存在及异常文件的权限是否正确。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤2 查看该告警“定位信息”中对应的“主机名”字段值，获取该告警产生的主机名，查看定位信息中对应的“路径名”字段值，获取异常的文件路径或者名称。

步骤3 以**root**用户登录告警所在节点。

步骤4 执行**ll** *路径名*命令，其中*路径名*为**步骤2**获取到的异常文件，获取到该文件或者目录在主机上的当前的用户，权限，用户组等信息。

步骤5 进入“${BIGDATA_HOME}/om-agent/nodeagent/etc/agent/autocheck”目录，然后执行**vi keyfile**命令，并搜索对应的异常文件名，可以看到该文件的正确权限。

说明

除keyfile中所列出的文件和目录外，为保证主备OMS配置同步正常，“$OMS_RUN_PATH/workspace/ha/module/hasync/plugin/conf/filesync.xml”中配置的文件、目录以及目录下的文件和子目录也会被监控，其中文件要求omm用户具有可读写权限，目录要求omm用户具有可读和可执行权限。

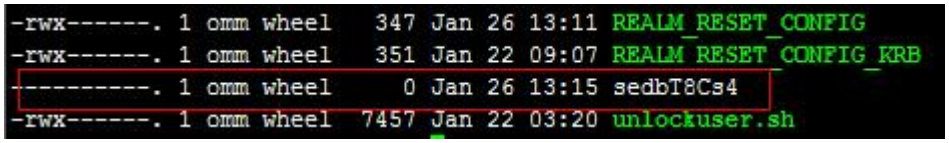
步骤6 对比当前主机上该文件的真实权限和**步骤5**中获取到的文件应有权限，对该文件进行正确的权限和用户，用户组信息的修改。

步骤7 等待一个小时，进入下一次检查，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤8**。

说明

如果集群安装目录所在磁盘分区已满，部分程序安装目录会由于sed命令执行失败，产生一些临时文件，且没有读写可执行权限。如果这些文件产生在该告警的监控范围内，那么系统会上报该告警，告警原因可以看到是由于产生的临时文件权限异常导致，可以参照上述告警处理流程处理该告警，或者确认权限异常文件为临时文件后，可以直接删除。sed命令产生的临时文件类似于下图。



收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.22 ALM-12042 关键文件配置异常

告警解释

系统每隔5分钟检查一次系统中关键的配置是否正确，如果不正常，则上报故障告警。

当检查到配置正确时，则告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12042	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
路径名	异常的文件路径或者名称。

对系统的影响

导致文件所属服务功能不正常。

可能原因

用户手动修改了文件配置或者系统异常下电等原因。

处理步骤

检查异常文件配置。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤2 查看该告警“定位信息”中对应的“主机名”字段值，获取该告警产生的主机名，查看定位信息中对应的“路径名”字段值，获取异常的文件路径或者名称。

步骤3 以**root**用户登录告警所在节点。

步骤4 查看日志文件“$BIGDATA_LOG_HOME/nodeagent/scriptlog/checkfileconfig.log”，根据错误日志分析原因。在**参考信息**中查找该文件的检查标准，并对照检查标准对文件进行进一步的手动检查和修改。

执行**vi** *文件名*命令进入编辑模式，按“Insert”键开始编辑。

修改完成后按“Esc”键退出编辑模式，并输入**:wq**保存退出。

例如：

vi /etc/ssh/sshd_config

步骤5 等待一个小时，进入下一次检查，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

- **/etc/fstab检查文件的检查标准**

  检查“/etc/fstab”文件中配置的分区，是否在“/proc/mounts”中能找到。

  检查在“fstab”中配置的swap分区，是否和“/proc/swaps”一一对应。

- **/etc/hosts检查文件的检查标准**

  通过命令**cat /etc/hosts**查看是否存在以下几种情况，如果是，则说明该配置文件配置异常。

  a. “/etc/hosts”文件不存在。
  b. 该主机的主机名不在文件中配置。
  c. 该主机名对应的IP不唯一。
  d. 该主机名对应的IP在**ifconfig**命令下的回显列表中不存在。
  e. 该文件中存在一个IP对应多个主机名的情况。

- **/etc/ssh/sshd_config检查文件的检查标准**

  通过命令**vi /etc/ssh/sshd_config**查看下面几个配置项是否正确。

  a. “UseDNS”项必须配置为“no”。

  b. “MaxStartups”必须配置为大于等于1000。

  c. “PasswordAuthentication”和“ChallengeResponseAuthentication”两个配置项中必须至少有一项没有配置或者至少有一项配置为“yes”。

10.13.23 ALM-12045 网络读包丢包率超过阈值

告警解释

系统每30秒周期性检测网络读包丢包率，并把实际丢包率和阈值（系统默认阈值0.5%）进行比较，当检测到网络读包丢包率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读包丢包率”修改阈值。

平滑次数为1，网络读包丢包率小于或等于阈值时，告警恢复；平滑次数大于1，网络读包丢包率小于或等于阈值的90%时，告警恢复。

该告警检测默认关闭。若需要开启，请根据“检查系统环境”步骤，确认该系统是否可以开启该告警发送。

告警属性

告警ID	告警级别	是否自动清除
12045	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
网口名	产生告警的网口名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务性能下降或者个别业务出现超时问题。

风险提示：在SUSE内核版本3.0以上或Red Hat 7.2版本，由于系统内核修改了网络读包丢包数的计数机制，在该系统下，即使网络正常运行，也可能会导致该告警出现，对业务无影响，建议优先按照“检查系统环境”进行排查。

可能原因

- 操作系统问题。
- 网卡配置了主备bond模式。
- 告警阈值配置不合理。
- 客户网络环境质量差。

处理步骤

查看网络丢包率

步骤1 打开FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，查看该告警的主机名称和网卡名称。

步骤2 以**omm**用户登录该告警所在节点，执行**/sbin/ifconfig** *网卡名称*命令检查网络中是否存在丢包。

```
omm@8-5-192-4:~> /sbin/ifconfig eth2
eth2      Link encap:Ethernet  HWaddr E4:35:C8:7B:B5:48
          inet addr:192.168.192.4  Bcast:192.168.255.255  Mask:255.255.0.0
          inet6 addr: fe80::e635:c8ff:fe7b:b548/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:5254854 errors:0 dropped:214676 overruns:0 frame:0
          TX packets:329443 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:354839633 (338.4 Mb)  TX bytes:25083094 (23.9 Mb)
```

说明

- 告警节点IP地址：根据告警定位信息中的“主机名”字段值，在FusionInsight Manager的“主机”查询对应的IP地址，管理平面和业务平面IP都需要检查。
- 丢包率=（dropped个数/RX packets总个数）*100%，如果丢包率大于该指标所设置的系统阈值（系统默认阈值0.5%），则认为网络读包存在丢包现象。

- 是，执行**步骤11**。
- 否，执行**步骤3**。

检查系统环境

步骤3 以**omm**用户登录主OMS节点或者告警所在节点。

步骤4 执行**cat /etc/*-release**命令，确认操作系统的类型。

- Red Hat，执行**步骤5**。
  ```
  # cat /etc/*-release
  Red Hat Enterprise Linux Server release 7.2 (Santiago)
  ```
- SUSE，执行**步骤6**。
  ```
  # cat /etc/*-release
  SUSE Linux Enterprise Server 11 (x86_64)
  VERSION = 11
  PATCHLEVEL = 3
  ```
- 其他，执行**步骤11**。

步骤5 执行**cat /etc/redhat-release**命令，查询操作系统版本是否为Red Hat 7.2（x86）或者Red Hat 7.4（TaiShan）。

```
# cat /etc/redhat-release
Red Hat Enterprise Linux Server release 7.2 (Santiago)
```

- 是，不能开启告警发送，执行**步骤7**。
- 否，执行**步骤11**。

步骤6 执行**cat /proc/version**命令，查询SUSE内核版本是否为3.0及以上。

```
# cat /proc/version
Linux version 3.0.101-63-default (geeko@buildhost) (gcc version 4.3.4 [gcc-4_3-branch revision 152973] (SUSE Linux) ) #1 SMP Tue Jun 23 16:02:31 UTC 2015 (4b89d0c)
```

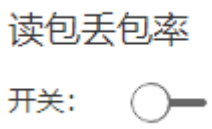
- 是，不能开启告警发送，执行**步骤7**。
- 否，执行**步骤11**。

步骤7 登录FusionInsight Manager，进入“运维 > 告警 > 阈值设置”页面。

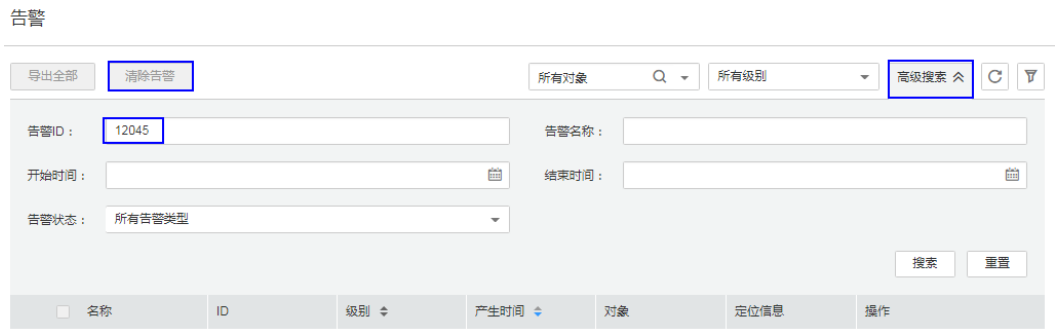
步骤8 在“阈值设置”页面左侧树形结构中选择“*待操作集群名称*>主机 > 网络读信息 > 读包丢包率”，查看发送告警开关指示是否打开。

- 是，说明开启了告警发送，执行**步骤9**。
- 否，已经关闭告警发送，执行**步骤10**。

步骤9 关闭该告警“开关”开关，屏蔽对“网络读包丢包率超过阈值”的检测，操作后的结果如下图所示。



步骤10 在FusionInsight Manager的“告警”界面，搜索“12045”告警，将未自动清除的该告警全部手动清除，处理完毕。



说明

“网络读包丢包率超过阈值”的告警ID是12045。

检查网卡是否配置了主备bond模式。

步骤11 以**omm**用户登录告警所在节点，执行**ls -l /proc/net/bonding**命令，查看该节点是否存在“/proc/net/bonding”目录。

- 是，如下所示，则该节点配置了bond模式，执行**步骤12**。

```
# ls -l /proc/net/bonding/
total 0
-r--r--r-- 1 root root 0 Oct 11 17:35 bond0
```

- 否，如下所示，则该节点未配置bond模式，执行**步骤14**。

```
# ls -l /proc/net/bonding/
ls: cannot access /proc/net/bonding/: No such file or directory
```

步骤12 执行**cat /proc/net/bonding/***bond0*命令，查看配置文件中**Bonding Mode**参数的值是否为**fault-tolerance**。

说明

bond0为bond配置文件名称，请以**步骤11**查询出的文件名称为准。

```
# cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v3.7.1 (April 27, 2011)

Bonding Mode: fault-tolerance (active-backup)
Primary Slave: eth1 (primary_reselect always)
Currently Active Slave: eth1
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 0
Down Delay (ms): 0

Slave Interface: eth0
MII Status: up
Speed: 1000 Mbps
Duplex: full
Link Failure Count: 1
Slave queue ID: 0

Slave Interface: eth1
MII Status: up
Speed: 1000 Mbps
Duplex: full
Link Failure Count: 1
Slave queue ID: 0
```

- 是，该环境的网卡为主备bond模式，执行**步骤13**。
- 否，执行**步骤14**。

步骤13 检查该告警中**NetworkCardName**参数对应的网卡是否为备网卡。

- 是，备网卡的告警无法自动恢复，请在告警管理页面手动清除该告警，处理完毕。
- 否，执行**步骤14**。

说明

备网卡判断方式：查看配置文件**/proc/net/bonding/bond0**，**NetworkCardName**参数对应的网卡名称等于其中一个**Slave Interface**，但是不等于**Currently Active Slave**（当前主网卡），则该网卡为备网卡。

检查阈值设置是否合理。

步骤14 登录FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读包丢包率”，查看该告警阈值是否合理（默认0.5%为合理值，用户可以根据自己的实际需求调整）。

- 是，执行**步骤17**。
- 否，执行**步骤15**。

步骤15 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读包丢包率”，单击“操作”列的“修改”，更改告警阈值。如图10-25所示。

图 10-25 设置告警阈值

步骤16 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤17**。

检查网络是否异常。

步骤17 联系系统管理员，检查网络是否存在异常。

- 是，恢复网络故障，执行**步骤18**。
- 否，执行**步骤19**。

步骤18 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤19**。

收集故障信息。

步骤19 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤20 在“服务”中勾选“OMS”，单击“确定”。

步骤21 设置“主机”为告警所在节点和主OMS节点。

步骤22 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤23 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.24 ALM-12046 网络写包丢包率超过阈值

告警解释

系统每30秒周期性检测网络写包丢包率，并把实际丢包率和阈值（系统默认阈值0.5%）进行比较，当检测到网络写包丢包率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写包丢包率”修改阈值。

平滑次数为1，网络写包丢包率小于或等于阈值时，告警恢复；平滑次数大于1，网络写包丢包率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12046	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
网口名	产生告警的网口名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务性能下降或者个别业务出现超时问题。

可能原因

- 告警阈值配置不合理。
- 客户网络环境质量差。

处理步骤

检查阈值设置是否合理。

步骤1 在FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写包丢包率”，查看该告警阈值是否合理（默认0.5%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写包丢包率”，单击“操作”列的“修改”更改告警阈值。

如**图10-26**所示：

图 10-26 设置告警阈值

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查网络是否异常。

步骤4 联系系统管理员，检查网络是否存在异常。

- 是，恢复网络故障，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“OMS”，单击“确定”。

步骤8 设置“主机”为告警所在节点和主OMS节点。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.25 ALM-12047 网络读包错误率超过阈值

告警解释

系统每30秒周期性检测网络读包错误率，并把实际错误率和阈值（系统默认阈值0.5%）进行比较，当检测到网络读包错误率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读包错误率”修改阈值。

平滑次数为1，网络读包错误率小于或等于阈值时，告警恢复；平滑次数大于1，网络读包错误率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12047	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
网口名	产生告警的网口名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

通信闪断，业务超时。

可能原因

- 告警阈值配置不合理。
- 客户网络环境质量差。

处理步骤

检查阈值设置是否合理。

步骤1 在FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读包错误率”，查看该告警阈值是否合理（默认0.5%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读包错误率”，单击“操作”列的“修改”更改告警阈值。

如**图10-27**所示：

图 10-27 设置告警阈值

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查网络是否异常。

步骤4 联系系统管理员，检查网络是否存在异常。

- 是，恢复网络故障，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“OMS”，单击“确定”。

步骤8 设置“主机”为告警所在节点和主OMS节点。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.26 ALM-12048 网络写包错误率超过阈值

告警解释

系统每30秒周期性检测网络写包错误率，并把实际错误率和阈值（系统默认阈值0.5%）进行比较，当检测到网络写包错误率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写包错误率”修改阈值。

平滑次数为1，网络写包错误率小于或等于阈值时，告警恢复；平滑次数大于1，网络写包错误率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12048	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
网口名	产生告警的网口名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

通信闪断，业务超时。

可能原因

- 告警阈值配置不合理。
- 客户网络环境质量差。

处理步骤

检查阈值设置是否合理。

步骤1 在FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写包错误率”，查看该告警阈值是否合理（默认0.5%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写包错误率”，单击“操作”列的“修改”更改告警阈值。

如**图10-28**所示：

图 10-28 设置告警阈值

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查网络是否异常。

步骤4 联系系统管理员，检查网络是否存在异常。

- 是，恢复网络故障，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“OMS”，单击“确定”。

步骤8 设置“主机”为告警所在节点和主OMS节点。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.27 ALM-12049 网络读吞吐率超过阈值

告警解释

系统每30秒周期性检测网络读吞吐率，并把实际吞吐率和阈值（系统默认阈值80%）进行比较，当检测到网络读吞吐率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读吞吐率”修改阈值。

平滑次数为1，网络读吞吐率小于或等于阈值时，告警恢复；平滑次数大于1，网络读吞吐率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12049	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
网口名	产生告警的网口名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务系统运行不正常或不可用。

可能原因

- 告警阈值配置不合理。
- 网口速率不满足当前业务需求。

处理步骤

检查阈值设置是否合理。

步骤1 在FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读吞吐率”，查看该告警阈值是否不合理（默认80%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络读信息 > 读吞吐率”，单击“操作”列的“修改”更改告警阈值。

如**图10-29**所示：

图 10-29 设置告警阈值

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查网口速率是否满足需求。

步骤4 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的，获取告警所在主机地址及网口名称。

步骤5 以**root**用户登录告警所在主机。

步骤6 执行命令**ethtool** *网口名称*，查看当前网口速率最大值Speed。

说明

对于虚拟机环境，通过命令可能无法查询到网口速率，建议直接联系系统管理确认网口速率是否满足需求。

步骤7 若网络读吞吐率超过阈值，直接联系系统管理员，提升网口速率。

步骤8 检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选“OMS”，单击“确定”。

步骤11 设置“主机”为告警所在节点和主OMS节点。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.28 ALM-12050 网络写吞吐率超过阈值

告警解释

系统每30秒周期性检测网络写吞吐率，并把实际吞吐率和阈值（系统默认阈值80%）进行比较，当检测到网络写吞吐率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写吞吐率”修改阈值。

平滑次数为1，网络写吞吐率小于或等于阈值时，告警恢复；平滑次数大于1，网络写吞吐率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12050	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
网口名	产生告警的网口名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

业务系统运行不正常或不可用。

可能原因

- 告警阈值配置不合理。
- 网口速率不满足当前业务需求。

处理步骤

检查阈值设置是否合理。

步骤1 在FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写吞吐率”，查看该告警阈值是否合理（默认80%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络写信息 > 写吞吐率”，单击“操作”列的“修改”更改告警阈值。

如**图10-30**所示：

图 10-30 设置告警阈值

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查网口速率是否满足需求。

步骤4 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的，获取告警所在主机地址及网口。

步骤5 以**root**用户登录告警所在主机。

步骤6 执行命令**ethtool** *网口名称*，查看当前网口速率最大值Speed。

说明

对于虚拟机环境，通过命令可能无法查询到网口速率，建议直接联系系统管理确认网口速率是否满足需求。

步骤7 若网络写吞吐率超过阈值，直接联系系统管理员，提升网口速率。

步骤8 检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选“OMS”，单击“确定”。

步骤11 设置“主机”为告警所在节点和主OMS节点。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.29 ALM-12051 磁盘 Inode 使用率超过阈值

告警解释

系统每30秒周期性检测磁盘Inode使用率，并把实际Inode使用率和阈值（系统默认阈值80%）进行比较，当检测到Inode使用率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > *磁盘* > *磁盘inode使用率*”修改阈值。

平滑次数为1，磁盘Inode使用率小于或等于阈值时，告警恢复；平滑次数大于1，磁盘Inode使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12051	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
设备分区	产生告警的磁盘分区。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

文件系统无法正常写入。

可能原因

磁盘写入的小文件过多。

处理步骤

磁盘写入的小文件过多。

步骤1 打开FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，获取告警所在主机地址和磁盘分区。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**df -i | grep -iE "***分区名称***Filesystem"**，查看磁盘当前Inode使用率。

```
# df -i | grep -iE "xvda2|Filesystem"
Filesystem            Inodes   IUsed   IFree IUse% Mounted on
/dev/xvda2            2359296  207420 2151876    9% /
```

步骤4 若Inode使用率超过阈值，手工排查该分区存在的小文件，确认是否能够删除这些文件。

说明

可使用命令**for i in /*; do echo $i; find $i|wc -l; done**查看分区下的文件个数，使用时请替换“/*”为需要检查的分区。

```
# for i in /srv/*; do echo $i; find $i|wc -l; done
/srv/BigData
```

```
4284
/srv/ftp
1
/srv/www
13
```

- 是，执行**rm -rf** *待删除文件或文件夹路径*命令，删除文件，执行**步骤5**。

  说明

  删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。

- 否，进行磁盘扩容，执行**步骤5**。

步骤5 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“OMS”，单击“确定”。

步骤8 设置“主机”为告警所在节点和主OMS节点。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.30 ALM-12052 TCP 临时端口使用率超过阈值

告警解释

系统每30秒周期性检测TCP临时端口使用率，并把实际使用率和阈值（系统默认阈值80%）进行比较，当检测到TCP临时端口使用率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 网络状态 > TCP临时端口使用率”修改阈值。

平滑次数为1，TCP临时端口使用率小于或等于阈值时，告警恢复；平滑次数大于1，TCP临时端口使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12052	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

主机上业务无法发起对外建立连接，业务中断。

可能原因

- 临时端口不满足当前业务需求。
- 系统环境异常。

处理步骤

扩大临时端口范围。

步骤1 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的，获取告警所在主机IP地址。

步骤2 以**omm**用户登录告警所在主机。

步骤3 执行**cat /proc/sys/net/ipv4/ip_local_port_range |cut -f 1**命令，获得开始端口值，执行**cat /proc/sys/net/ipv4/ip_local_port_range |cut -f 2**命令，获得结束端口值，相减得到临时端口总数，若临时端口总数小于28232，说明操作系统随机端口范围太小，需要联系系统管理员扩大端口范围。

步骤4 执行命令**ss -ant 2>/dev/null | grep -v LISTEN | awk 'NR > 2 {print $4}'|cut -d ':' -f 2 | awk '$1 >"*开始端口值*" {print $1}' | sort -u | wc -l**，计算临时端口使用数。

步骤5 使用公式计算临时端口使用率，临时端口使用率=（临时端口使用数/临时端口总数）*100，确认临时端口使用率是否超过阈值。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

步骤6 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

检查系统环境是否异常。

步骤7 执行以下命令导入临时文件，并查看“port_result.txt”文件中高使用率端口。

netstat -tnp|sort > $BIGDATA_HOME/tmp/port_result.txt

```
netstat -tnp|sort

Active Internet connections (w/o servers)

Proto Recv Send LocalAddress ForeignAddress State PID/ProgramName tcp   0   0 10-120-85-154:45433
10-120-85-154:9866 CLOSE_WAIT 94237/java
tcp   0   0 10-120-85-154:45434  10-120-85-154:9866 CLOSE_WAIT 94237/java
tcp   0   0 10-120-85-154:45435  10-120-85-154:9866 CLOSE_WAIT 94237/java
...
```

步骤8 执行如下命令，查看占用大量端口的进程。

ps -ef |grep *PID*

说明

- PID为**步骤7**查询出所属端口的进程号。
- 可以执行如下命令，收集系统所有进程信息，查看占用大量端口的进程。
  ps -ef > $BIGDATA_HOME/tmp/ps_result.txt

步骤9 请系统管理员确认后，清除大量占用端口的进程，等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤10**。

收集故障信息。

步骤10 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选“OMS”，单击“确定”。

步骤12 设置“主机”为告警所在节点和主OMS节点。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤14 请联系运维人员，发送已收集的故障日志信息及“ port_result.txt”和“ ps_result.txt”文件，并删除环境中残留的两个临时文件。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.31 ALM-12053 主机文件句柄使用率超过阈值

告警解释

系统每30秒周期性检测主机文件句柄使用率，并把实际使用率和阈值（系统默认阈值80%）进行比较，当检测到主机文件句柄使用率连续多次（默认值为5）超过阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 主机状态 > 主机文件句柄使用率”修改阈值。

平滑次数为1，主机文件句柄使用率小于或等于阈值时，告警恢复；平滑次数大于1，主机文件句柄使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12053	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

系统应用无法打开文件、网络等IO操作，程序异常。

可能原因

- 应用进程存在异常，如打开的文件或socket没有关闭。
- 文件句柄数不满足当前业务需求。
- 系统环境异常。

处理步骤

查看进程打开文件情况。

步骤1 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的，获取告警所在主机IP地址。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**lsof -n|awk '{print $2}'|sort|uniq -c|sort -nr|more**，查看文件句柄占用较多的进程。

步骤4 分析打开文件数目较多的进程，分析该进程是否存在异常，如打开的文件或socket没有关闭。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 文件句柄占用多的异常进程进行确认释放。

步骤6 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

增大文件句柄数。

步骤7 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的，获取告警所在主机IP地址。

步骤8 以**root**用户登录告警所在主机。

步骤9 联系系统管理员，增大系统文件句柄数。

步骤10 执行**cat /proc/sys/fs/file-nr**查看已使用句柄数和最大句柄数。第一个值为已使用句柄数，第三个值为最大句柄数，计算使用率是否超过设定阈值。

```
# cat /proc/sys/fs/file-nr
12704    0    640000
```

- 是，执行**步骤9**。
- 否，执行**步骤11**。

步骤11 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤12**。

检查系统环境是否异常。

步骤12 联系系统管理员，检查操作系统是否存在异常。

- 是，恢复操作系统故障，执行**步骤13**。
- 否，执行**步骤14**。

步骤13 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤14**。

收集故障信息。

步骤14 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤15 在“服务”中勾选“OMS”，单击“确定”。

步骤16 设置“主机”为告警所在节点和主OMS节点。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.32 ALM-12054 证书文件失效

告警解释

系统在每天二十三点检查当前系统中的证书文件是否失效（即当前集群中的证书文件是否过期，或者尚未生效）。如果证书文件失效，产生该告警。

当重新导入一个正常证书，并且状态不为失效状态，该告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12054	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

提示用户证书文件已经失效，部分功能受限，无法正常使用。

可能原因

系统未导入证书（CA证书、HA根证书、HA用户证书、Gaussdb根证书或者Gaussdb用户证书等）、导入证书失败、证书文件失效。

处理步骤

查看告警原因。

步骤1 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的。

查看“附加信息”，获取告警附加信息。

- 告警附加信息中显示“CA Certificate”，以**omm**用户登录主OMS管理节点，执行**步骤2**。
- 告警附加信息中显示“HA root Certificate”，查看“定位信息”获取告警所在节点主机名，以**omm**用户登录该主机，执行**步骤3**。
- 告警附加信息中显示“HA server Certificate”，查看“定位信息”获取告警所在节点主机名，以**omm**用户登录该主机，执行**步骤4**。

检查系统中合法证书文件的有效期。

步骤2 查看当前系统时间是否在CA证书的有效期内。

执行命令**bash ${CONTROLLER_HOME}/security/cert/conf/querycertvalidity.sh**可以查看CA根证书的生效时间与失效时间。

- 是，执行**步骤7**。
- 否，执行**步骤5**。

步骤3 查看当前系统时间是否在HA根证书的有效期内。

执行命令**openssl x509 -noout -text -in ${CONTROLLER_HOME}/security/certHA/root-ca.crt**可以查看HA根证书的生效时间与失效时间。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

步骤4 查看当前系统时间是否在HA用户证书的有效期内。

执行命令**openssl x509 -noout -text -in ${CONTROLLER_HOME}/security/certHA/server.crt**可以查看HA用户证书的生效时间与失效时间。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

CA或者HA证书的“生效时间”和“失效时间” 示例：

```
Certificate:
    Data:
        Version: 3 (0x2)
        Serial Number:
            97:d5:0e:84:af:ec:34:d8
        Signature Algorithm: sha256WithRSAEncryption
        Issuer: C=CN, ST=xxx, L=yyy, O=zzz, OU=IT, CN=HADOOP.COM
        Validity
            Not Before: Dec 13 06:38:26 2016 GMT          //生效时间
            Not After : Dec 11 06:38:26 2026 GMT          //失效时间
```

导入证书文件。

步骤5 导入新的CA证书文件。

申请或生成新的CA证书文件并导入。导入CA证书后该告警信息会自动清除，查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤7**。
- 否，处理完毕。

步骤6 导入新的HA证书文件。

申请或生成新的HA证书文件并导入。导入CA证书后该告警信息会自动清除，查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤7**。
- 否，处理完毕。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“Controller”、“OmmServer”、“OmmCore”和“Tomcat”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.33 ALM-12055 证书文件即将过期

告警解释

系统每天二十三点检查一次当前系统中的证书文件，如果当前时间距离过期时间不足告警阈值天数，则证书文件即将过期，产生该告警。

当重新导入一个正常证书，并且状态不为即将过期，该告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12055	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

提示用户证书文件即将过期，如果证书文件过期，则会导致部分功能受限，无法正常使用。

可能原因

系统证书文件（CA证书、HA根证书、HA用户证书、Gaussdb根证书或者Gaussdb用户证书等）剩余有效期小于证书的告警阈值。

处理步骤

查看告警原因。

步骤1 打开FusionInsight Manager页面，在实时告警列表中，单击此告警所在行的。

查看“附加信息”，获取告警附加信息。

- 告警附加信息中显示“CA Certificate”，以**omm**用户登录主OMS管理节点，执行**步骤2**。
- 告警附加信息中显示“HA root Certificate”，查看“定位信息”获取告警所在节点主机名，以**omm**用户登录该主机，执行**步骤3**。
- 告警附加信息中显示“HA server Certificate”， 查看“定位信息”获取告警所在节点主机名，以**omm**用户登录该主机，执行**步骤4**。

检查系统中合法证书文件的有效期。

步骤2 查看当前CA证书剩余有效期是否小于证书的告警阈值。

执行命令**bash ${CONTROLLER_HOME}/security/cert/conf/querycertvalidity.sh**可以查看CA根证书的生效时间与失效时间。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤3 查看当前HA根证书剩余有效期是否小于证书的告警阈值。

执行命令**openssl x509 -noout -text -in ${CONTROLLER_HOME}/security/certHA/root-ca.crt**可以查看HA根证书的生效时间与失效时间。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤4 查看当前HA用户证书剩余有效期是否小于证书的告警阈值。

执行命令**openssl x509 -noout -text -in ${CONTROLLER_HOME}/security/certHA/server.crt**可以查看HA用户证书的生效时间与失效时间。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

CA或者HA证书的“生效时间”和“失效时间” 示例：

```
Certificate:
    Data:
        Version: 3 (0x2)
        Serial Number:
            97:d5:0e:84:af:ec:34:d8
        Signature Algorithm: sha256WithRSAEncryption
        Issuer: C=CN, ST=xxx, L=yyy, O=zzz, OU=IT, CN=HADOOP.COM
        Validity
            Not Before: Dec 13 06:38:26 2016 GMT          //生效时间
            Not After : Dec 11 06:38:26 2026 GMT          //失效时间
```

导入证书文件。

步骤5 导入新的CA证书文件。

申请或生成新的CA证书文件并导入。手动清除该告警信息，查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤7**。
- 否，处理完毕。

步骤6 导入新的HA证书文件。

申请或生成新的HA证书文件并导入。手动清除该告警信息，查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤7**。
- 否，处理完毕。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”

步骤8 在“服务”中勾选“Controller”、“OmmServer”、“OmmCore”和“Tomcat”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.34 ALM-12057 元数据未配置周期备份到第三方服务器的任务

告警解释

系统安装完成后会检查元数据是否有周期备份到第三方服务器的任务，然后每1小时会检查一次。如果元数据未配置周期备份到第三方服务器的任务，将发送重要告警。

在用户创建元数据周期备份到第三方服务器的任务后，告警消除。

告警属性

告警ID	告警级别	是否自动清除
12057	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

如果没有将元数据备份到第三方服务器，当集群主备管理节点同时故障且本地备份数据丢失时，导致元数据无法恢复。

可能原因

元数据未配置周期备份到第三方服务器任务。

处理步骤

查看元数据是否配置周期备份。

步骤1 在FusionInsight Manager管理界面，选择“运维 > 告警 > 告警”。

步骤2 在告警列表中单击该告警的，从“附加信息”中获取产生告警的数据模块。

步骤3 选择“运维 > 备份恢复 > 备份管理 > 创建”。

步骤4 配置备份任务，需要配置的备份数据与该告警的附加信息保持一致。

步骤5 创建备份任务成功后，等待2分钟，检查告警是否消除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”，单击“确定”。

步骤8 单击右上角的设置日志收集的时间范围，一般为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.35 ALM-12061 进程使用率超过阈值

告警解释

系统每30秒周期性检测omm进程使用情况，执行**ps -o nlwp,pid,args, -u omm | awk '{sum+=$1} END {print "", sum}'**命令，获取当前omm用户并发的所有进程数，在omm用户下，执行**ulimit -u**，获取omm用户可以同时打开的进程最大数。

结果相除，获取到对应的omm用户进程使用率。进程使用率默认提供一个阈值范围。当检测到进程使用率超出阈值范围时产生该告警。

平滑次数为3，进程使用率小于或等于阈值时，告警恢复；如果当前平滑次数大于1，进程使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12061	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

- 无法切换到omm用户。
- 无法创建新的omm线程。

可能原因

- 告警阈值配置不合理。
- omm用户可以同时打开的进程（包括线程）的最大个数配置不合理。
- 同时打开的进程过多。

处理步骤

检查告警阈值配置或者平滑次数配置是否合理。

步骤1 在FusionInsight Manager界面，基于实际CPU使用情况，修改告警阈值和平滑次数配置项。

根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 进程 > omm 进程使用率”中更改告警的平滑次数，如图10-31所示。

说明

该选项的含义为告警检查阶段，“平滑次数”为连续检查多少次超过阈值，则发送告警。

图 10-31 设置告警平滑次数

根据实际服务的使用情况在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > 主机 > 进程 > omm 进程使用率”中修改对应规则的阈值，如图10-32所示。

图 10-32 设置告警阈值

步骤2 等待2分钟，查看告警是否自动恢复。

- 是，处理完毕。
- 否，执行**步骤3**。

检查系统omm用户同时打开的进程（包括线程）最大数的配置是否合理。

步骤3 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址

步骤4 以**root**用户登录告警所在主机。

步骤5 执行命令**su - omm**，切换到**omm**用户。

步骤6 执行命令**ulimit -u**，获取到当前配置的omm用户同时打开的线程最大数的配置值，查看该值是否大于等于60000。

- 是，执行**步骤8**。
- 否，执行**步骤7**。

步骤7 执行命令**ulimit -u 60000**，将**omm**用户的该配置修改为60000，等待2分钟，查看告警是否消失。

- 是，处理完毕。

- 否，执行**步骤12**。

检查是否同时打开的进程过多。

步骤8 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址。

步骤9 以**root**用户登录告警所在主机。

步骤10 执行命令**ps -o nlwp,pid,lwp,args, -u omm|sort -n**，查看系统当前使用的线程数量。

命令回显结果是基于线程数排序的，分析线程数最大的top5线程，结合业务分析是否异常使用，如果是，则需要联系相关维护人员修复该异常，如果所有线程均正常使用，则需要执行**ulimit -u**命令，将该值调整到大于60000。

步骤11 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤12**。

收集故障信息。

步骤12 在集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤13 在“服务”中勾选“OmmServer”和“NodeAgent”，单击“确定”。

步骤14 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤15 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.36 ALM-12062 OMS 参数配置同集群规模不匹配

告警解释

系统每一个小时，整点检查一次OMS参数配置和集群规模是否匹配，如果检查OMS配置的参数不足以支撑当前的集群规模，系统将发送此告警。待用户修改OMS参数配置，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
12062	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

安装集群或者系统扩容节点未同步修改相应的OMS配置。

可能原因

OMS配置同集群规模不匹配。

处理步骤

检查OMS配置同集群规模是否匹配。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**su - omm**，切换到**omm**用户。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/controller/scriptlog/modify_manager_param.log**打开对应日志，搜索日志“**Current oms configurations can not support *xx* nodes**”，其中*xx*为当前集群节点个数。

步骤5 参考**根据集群节点数优化Manager配置**，对当前集群配置进行优化。

步骤6 配置完成后等待1小时后，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“Controller”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

根据集群节点数优化Manager配置

步骤1 以**omm**用户登录主管理节点。

步骤2 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/sbin

步骤3 执行以下命令查看当前集群Manager相关配置。

sh oms_config_info.sh -q

步骤4 执行以下命令指定当前集群的节点数。

命令格式：**sh oms_config_info.sh -s** *节点数*

例如：

sh oms_config_info.sh -s 1000

根据界面提示，输入“y”：

```
The following configurations will be modified:
    Module      Parameter         Current             Target
    Controller   controller.Xmx    4096m            =>  16384m
    Controller   controller.Xms    1024m            =>  8192m       Controller
controller.node.heartbeat.error.threshold    30000                 =>   60000
    Pms          pms.mem           8192m            =>  10240m
Do you really want to do this operation? (y/n):
```

界面提示以下信息表示配置更新成功：

```
...
Operation has been completed. Now restarting OMS server.                [done]
Restarted oms server successfully.
```

说明

- 配置更新过程中，OMS会自动重启。
- 相近数量的节点规模对应的Manager相关配置是通用的，例如100节点变为101节点，并没有新的配置项需要刷新。

----结束

10.13.37 ALM-12063 磁盘不可用

告警解释

系统每一个小时，整点检查一次当前主机的磁盘是否可用，只检查数据盘，在磁盘对应的挂载目录下执行创建文件，写文件和删文件等操作，如果能够成功则认为磁盘可用，发送恢复告警，如果不能成功，则发送故障告警。

告警属性

告警ID	告警级别	是否自动清除
12063	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
磁盘名	产生告警的磁盘名称。

对系统的影响

数据盘不可写或者不可读，会导致业务异常。

可能原因

磁盘挂载目录权限异常或磁盘坏道。

处理步骤

检查磁盘挂载目录权限是否正常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址和告警的磁盘名称DiskName。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**df -h |grep DiskName**，获取对应的挂载点，查看挂载目录的权限，是否存在不可写或者不可读。

- 是，执行**步骤4**。
- 否，执行**步骤8**。

说明

如果挂载目录权限为000，或者属主为root，则表示当前状态为不可读不可写。

步骤4 修改目录权限为合适的目录权限。

步骤5 等待一小时，查看告警是否恢复。

- 是，操作结束。
- 否，执行**6**。

步骤6 联系硬件工程师，修复磁盘故障。

步骤7 等待一小时，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.38 ALM-12064 主机随机端口范围配置与集群使用端口冲突

告警解释

系统每一个小时检查一次主机随机端口配置范围是否与集群使用端口范围冲突，如果有冲突，则发送此告警。待客户重新修改该主机的随机端口范围配置到正常范围，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
12064	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

集群系统默认端口被占用，导致某些进程启动失败。

可能原因

随机端口范围配置被修改。

处理步骤

检查系统当前的随机端口范围。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**cat /proc/sys/net/ipv4/ip_local_port_range**，获取该主机的随机端口范围配置，查看最小值是否小于32768。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 执行命令**vim /etc/sysctl.conf**，修改配置项**net.ipv4.ip_local_port_range**的值为**32768 61000**，如果没有该配置项，则新增net.ipv4.ip_local_port_range = 32768 61000。

步骤5 执行命令**sysctl -p /etc/sysctl.conf**使修改的配置生效。

步骤6 配置完成后等待1小时后，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.39 ALM-12066 节点间互信失效

告警解释

系统每一个小时检查一次主OMS节点和其他Agent节点间的互信是否正常，如果存在互信失效的节点，则发送告警。待客户修复改问题，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
12066	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

可能会导致管理面的一些操作异常。

可能原因

- /etc/ssh/sshd_config配置文件被破坏。
- omm密码过期。

处理步骤

查看/etc/ssh/sshd_config配置文件状态。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看告警详情中涉及的主机列表。

步骤2 以**omm**用户登录主OMS管理节点。

步骤3 依次在告警详情中的节点执行**ssh**命令：**ssh** *host2*（host2为告警详情中OMS节点之外的其它节点），看是否连接失败。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 打开host2主机上的“/etc/ssh/sshd_config”配置文件，查看另外节点是否配置在AllowUsers 、DenyUsers等白名单或者黑名单中。

- 是，执行**步骤5**。
- 否，联系OS专家处理。

步骤5 修改白名单或者黑名单设置，保证**omm**用户在白名单中或者不在黑名单中。然后持续一段时间观察告警是否清除。

- 是，操作结束。
- 否，执行**步骤6**。

查看omm密码状态。

步骤6 查看ssh命令的交互信息。

- 要求输入**omm**用户的密码（Password:），执行**步骤7**。
- 要求输入密码短语（Enter passphrase for key '/home/omm/.ssh/id_rsa':），执行**步骤9**。

步骤7 排查OMS节点和host2节点**omm**用户的信任清单（/home/omm/.ssh/authorized_keys），查看是否包含对端主机**omm**用户的公钥文件（/home/omm/.ssh/id_rsa.pub）。

- 是，联系OS专家处理。
- 否，把对端主机**omm**用户的公钥添加到本机的信任清单中。

步骤8 把对端主机**omm**用户的公钥添加到本机的信任清单中，然后依次在告警详情中的节点执行**ssh**命令：**ssh** *host2*（host2为告警详情中OMS节点之外的其它节点），看是否连接失败。

- 是，执行**步骤9**。
- 否，持续一段时间观察告警是否清除，如果清除则操作结束，如果未清除请执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选“Controller”，单击“确定”。

步骤11 单击右上角的设置日志收集的时间范围，一般为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

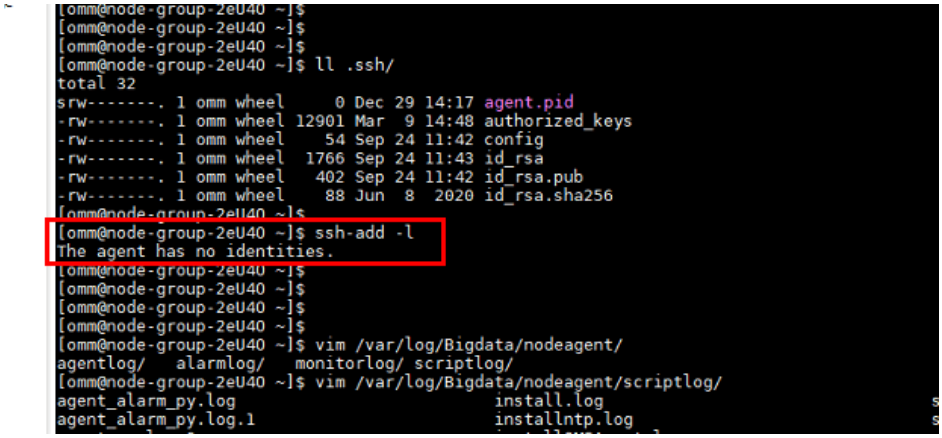
参考信息

节点互信异常处理方法如下：

须知

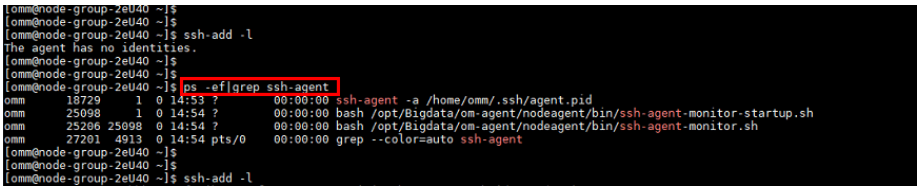
- 本此操作需使用**omm**用户执行。
- 如果节点间网络不通，请先解决网络不通的问题，可以检查两个节点是否通一个安全组，是否有设置hosts.deny、hosts.allow 等。

1. 在两端节点执行**ssh-add -l** 确认是否有identities信息。

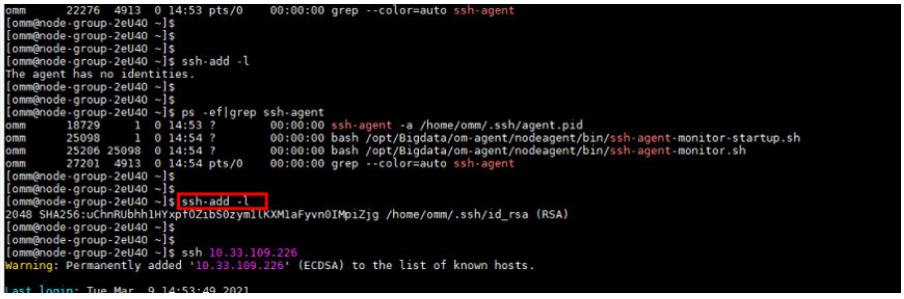


   - 是，执行**4**。
   - 否，执行**2**。

2. 如果没有identities信息，执行**ps -ef|grep ssh-agent**找到**ssh-agent**进程，并停止该进程并等待该进程自动重启。



3. 执行**ssh-add -l** 查看是否已经添加identities信息，如果已经添加手动**ssh**确认是否互信正常。



4. 如果有identities信息，需要确认“/home/omm/.ssh/authorized_keys”中是否有对端节点“/home/omm/.ssh/id_rsa.pub”文件中的信息，如果没有手动添加。
5. 检查“/home/omm/.ssh”目录下的文件权限是否被修改。
6. 排查如下日志文件“/var/log/Bigdata/nodeagent/scriptlog/ssh-agent-monitor.log”。
7. 如果用户把**omm**的“/home”目录删除了，请联系MRS支撑人员修复。

10.13.40 ALM-12067 tomcat 资源异常

告警解释

HA每85秒周期性检测Manager的tomcat资源。当HA连续2次都检测到tomcat资源异常时，产生该告警。

当HA检测到tomcat资源正常后，告警恢复。

tomcat资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的tomcat资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12067	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- tomcat持续重启。

可能原因

- tomcat目录权限异常，tomcat进程异常。

处理步骤

检查tomcat目录权限是否正常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址。

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**su - omm**，切换到**omm**用户。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/ha/scriptlog/tomcat.log**，查看ha的tomcat资源日志，是否有如下关键字**“Cannot find *XXX*”**，根据如下关键字修复对应文件的权限。

步骤5 等待5分钟，查看告警是否自动清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“OmmServer”和 “Tomcat”，单击“确定”。

步骤8 单击右上角的设置日志收集的时间范围，一般为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.41 ALM-12068 acs 资源异常

告警解释

HA每80秒周期性检测Manager的acs资源。当HA连续2次都检测到acs资源异常时，产生该告警。

当HA检测到acs资源正常后，告警恢复。

acs资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的acs资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12068	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- acs进程持续重启，可能引起无法登录FusionInsight Manager。

可能原因

acs进程异常。

处理步骤

检查acs进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的acs资源状态是否正常（单机模式下面，acs资源为normal状态；双机模式下，acs资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/ha/scriptlog/acs.log**，查看ha的acs资源日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤5 等待五分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.42 ALM-12069 aos 资源异常

告警解释

HA每81秒周期性检测Manager的aos资源。当HA连续2次检测到aos资源异常时，产生该告警。

当HA检测到aos资源正常后，告警恢复。

aos资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的acs资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12069	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- aos进程持续重启，可能引起无法登录FusionInsight Manager。

可能原因

aos进程异常。

处理步骤

检查aos进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的aos资源状态是否正常（单机模式下面，aos资源为normal状态；双机模式下，aos资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/ha/scriptlog/aos.log**，查看ha的aos资源日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤5 等待五分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.43 ALM-12070 controller 资源异常

告警解释

HA每80秒周期性检测Manager的controller资源。当HA连续2次检测到controller资源异常时，产生该告警。

当HA检测到controller资源正常后，告警恢复。

controller资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的controller资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12070	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- controller进程持续重启，可能引起无法登录FusionInsight Manager。

可能原因

controller进程异常。

处理步骤

检查controller进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的controller资源状态是否正常（单机模式下面，controller资源为normal状态；双机模式下，controller资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/ha/scriptlog/controller.log**，查看ha的controller资源日志，执行命令**vi $BIGDATA_LOG_HOME/controller/controller.log**，查看controller运行日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤5 等待五分钟，查看告警是否恢复。

- 是，操作结束。

- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.44 ALM-12071 httpd 资源异常

告警解释

HA每120秒周期性检测Manager的httpd资源。当HA连续10次检测到httpd资源异常时，产生该告警。

当HA检测到httpd资源正常后，告警恢复。

httpd资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的httpd资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12071	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- httpd进程持续重启，可能引起无法访问服务原生UI界面。

可能原因

httpd进程异常。

处理步骤

检查httpd进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，切换至omm用户。

步骤4 执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的httpd资源状态是否正常（单机模式下面，httpd资源为normal状态；双机模式下，httpd资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤7**。
- 否，执行**步骤5**。

步骤5 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/ha/scriptlog/httpd.log**，查看ha的httpd资源日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤6 等待五分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.45 ALM-12072 floatip 资源异常

告警解释

HA每9秒周期性检测Manager的floatip资源。当HA连续3次检测到floatip资源异常时，产生该告警。

当HA检测到floatip资源正常后，告警恢复。

floatip资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的floatip资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12072	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- floatip进程持续重启，可能引起无法访问服务原生UI界面。

可能原因

浮动IP地址异常。

处理步骤

检查主管理节点的浮动IP地址状态。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址及资源名称。

步骤2 以**root**用户登录主管理节点。

步骤3 执行以下命令进入“ ${BIGDATA_HOME}/om-server/om/sbin/”目录。

su - omm

cd ${BIGDATA_HOME}/om-server/om/sbin/

步骤4 执行“**sh status-oms.sh**”命令，执行**status-oms.sh**脚本检查主Manager的浮动IP是否正常，查看回显中，主管理节点的“ResName”为“floatip”的一行，是否显示以下信息：

例如：

```
10-10-10-160   floatip   Normal   Normal   Single_active
```

- 是，执行**步骤8**。
- 否，执行**步骤5**。

步骤5 执行**ifconfig**命令检查浮动IP地址的网卡是否存在。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 执行命令**ifconfig** *网卡名称 浮动IP地址* **netmask** *子网掩码*重新配置浮动IP网卡（例如，**ifconfig eth0 10.10.10.102 netmask 255.255.255.0**）。

步骤7 等待5分钟，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.46 ALM-12073 cep 资源异常

告警解释

HA每60秒周期性检测Manager的cep资源。当HA连续2次检测到cep资源异常时，产生该告警。

当HA检测到cep资源正常后，告警恢复。

cep资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的cep资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12073	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- cep进程持续重启，可能会导致监控数据异常。

可能原因

cep进程异常。

处理步骤

检查cep进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的cep资源状态是否正常（单机模式下面，cep资源为normal状态；双机模式下，cep资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/cep/cep.log**和**vi $BIGDATA_LOG_HOME/omm/oms/cep/scriptlog/cep_ha.log**，查看ha的cep资源日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤5 等待五分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.47 ALM-12074 fms 资源异常

告警解释

HA每60秒周期性检测Manager的fms资源。当HA连续2次检测到fms资源异常时，产生该告警。

当HA检测到fms资源正常后，告警恢复。

fms资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的fms资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12074	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

- Manager主备倒换。
- fms进程持续重启，可能导致告警信息无法正常上报。

可能原因

fms进程异常。

处理步骤

检查fms进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的fms资源状态是否正常（单机模式下面，fms资源为normal状态；双机模式下，fms资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/fms/fms.log** 和**vi $BIGDATA_LOG_HOME/omm/oms/fms/scriptlog/fms_ha.log** 查看ha的fms资源日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤5 等待五分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.48 ALM-12075 pms 资源异常

告警解释

HA每55秒周期性检测Manager的pms资源。当HA连续3次检测到pms资源异常时，产生该告警。

当HA检测到pms资源正常后，告警恢复。

pms资源为单主资源，一般资源异常会导致主备倒换，看到告警时，基本已经主备倒换，并在新主环境上启动新的pms资源，告警恢复。该告警用于提示用户，Manager主备倒换的原因。

告警属性

告警ID	告警级别	是否自动清除
12075	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称
角色名	产生告警的角色名称
主机名	产生告警的主机名

对系统的影响

- Manager主备倒换。
- pms进程持续重启，可能会导致监控信息异常。

可能原因

pms进程异常。

处理步骤

检查pms进程是否异常。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机名称。

步骤2 以**root**用户登录该告警的主机地址。

步骤3 执行命令**su - omm**，执行**sh ${BIGDATA_HOME}/om-server/OMS/workspace0/ha/module/hacom/script/status_ha.sh**，查询当前HA管理的pms资源

状态是否正常（单机模式下面，pms资源为normal状态；双机模式下，pms资源在主节点为normal状态，在备节点为stopped状态。）

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 执行命令**vi $BIGDATA_LOG_HOME/omm/oms/pms/pms.log** 和**vi $BIGDATA_LOG_HOME/omm/oms/pms/scriptlog/pms_ha.log**，查看ha的pms资源日志，是否有关键字“ERROR”，分析日志查看资源异常原因并修复。

步骤5 等待五分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”和“OmmServer”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.49 ALM-12076 gaussDB 资源异常

告警解释

HA软件每10秒周期性检测Manager的数据库。当HA软件连续3次检测到数据库异常时，产生该告警。

当HA检测到数据库正常后，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12076	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

如果数据库异常，所有核心业务和相关业务进程，例如告警和监控功能，都会受影响。

可能原因

数据库异常。

处理步骤

检查主备管理节点的数据库状态。

步骤1 以**root**用户分别登录主备管理节点，执行**su - ommdba**命令切换到**ommdba**用户，执行**gs_ctl query**命令。查看回显是否显示以下信息。

主管理节点的回显：

```
Ha state:
        LOCAL_ROLE                : Primary
        STATIC_CONNECTIONS        : 1
        DB_STATE                  : Normal
        DETAIL_INFORMATION        : user/password invalid
Senders info:
        No information
Receiver info:
        No information
```

备管理节点的回显：

```
Ha state:
        LOCAL_ROLE                : Standby
        STATIC_CONNECTIONS        : 1
        DB_STATE                  : Normal
        DETAIL_INFORMATION        : user/password invalid
Senders info:
        No information
Receiver info:
        No information
```

- 是，执行**步骤3**。
- 否，执行**步骤2**。

步骤2 联系网络管理员查看是否为网络故障，并修复故障。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 等待5分钟，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤4**。

步骤4 分别登录主备管理节点，执行**su - omm**命令切换到**omm**，用户进入“${BIGDATA_HOME}/om-server/om/sbin/”目录，并执行**status-oms.sh**脚本检查主备Manager的floatip资源和gaussDB资源是否如下图所示的状态：

```
acs                                 Normal              Normal          Single_active
aos                                 Normal              Normal          Single_active
cep                                 Normal              Normal          Single_active
controller                          Normal              Normal          Single_active
feed_watchdog                       Normal              Normal          Double_active
floatip                             Normal              Normal          Single_active
fms                                 Normal              Normal          Single_active
gaussDB                             Active_normal       Normal          Active_standby
heartBeatCheck                      Normal              Normal          Single_active
httpd                               Normal              Normal          Single_active
iam                                 Normal              Normal          Single_active
ntp                                 Active_normal       Normal          Active_standby
okerberos                           Normal              Normal          Double_active
oldap                               Active_normal       Normal          Active_standby
pms                                 Normal              Normal          Single_active
tomcat                              Normal              Normal          Single_active
acs                                 Stopped             Normal          Single_active
aos                                 Stopped             Normal          Single_active
cep                                 Stopped             Normal          Single_active
controller                          Stopped             Normal          Single_active
feed_watchdog                       Normal              Normal          Double_active
floatip                             Stopped             Normal          Single_active
fms                                 Stopped             Normal          Single_active
gaussDB                             Standby_normal      Normal          Active_standby
heartBeatCheck                      Stopped             Normal          Single_active
httpd                               Stopped             Normal          Single_active
```

- 是，在告警列表中找到该告警，手工清除该告警。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选“OmmServer”，单击“确定”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.50 ALM-12077 omm 用户过期

告警解释

系统每天零点开始，每8小时检测当前系统中omm用户是否过期，如果用户过期，则发送告警。

当系统中omm用户过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12077	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

omm用户过期，Manager各节点互信不可用，无法对服务提供管理功能。

可能原因

omm用户过期。

处理步骤

检查系统中omm用户是否过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l omm**命令来查看当前**omm**用户密码设置信息。

步骤2 查找“Account expires”对应值，查看用户设置是否过期。

说明

如果参数值为“never”，则代表永不过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -E** *'yyyy-MM-dd'* **omm**命令设置**omm**用户过期的期限，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.51 ALM-12078 omm 密码过期

告警解释

系统每天零点开始，每8小时检测当前系统中**omm**密码是否过期，如果密码过期，则发送告警。

当系统中**omm**密码过期的期限修改，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12078	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

omm密码过期，Manager各节点互信不可用，无法对服务提供管理功能。

可能原因

omm密码过期。

处理步骤

检查系统中omm密码是否过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l omm**命令来查看当前**omm**用户密码设置信息。

步骤2 查找“Password expires”对应值，查看密码设置是否过期。

说明

如果参数值为“never”，则代表永不过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -M** *'天数'* **omm**命令设置**omm**密码的有效天数，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.52 ALM-12079 omm 用户即将过期

告警解释

系统每天零点开始，每8小时检测当前系统中**omm**用户是否即将过期，如果当前时间与用户过期时间剩余不足15天，则发送告警。

当系统中**omm**用户过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12079	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

omm用户过期，Manager各节点互信不可用，无法对服务提供管理功能。

可能原因

该主机**omm**用户即将过期。

处理步骤

检查系统中omm用户是否即将过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l omm**命令来查看当前**omm**用户密码设置信息。

步骤2 查找“Account expires”对应值，查看用户设置是否即将过期。

说明

如果参数值为“never”，则代表永不过期；如果为日期值，则查看是否在15天内过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -E** *'yyyy-MM-dd'* **omm**命令设置**omm**用户过期的期限，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.53 ALM-12080 omm 密码即将过期

告警解释

系统每天零点开始，每8小时检测当前系统中**omm**密码是否即将过期，如果当前时间与密码过期时间剩余不足15天，则发送告警。

当系统中**omm**密码过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12080	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

omm密码过期，Manager各节点互信不可用，无法对服务提供管理功能。

可能原因

该主机**omm**密码即将过期。

处理步骤

检查系统中omm密码是否即将过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l omm**命令来查看当前**omm**用户密码设置信息。

步骤2 查找“Password expires”对应值，查看密码设置是否即将过期。

说明

如果参数值为“never”，则代表永不过期；如果为日期值，则查看是否在15天内过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -M** *'天数'* **omm**命令设置**omm**密码的有效天数，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.54 ALM-12081 ommdba 用户过期

告警解释

系统每天零点开始，每8小时检测当前系统中**ommdba**用户是否过期，如果用户过期，则发送告警。

当系统中**ommdba**用户过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12081	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

ommdba用户过期，OMS数据库无法管理，数据不能访问。

可能原因

该主机**ommdba**用户过期。

处理步骤

检查系统中ommdba用户是否过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l ommdba**命令查看当前**ommdba**用户密码设置信息。

步骤2 查找“Account expires”对应值，查看用户设置是否过期。

说明

如果参数值为“never”，则代表永不过期；如果为日期值，则查看是否过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -E** *'yyyy-MM-dd'* **ommdba**命令设置**ommdba**用户过期的期限，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.55 ALM-12082 ommdba 用户即将过期

告警解释

系统每天零点开始，每8小时检测当前系统中**ommdba**用户是否即将过期，如果用户即将在15天内过期，则发送告警。

当系统中**ommdba**用户过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12082	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

ommdba用户过期，OMS数据库无法管理，数据不能被访问。

可能原因

该主机**ommdba**用户即将过期。

处理步骤

检查系统中ommdba用户是否即将过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l ommdba**命令来查看当前**ommdba**用户设置信息。

步骤2 查找“Account expires”对应值，查看用户设置是否即将过期。

说明

如果参数值为“never”，则代表永不过期；如果为日期值，则查看是否在15天内过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -E** *'yyyy-MM-dd'* **ommdba**命令设置**ommdba**用户过期的期限，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.56 ALM-12083 ommdba 密码即将过期

告警解释

系统每天零点开始，每8小时检测当前系统中ommdba密码是否即将过期，如果当前时间与ommdba密码过期时间剩余不足15天，则发送告警。

当系统中**ommdba**用户密码过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12083	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

ommdba密码过期，OMS数据库无法管理，数据不能访问。

可能原因

该主机**ommdba**密码即将过期。

处理步骤

检查系统中ommdba密码是否即将过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l ommdba**命令来查看当前**ommdba**用户密码设置信息。

步骤2 查找“Password expires”对应值，查看密码设置是否即将过期。

说明

如果参数值为“never”，则代表永不过期；如果为日期值，则查看是否在15天内过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -M** *'天数'* **ommdba**命令设置**ommdba**密码的有效天数，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.57 ALM-12084 ommdba 密码过期

告警解释

系统每天零点开始，每8小时检测当前系统中**ommdba**密码是否过期，如果过期，则发送告警。

当系统中**ommdba**密码过期的期限重置，当前状态为正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12084	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

ommdba密码过期，Manager各节点互信不可用，无法对服务提供管理功能。

可能原因

该主机**ommdba**密码过期。

处理步骤

检查系统中ommdba密码是否过期。

步骤1 以**root**用户登录集群故障节点。

执行**chage -l ommdba**命令来查看当前**ommdba**用户密码设置信息。

步骤2 查找“Password expires”对应值，查看密码设置是否过期。

说明

如果参数值为“never”，则代表永不过期；如果为日期值，则查看是否已经过期。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 执行**chage -M** *'天数'* **ommdba**命令设置**ommdba**密码的有效天数，等待8小时，观察告警是否自动清除。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选“NodeAgent”，单击“确定”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.58 ALM-12085 服务审计日志转储失败

告警解释

系统每天凌晨三点启动服务审计日志转储，将服务审计日志备份到OMS节点，如果转储失败，则发送告警。当下一次转储成功，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12085	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

服务审计日志有可能丢失。

可能原因

- 服务审计日志过大。
- OMS备份路径存储空间不足。
- 服务所在某一个主机的存储空间不足。

处理步骤

检查是否服务审计日志过大。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的主机地址

步骤2 以**root**用户登录告警所在主机。

步骤3 执行命令**vi ${BIGDATA_LOG_HOME}/controller/scriptlog/getLogs.log**，检索关键字 "LOG SIZE is more than 5000MB"。是否能够检索到此关键字。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 查看是否有异常导致服务审计日志过大。

OMS备份路径存储空间不足。

步骤5 执行命令**vi ${BIGDATA_LOG_HOME}/controller/scriptlog/getLogs.log**，检索关键字 "Collect log failed, too many logs on"。是否能够检索到此关键字。

- 是，获取Collect log failed, too many logs on关键字后面的主机IP地址，执行**步骤6**。
- 否，执行**步骤10**。

步骤6 以**root**用户登录**步骤5**中获取到的主机IP地址。

步骤7 执行命令**vi {BIGDATA_LOG_HOME}/nodeagent/scriptlog/collectLog.log**，是否能够检索到此关键字“log size exceeds”。

- 是，执行**步骤8**。
- 否，执行**步骤10**。

步骤8 对OMS节点进行磁盘扩容。

步骤9 等待下一个执行周期（凌晨三点），查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤10**。

检查服务所在某一个主机的空间是否不足

步骤10 执行命令**vi ${BIGDATA_LOG_HOME}/controller/scriptlog/getLogs.log**，检索关键字 "Collect log failed, no enough space on *hostIp*"。是否能够检索到此关键字。

- 是，获取*hostIp*作为异常主机IP，执行**步骤11**。
- 否，执行**步骤14**。

步骤11 以**root**用户登录获取到的主机IP，执行命令**df "$BIGDATA_HOME/tmp" -lP | tail -1 | awk '{print ($4/1024)}'**，获取该主机日志目录剩余空间，查看该值是否小于1000M。

- 是，执行**步骤12**。
- 否，执行**步骤14**。

步骤12 对该节点进行磁盘扩容。

步骤13 等待下一个执行周期，凌晨3点，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤14**。

收集故障信息。

步骤14 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤15 在“服务”中勾选 "Controller”，单击“确定”。

步骤16 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分，单击“下载”。

步骤17 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.59 ALM-12087 系统处于升级观察期

告警解释

系统定时在每天零点查看当前系统是否处于升级观察期，同时检查进入升级观察时间是否超过了为客户预留的升级观察期时间（默认为10天）。当系统处于升级观察期，并且进入升级观察期时间超过了为客户预留的升级观察期时间（默认时间为10天）时，系统触发此告警。如果用户进行了回滚或者提交操作，使得系统退出升级观察期，该告警将会自动清除。

告警属性

告警ID	告警级别	是否自动清除
12087	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Upgrade Observation Period (Days)	进入升级观察期的时间（天）。

对系统的影响

会导致下一次升级或者补丁失败。

可能原因

系统升级之后超过一定时间（默认为10天）未做升级提交。

处理步骤

查看系统是否处于升级观察期。

步骤1 以**root**用户登录主管理节点。

步骤2 执行以下命令，切换至omm用户并登录omm数据库。

su - omm

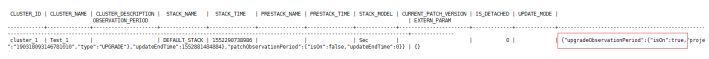
gsql -U omm -W *omm数据库密码* **-p 20015**

步骤3 执行命令 **select * from OM_CLUSTERS;**，查看集群信息。

步骤4 查看图中upgradObservationPeriod isON值是否为true，如**图10-33**所示。

- 是，表示系统当前处于升级观察期，请使用UpdateTool工具提交该次升级，具体操作参考相应版本的升级指导书。
- 否，执行**步骤6**。

图 10-33 集群信息



步骤5 等待第二天凌晨，查看告警是否清除。

- 是，操作结束。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选“Controller”，单击“确定”。

步骤8 单击右上角的设置日志收集的时间范围，一般为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.60 ALM-12089 节点间网络互通异常

告警解释

告警模块按10s周期检测集群间节点的网络健康状态。当检测到某两台节点之间网络不可达或者网络状态不稳定，产生该告警。

当网络恢复正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
12089	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

集群节点间网络健康状态不好时，会影响某些组件的功能使用，如HDFS，ZooKeeper等。

可能原因

- 节点宕机。
- 网络故障。

处理步骤

确认网络健康状态。

步骤1 打开FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看附加信息中的描述信息。明确具体发生告警源IP地址及目标IP，并记录两个IP地址。

步骤2 登录告警上报节点，在告警上报节点上使用**ping**命令，向目标节点手动发起ping请求，检查两个节点之间的网络状态是否正常。

- 是，执行**6**
- 否，执行**3**。

确认节点状态。

步骤3 在FusionInsight Manager界面，单击“主机”查看主机列表中是否包含故障节点，确认故障节点是否已从集群中移除。

- 是，执行**5**。
- 否，执行**4**。

步骤4 查看故障节点运行状态，判断是否处于关机状态。

- 是，启动故障节点，执行**步骤2**。

- 否，联系相关工作人员定位问题，若需要从集群中移除故障节点，执行5，否则执行6。

步骤5 将故障节点从集群所有节点的$NODE_AGENT_HOME/etc/agent/hosts.ini文件中移除，并清空/var/log/Bigdata/unreachable/unreachable_ip_info.log文件内容，同时手动清除告警。

步骤6 等待30s查看告警是否自动清除。

- 是，处理完毕。
- 否，执行7。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选“OmmAgent”，单击“确定”。

步骤9 单击右上角的设置日志收集的时间范围，一般为告警产生时间的前后10秒钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.61 ALM-12101 AZ 不健康

告警解释

AZ容灾开启后，系统每隔5分钟检查一次当前系统上AZ的健康状态，当检测到AZ健康状态为亚健康或者不健康时产生告警。AZ健康状态恢复健康时，告警清除。

告警属性

告警ID	告警级别	是否自动清除
12101	重要	是

告警参数

告警参数	参数含义
来源	产生告警的集群或系统名称。
服务名	产生告警的服务名称。

告警参数	参数含义
AZ名	产生告警的AZ名称。
主机名	产生告警的主机名。

对系统的影响

AZ的健康状态由AZ内的存储资源（HDFS）、计算资源（Yarn）和关键角色的健康度是否超过配置阈值决定。

AZ亚健康有两种：

- 计算资源（Yarn）不健康，存储资源（HDFS）健康，任务无法提交到本AZ，但是数据可以继续往本AZ内读写。
- 计算资源（Yarn）健康，存储资源（HDFS）部分不健康，任务可以提交到本AZ，部分数据可以在本AZ内读写，依赖于Spark/Hive调度感知数据的本地性。

AZ不健康有三种：

- 计算资源（Yarn）健康，存储资源（HDFS）不健康，任务虽然可以提交到本AZ，但是数据无法在本AZ内读写，导致任务提交到本AZ无意义。
- 计算资源（Yarn）不健康，存储资源（HDFS）不健康，任务无法提交到本AZ，数据也无法往本AZ内读写。
- 除Yarn与HDFS以外，关键角色的健康度低于配置阈值。

可能原因

- 计算资源（Yarn）不健康。
- 存储资源（HDFS）不健康。
- 存储资源（HDFS）部分不健康。
- 除Yarn与HDFS以外，关键角色不健康。

处理步骤

关闭容灾演练。

步骤1 在FusionInsight Manager页面，选择“集群 > *待操作集群的名称* > 跨AZ高可用”，打开跨AZ高可用页面。

步骤2 检查AZ容灾列表中健康状态为“非健康”的AZ所在行的操作列中的“容灾演练”是否为灰色。

- 是，执行**步骤4**。
- 否，执行**步骤3**。

步骤3 单击目标AZ行“操作”列中的“恢复”，待恢复后。等待2分钟，刷新页面查看该AZ健康状态。查看是否健康恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

收集故障信息。

步骤4 以**root**用户登录主管理节点。

步骤5 查看不健康服务的日志信息。

- HDFS的日志文件存储路径为“/var/log/Bigdata/hdfs/nn/hdfs-az-state.log”。
- Yarn的日志文件存储路径为“/var/log/Bigdata/yarn/rm/yarn-az-state.log”。
- 其余服务请查看对应服务日志目录下的服务健康检查日志。

步骤6 请联系运维人员，并提供日志文件详细信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.62 ALM-12102 AZ 高可用组件未按容灾需求部署

告警解释

告警模块按照5分钟周期检测AZ高可用组件部署状态。当开启AZ后，支持容灾的组件未按容灾需求部署时产生该告警。组件恢复按容灾需求部署时，告警清除。

告警属性

告警ID	告警级别	是否自动清除
12102	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

对系统的影响

影响单集群跨AZ的高可用能力。

可能原因

支持容灾的组件角色未按容灾需求部署。

处理步骤

获取告警的信息。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”。

步骤2 在告警列表，单击此告警所在行的，从“附加信息”查看未按容灾需求部署的角色名。

重新部署角色实例。

步骤3 选择“集群 > 服务 > *待操作服务名* > 实例”，在实例页面，重新部署或调整该角色实例。

步骤4 等待10分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，请联系运维人员。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.63 ALM-12110 获取 ECS 临时 ak/sk 失败

告警解释

meta服务会周期性地获取ECS临时ak/sk，当调用ECS的meta服务获取临时ak/sk失败时，会产生该告警。

告警属性

告警ID	告警级别	是否自动清除
12110	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名称。

对系统的影响

存算分离场景下，集群由于获取不到最新的临时ak/sk，可能导致访问OBS失败。

可能原因

- MRS集群meta角色状态异常。
- 集群绑定过委托且访问过OBS，但是已经解绑了，现在未绑定委托。

处理步骤

检查meta角色状态。

步骤1 在集群的FusionInsight Manager页面，选择“运维 > 告警 > 告警”，单击此告警所在行的，确定该告警的主机地址。

步骤2 在集群的FusionInsight Manager页面，选择“集群 > 服务 > Meta”，单击“实例”，查看告警产生的主机对应的meta角色状态是否正常。

- 是，执行**步骤4**。
- 否，执行**步骤3**。

步骤3 勾选状态异常的角色，选择“更多 > 重启实例”重启异常状态的meta角色，重启完成后等待几分钟，查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤4**。

重新绑定委托

步骤4 登录MapReduce服务管理控制台。

步骤5 选择“集群列表 > 现有集群”，单击集群名称，进入集群概览页面，查看集群是否绑定委托。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

步骤6 单击“委托管理”，重新绑定委托，等待几分钟后查看告警是否恢复。

- 是，操作结束。
- 否，执行**步骤7**。

步骤7 联系运维人员。

----结束

10.13.64 ALM-13000 ZooKeeper 服务不可用

告警解释

系统每60秒周期性检测ZooKeeper服务状态，当检测到ZooKeeper服务不可用时产生该告警。

ZooKeeper服务恢复时，告警清除。

告警属性

告警ID	告警级别	是否自动清除
13000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

ZooKeeper无法为上层组件提供协调服务，依赖ZooKeeper的组件可能无法正常运行。

可能原因

- ZooKeeper节点上安装了DNS。
- 网络故障。
- KrbServer服务异常。
- ZooKeeper实例状态异常。
- 磁盘容量不足。

处理步骤

检查DNS。

步骤1 查看ZooKeeper实例所在节点上是否安装DNS。在ZooKeeper实例所在Linux节点使用命令**cat /etc/resolv.conf**，看该文件是否为空。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 运行命令**service named status**查看DNS是否启动。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 运行命令**service named stop**将DNS服务停掉，如果出现“Shutting down name server BIND waiting for named to shut down (28s)”结果，即说明DNS服务停止成功。然后将“/etc/resolv.conf”文件的内容（若不为空）全部注释。

步骤4 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

检查网络状态。

步骤5 在ZooKeeper实例所在Linux节点使用**ping**命令，看能否**ping**通其他ZooKeeper实例所在节点的主机名。

- 是，执行**步骤9**。
- 否，执行**步骤6**。

步骤6 修改“/etc/hosts”中的IP信息，添加主机名与IP地址的对应关系。

步骤7 再次执行**ping**命令，查看能否在该ZooKeeper实例节点**ping**通其他ZooKeeper实例节点的主机名。

- 是，执行**步骤8**。
- 否，执行**步骤23**。

步骤8 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

检查KrbServer服务状态（普通模式集群跳过此步骤）。

步骤9 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务”。

步骤10 查看KrbServer服务是否正常。

- 是，执行**步骤13**。
- 否，执行**步骤11**。

步骤11 参考“ALM-25500 KrbServer服务不可用”进行处理，查看KrbServer服务是否能够恢复。

- 是，执行**步骤12**。
- 否，执行**步骤23**。

步骤12 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

检查ZooKeeper服务实例状态。

步骤13 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > quorumpeer”。

步骤14 查看ZooKeeper各实例是否正常。

- 是，执行**步骤18**。
- 否，执行**步骤15**。

步骤15 选中运行状态不为良好的实例，选择“更多 > 重启实例”。

步骤16 查看实例重启后运行状态是否为良好。

- 是，执行**步骤17**。

- 否，执行**步骤18**。

步骤17 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤18**。

检查磁盘状态。

步骤18 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > quorumpeer”查看ZooKeeper实例所在的各节点主机信息。

步骤19 在FusionInsight Manager首页，单击“主机”。

步骤20 在“磁盘”列，检查ZooKeeper实例所在的各节点数据磁盘空间是否不足（使用率超过百分之80）。

- 是，执行**步骤21**。
- 否，执行**步骤23**。

步骤21 参考“ALM-12017 磁盘容量不足”进行处理，对磁盘进行扩容。

步骤22 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤23**。

收集故障信息。

步骤23 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤24 在“服务”中勾选待操作集群的如下节点信息。（普通模式集群不需要下载KrbServer日志。）

- ZooKeeper
- KrbServer

步骤25 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤26 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.65 ALM-13001 ZooKeeper 可用连接数不足

告警解释

系统每60秒周期性检测ZooKeeper服务连接数状态，当检测到ZooKeeper实例连接数超出阈值（最大连接数的80%）时产生该告警。

平滑次数为1，ZooKeeper可用连接数小于或等于阈值时，告警恢复；平滑次数大于1，ZooKeeper可用连接数小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
13001	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

ZooKeeper可用连接数不足，当连接率超过100%时无法处理外部连接。

可能原因

该节点ZooKeeper连接量过大，超过阈值。某些连接进程存在连接泄露，或配置的最大连接数不符合实际使用场景。

处理步骤

检查连接状态。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ZooKeeper可用连接数不足”所在行的下拉菜单，在定位信息中确认告警上报的主机名所在的节点IP地址。

步骤2 获取ZooKeeper进程pid。以**root**用户登录到告警上报的节点，执行命令：**pgrep -f proc_zookeeper**。

步骤3 是否正常获取pid。

- 是，执行**步骤4**。
- 否，执行**步骤15**。

步骤4 获取所有与当前ZooKeeper实例连接的IP及连接数量，取连接数最多的前十个进行检查。根据获取到的pid值，执行命令**lsof -i|grep** *$pid* **| awk '{print $9}' | cut -d : -f 2 | cut -d \> -f 2 | awk '{a[$1]++} END {for(i in a){print i,a[i] | "sort -r -g -k 2"}}' | head -10** 。（$pid为上一步获取的pid值）

步骤5 获取节点IP与连接数是否成功。

- 是，执行**步骤6**。
- 否，执行**步骤15**。

步骤6 获取连接进程的端口号。根据获取到的pid与IP值，执行命令**lsof -i|grep** *$pid* **| awk '{print $9}'|cut -d \> -f 2 |grep** *$IP*| **cut -d : -f 2** 。（$pid与$IP为上一步获取的pid值与IP值）

步骤7 获取端口号port成功。

- 是，执行**步骤8**。
- 否，执行**步骤15**。

步骤8 获取连接进程的进程号。依次登录到各IP，根据获取到的port号，执行命令**lsof -i| grep** *$port*。（$port为上一步获取端口号）

步骤9 获取进程号成功。

- 是，执行**步骤10**。
- 否，执行**步骤15**。

步骤10 根据获取到的进程号，查看进程是否存在连接泄露。

- 是，执行**步骤11**。
- 否，执行**步骤12**。

步骤11 将存在连接泄露的进程关掉，观察界面上告警是否消除。

- 是，处理完毕。
- 否，执行**步骤12**。

步骤12 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > quorumpeer > 性能”中，将“maxCnxns”的值根据实际情况调大。

步骤13 保存配置，并重启ZooKeeper服务。

步骤14 界面上告警是否消除。

- 是，处理完毕。
- 否，执行**步骤15**。

收集故障信息。

步骤15 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤16 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.66 ALM-13002 ZooKeeper 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测ZooKeeper服务直接内存使用状态，当检测到ZooKeeper实例直接内存使用率超出阈值（最大内存的80%）时产生该告警。

平滑次数为1，ZooKeeper直接内存使用率小于阈值时，告警恢复；平滑次数大于1，ZooKeeper直接内存使用率小于阈值的80%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
13002	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

ZooKeeper可用内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点ZooKeeper实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ZooKeeper直接内存使用率超过阈值”所在行的下拉菜单。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 实例 > quorumpeer（对应上报告警实例ip）”。单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“ZooKeeper堆内存与直接内存使用率”，单击“确定”，查看直接内存使用情况。

步骤3 查看ZooKeeper使用的直接内存是否已达到ZooKeeper设定的最大直接内存的80%?

- 是，执行**步骤4**。
- 否，执行**步骤8**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > quorumpeer > 系统”。查看“GC_OPTS”参数中是否存在“-XX:MaxDirectMemorySize”。

- 是，在“GC_OPTS”中把参数“-XX:MaxDirectMemorySize”删除。执行**步骤5**。
- 否，执行**步骤6**。

步骤5 保存配置，并重启ZooKeeper服务。

步骤6 查看告警信息，是否存在“ALM-13004 ZooKeeper堆内存使用率超过阈值”告警。

- 是，按照“ALM-13004 ZooKeeper堆内存使用率超过阈值”告警进行处理。
- 否，执行**步骤7**。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.67 ALM-13003 ZooKeeper 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测ZooKeeper进程的垃圾回收（GC）占用时间，当检测到ZooKeeper进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
13003	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

ZooKeeper进程的垃圾回收时间过长，可能影响该ZooKeeper进程正常提供服务。

可能原因

该节点ZooKeeper实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ZooKeeper进程垃圾回收（GC）时间超过阈值”所在行的下拉菜单。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 实例 > quorumpeer”。单击图表区域右上角的下拉菜单，选择“定制 >

GC”，勾选“ZooKeeper垃圾回收（GC）时间”，单击“确定”，查看ZooKeeper每分钟的垃圾回收时间统计情况。

步骤3 查看ZooKeeper每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤8**。

步骤4 请先排查应用程序是否存在内存泄露等问题。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > quorumpeer > 系统”。将“GC_OPTS”参数值根据实际情况调大。

说明

-Xmx一般配置为ZooKeeper数据容量的2倍，如果ZooKeeper容量达到2G，则GC_OPTS建议配置为：

-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=512M -XX:MetaspaceSize=64M -XX:MaxMetaspaceSize=64M -XX:CMSFullGCsBeforeCompaction=1

步骤6 保存配置，并重启ZooKeeper服务。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.68 ALM-13004 ZooKeeper 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测ZooKeeper服务堆内存使用状态，当检测到ZooKeeper实例堆内存使用率超出阈值（最大内存的95%）时产生该告警。

堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
13004	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

ZooKeeper可用内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点ZooKeeper实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ZooKeeper堆内存使用率超过阈值”所在行的下拉菜单，在定位信息中确认告警上报的主机名所在的节点IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 实例”，单击对应IP地址的“角色”列的“quorumpeer”。单击图表区域右上角的下拉菜单，选择“定制 > CPU 和内存”，勾选“ZooKeeper堆内存与直接内存使用率”，单击“确定”，查看堆内存使用情况。

步骤3 查看ZooKeeper使用的堆内存是否已达到ZooKeeper设定的最大堆内存的95%。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > quorumpeer > 系统”。将GC_OPTS参数中-Xmx的值根据实际情况调大，具体调整方案如下：

1. 单击“实例”，选择对应IP地址的“角色”列的“quorumpeer”，单击图表区域右上角的下拉菜单，选择“定制 > CPU 和内存”，勾选“ZooKeeper堆内存与直接内存资源状况”，单击“确定”，查看ZooKeeper实际使用的堆内存大小。
2. 根据堆内存实际使用量，修改GC_OPTS参数中的-Xmx值，该值一般为Zookeeper数据容量的2倍。例如当前ZooKeeper堆内存使用达到2G，则GC_OPTS建议配置为“-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=512M -XX:MetaspaceSize=64M -XX:MaxMetaspaceSize=64M -XX:CMSFullGCsBeforeCompaction=1”。

步骤5 保存配置，并重启ZooKeeper服务。

步骤6 观察界面告警是否清除?

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.69 ALM-13005 ZooKeeper 中组件顶层目录的配额设置失败

告警解释

系统每5小时周期性为组件和“customized.quota”配置项中的每个ZooKeeper顶层目录设置配额，当设置某个目录的配额失败时，会产生该告警。

当设置失败的目录重新设置配额成功时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
13005	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
服务目录	产生告警的目录名称。
Trigger Condition	产生告警的具体原因。

对系统的影响

组件可以向对应的ZooKeeper顶层目录中写入大量数据，导致Zookeeper服务不可用。

可能原因

告警目录对应的配额值不合理。

处理步骤

检查告警目录对应的配额值是否合理。

步骤1 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > 配额”。查看“customized.quota”配置项中，是否有产生该告警的告警目录及对应的配额值。

- 是，执行**步骤5**。
- 否，执行**步骤2**。

步骤2 查看下表中的组件告警目录列中，是否有产生该告警的告警目录。

表 10-87 各组件告警目录

组件名称	组件告警目录
Hbase	/hbase
Hive	/beelinesql
Yarn	/rmstore
Storm	/stormroot
Streaming	/storm
Kafka	/kafka

- 是，执行**3**。
- 否，执行**7**。

步骤3 查看该表中告警目录对应的组件名称，并打开其相应的服务界面，选择“配置 > 全部配置”，右上角搜索框输入“zk.quota”，搜索结果就是该告警目录对应的配额值。

步骤4 检查产生告警的目录对应的配额值是否不合理。合理的配额值应该大于等于目录当前的实际使用值，该值可以在告警参数“Trigger Condition”中获取。

步骤5 根据告警信息的提示，修改不合理的配额值，并保存配置。

步骤6 等待配置项“service.quotas.auto.check.cron.expression”中指定的定时时长后，查看告警是否消失。

- 是，处理完毕。
- 否，执行**7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.70 ALM-13006 Znode 数量或容量超过阈值

告警解释

系统每4小时周期性检测ZooKeeper服务数据目录下二级znode状态，当检测到二级Znode数量或者容量超过阈值时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
13006	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
服务目录	产生告警的目录名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

向ZooKeeper数据目录空间写入大量数据，导致ZooKeeper无法对外正常提供服务。

可能原因

往ZooKeeper数据目录空间写入大量数据，或者自定义阈值设置不合理。

处理步骤

检查告警目录是否写入大量数据

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“Znode数量或容量超过阈值”所在行的下拉菜单，在定位信息中确认告警上报的Znode。

步骤2 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper”，单击“资源”，在表“资源使用（按二级Znode）”中，查看告警对应Znode是否被写入较多数据。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 登录ZooKeeper客户端，删除告警对应Znode下的无用数据。

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper”，在“资源”的“资源使用(按二级Znode)”中，选择“ > 按Znode数量”，进入“按Znode数量”的“阈值设置”页面，单击“操作”下的“修改”。参考“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > 配额”中参数“max.znode.count”的值，调大阈值。

步骤5 在“资源使用(按二级Znode)”中，选择“ > 按Znode数量”，进入“按容量”的“阈值设置”页面，单击“操作”下的“修改”。参考“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > 配额”中参数“max.data.size”的值，调大阈值。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.71 ALM-13007 ZooKeeper 客户端可用连接数不足

告警解释

系统每60秒周期性检测ZooKeeper客户端连接到ZooKeeper服务器上的活动进程数，当检测到连接数目超过阈值时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
13007	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
ClientIP	客户端IP。
ServerIP	服务端IP。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

大量进程连接到ZooKeeper，导致ZooKeeper连接数被占满，无法对外正常提供服务。

可能原因

客户端大量进程连接到ZooKeeper，或者自定义阈值设置不合理。

处理步骤

检查客户端是否存在大量进程连接ZooKeeper的情况

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ZooKeeper客户端可用连接数不足”所在行的下拉菜单，在定位信息中确认告警上报的主机名所在的节点IP地址。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper”，单击“资源”进入资源页面，在表“连接数（按客户端IP）”中查看告警对应客户端IP的连接数是否较大。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 请确认并排查该客户端是否存在进程连接泄露的情况。

步骤4 单击“连接数（按客户端IP）”中的，进入“阈值设置”页面，单击“操作”下的“修改”。参考“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置 > quorumpeer”中参数“ maxClientCnxns”的值，调大阈值。

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.72 ALM-13008 ZooKeeper Znode 数量使用率超出阈值

告警解释

系统每小时周期性检测ZooKeeper服务数据目录下二级znode状态，当检测到二级znode的总数量超过阈值时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
13008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
服务目录	产生告警的目录名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

向ZooKeeper数据目录空间写入大量数据，导致ZooKeeper无法对外正常提供服务。

可能原因

- 往ZooKeeper数据目录空间写入大量数据。
- 自定义阈值设置不合理。

处理步骤

检查告警目录是否写入大量数据

步骤1 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper”，单击“资源”，在精细化监控“资源使用（按二级Znode）”中单击“按Znode数量”，查看监控中是否有顶级Znode被写入较多数据。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 登录FusionInsight Manager，选择“运维 > 告警 > 告警”，打开告警“ALM-13008 ZooKeeper Znode数量使用率超出阈值”左侧下拉菜单，在“定位信息”的“服务目录”中获取告警的Znode路径。

步骤3 以集群用户登录ZooKeeper客户端，删除告警对应Znode下的无用数据。

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置”，搜索“max.znode.count”，即ZooKeeper目录的数量配额的最大值，告警阈值为该值的80%，修改调大该配置项，单击“保存”，重启服务使配置生效。

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.73 ALM-13009 ZooKeeper Znode 容量使用率超出阈值

告警解释

系统每小时周期性检测ZooKeeper服务数据目录下二级znode状态，当检测到二级znode的总容量超过阈值时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
13009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
服务目录	产生告警的目录名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

向ZooKeeper数据目录空间写入大量数据，导致ZooKeeper无法对外正常提供服务。

可能原因

- 往ZooKeeper数据目录空间写入大量数据。
- 自定义阈值设置不合理。

处理步骤

检查告警目录是否写入大量数据

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ALM-13009 ZooKeeper Znode容量使用率超出阈值”所在行的下拉菜单，在定位信息中确认告警上报的Znode。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper”，单击“资源”，在精细化监控“资源使用（按二级Znode）”中，单击“按容量”查看顶级Znode目录是否被写入较多数据。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 登录FusionInsight Manager，选择“运维 > 告警 > 告警”，打开告警“ALM-13009 ZooKeeper Znode容量使用率超出阈值”左侧下拉菜单，在“定位信息”的“服务目录”中获取告警的Znode路径。

步骤4 以集群用户登录ZooKeeper客户端，删除告警对应Znode下的无用数据。

步骤5 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置 > 全部配置”，然后搜索“max.data.size”即“ZooKeeper目录的容量配额的最大值”，单位为Byte。然后搜索“GC_OPTS”配置项,查看其中“Xmx”的值。

步骤6 比较“max.data.size”和“Xmx*0.65”的值的大小，较小的值乘以80%为ZooKeeper Znode容量的阈值，可适当修改这两项配置，增大阈值。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.74 ALM-13010 配置 quota 的目录 Znode 使用率超出阈值

告警解释

系统每小时周期性检测配置quota的所有服务目录的znode数量，当检测到某个二级znode的数量使用率超过阈值时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
13010	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
服务目录	产生告警的目录名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

向ZooKeeper数据目录空间写入大量数据，导致ZooKeeper无法对外正常提供服务。

可能原因

- 往ZooKeeper数据目录空间写入大量数据。
- 自定义阈值设置不合理。

处理步骤

检查告警目录是否写入大量数据

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ALM-13010 配置quota的目录Znode使用率超出阈值”所在行的下拉菜单，在定位信息中确认告警上报的Znode。

步骤2 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper”，单击“资源”，在表“资源使用（按二级Znode）”中，查看告警对应顶级Znode是否被写入较多数据。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤3 登录FusionInsight Manager，选择“运维 > 告警 > 告警”，打开告警“ALM-13010 配置quota的目录Znode使用率超出阈值”左侧下拉菜单，在“定位信息”的“服务目录”中获取告警的Znode路径。

步骤4 以集群用户登录ZooKeeper客户端，删除告警对应Znode下的无用数据。

步骤5 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > *告警对应的顶级Znode服务组件*”，在该服务的“配置”页面中，单击“全部配置”，搜索“zk.quota.number”配置项，调大服务在ZooKeeper上的顶层目录的数量配额，单击“保存”。

须知

如果告警对应的顶级Znode服务组件为ClickHouse，则请修改“clickhouse.zookeeper.quota.node.count”参数的配置项。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“ZooKeeper”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.75 ALM-14000 HDFS 服务不可用

告警解释

系统每60秒周期性检测NameService的服务状态，当检测到所有的NameService服务都异常时，就会认为HDFS服务不可用，此时产生该告警。

至少一个NameService服务正常后，系统认为HDFS服务恢复，告警清除。

告警属性

告警ID	告警级别	是否自动清除
14000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

无法为基于HDFS服务的HBase和MapReduce等上层部件提供服务。用户无法读写文件。

可能原因

- ZooKeeper服务异常。
- 所有NameService服务异常。

处理步骤

检查ZooKeeper服务状态。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”页面，查看系统是否上报“ALM-13000 ZooKeeper服务不可用”告警。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 参考“ALM-13000 ZooKeeper服务不可用”对ZooKeeper服务状态异常进行处理，然后查看ZooKeeper服务的运行状态是否恢复为“良好”。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在“运维 > 告警”页面，查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

处理NameService服务异常告警。

步骤4 在FusionInsight Manager首页，选择“运维 > 告警”页面，查看是否有“ALM-14010 NameService服务异常”告警。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 按照“ALM-14010 NameService服务异常”的处理方法，依次对这些服务异常的NameService进行处理，然后查看是否消除各个NameService服务异常告警。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤6 在“运维 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的如下节点信息。

- ZooKeeper
- HDFS

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.76 ALM-14001 HDFS 磁盘空间使用率超过阈值

告警解释

系统每30秒周期性检测HDFS磁盘空间使用率，并把实际的HDFS磁盘空间使用率和阈值相比较。HDFS磁盘使用率指标默认提供一个阈值范围。当HDFS磁盘空间使用率超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

平滑次数为1，HDFS磁盘使用率小于或等于阈值时，告警恢复；平滑次数大于1，HDFS磁盘使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14001	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS集群磁盘容量不足，会影响到HDFS的数据写入。

可能原因

HDFS集群配置的磁盘空间不足。

处理步骤

查看磁盘容量，清除无用文件。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤2 在“图表”区域“实时”栏中，通过监控项“HDFS磁盘容量比率”查看HDFS磁盘使用率是否超过阈值（默认为80%）。若未查看到该监控项，可单击图表区域右上角的下拉菜单，选择“定制 > 磁盘”，勾选“HDFS磁盘容量比率”。

- 是，执行**步骤3**。
- 否，执行**步骤11**。

步骤3 在“基本信息”区域，单击发生故障的NameService的“NameNode(主)”，进入HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤4 在HDFS WebUI，单击“Datanodes”，在“Block pool used”列查看所有DataNode节点的磁盘使用率，判断有无DataNode节点的磁盘使用率超过阈值。

- 是，执行**步骤6**。
- 否，执行**步骤11**。

步骤5 以**root**用户登录集客户端所在节点的主机。

步骤6 执行命令**cd /opt/Bigdata/client**进入客户端安装目录，然后执行**source bigdata_env**。如果集群采用安全版本，要进行安全认证。执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤7 执行**hdfs dfs -rm -r** *文件或目录路径*命令，确认删除无用的文件。

步骤8 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

对系统进行扩容。

步骤9 对磁盘进行扩容。

步骤10 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的如下节点信息。

- ZooKeeper
- HDFS

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.77 ALM-14002 DataNode 磁盘空间使用率超过阈值

告警解释

系统每30秒周期性检测DataNode磁盘空间使用率，并把实际磁盘使用率和阈值相比较。DataNode磁盘空间使用率指标默认提供一个阈值范围。当检测到DataNode磁盘空间使用率指标超出阈值范围时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

平滑次数为1，DataNode磁盘空间使用率指标的值小于或等于阈值时，告警恢复；平滑次数大于1，DataNode磁盘空间使用率指标的值小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14002	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

DataNode容量不足，会影响到HDFS的数据写入。

可能原因

- 集群磁盘容量已满。
- DataNode节点间数据倾斜。

处理步骤

检查集群磁盘容量是否已满。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”页面，查看是否存在“ALM-14001 HDFS磁盘空间使用率超过阈值”告警。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 参考“ALM-14001 HDFS磁盘空间使用率超过阈值”进行处理，查看对应告警是否清除。

- 是，执行**步骤3**。
- 否，执行**步骤11**。

步骤3 在“运维 > 告警 > 告警”页面查看本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤4**。

检查DataNode节点平衡状态。

步骤4 在FusionInsight Manager首页，单击“主机”，查看各个机架上的DatNode节点数目分布是否大致相等，如果差异过大，调整DataNode节点所属机架，保证各个机架上的DataNode数量大致相等。重启HDFS服务生效。

步骤5 选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤6 在“基本信息”区域，单击“NameNode(主)”，进入HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤7 在HDFS WebUI的“Summary”区域，查看“DataNodes usages”中“Max”的值是否比“Median”的值大10%。

- 是，执行**步骤8**。
- 否，执行**步骤11**。

步骤8 数据倾斜，需要均衡集群中的数据。以**root**用户登录MRS客户端。如果集群为普通模式，执行**su - omm**切换到**omm**用户。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。如果集群采用安全版本，要进行安全认证。执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤9 执行以下命令，均衡数据分布：

hdfs balancer -threshold 10

步骤10 等待几分钟，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的“HDFS”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.78 ALM-14003 丢失的 HDFS 块数量超过阈值

告警解释

系统每30秒周期性检测丢失的块数量，并把丢失的块数量和阈值相比较。丢失的块数量指标默认提供一个阈值范围。当检测到丢失的HDFS块数量超出阈值范围时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

平滑次数为1，丢失的HDFS块数量小于或等于阈值时，告警恢复；平滑次数大于1，丢失的HDFS块数量小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14003	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS存储数据丢失，HDFS可能会进入安全模式，无法提供写服务。丢失的块数据无法恢复。

可能原因

- DataNode实例异常。
- 数据被删除。

处理步骤

检查DataNode实例。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”。

步骤2 查看所有DataNode实例的状态是否为“良好”。

- 是，执行**步骤11**。
- 否，执行**步骤3**。

步骤3 重启DataNode实例，查看能否成功启动。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 选择“运维 > 告警 > 告警”，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

删除被破坏的文件。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > NameNode(主)”，在HDFS的WebUI页面，查看列出的丢失块信息。

说明

- 如果有丢块，WebUI上会有一行红字显示。
- **admin**用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤6 用户确认丢失块所在的文件是否有用。

说明

MapReduce任务运行过程中在“/mr-history”、“/tmp/hadoop-yarn”、“/tmp/logs”这三个目录中生成的文件不属于有用文件。

- 是，执行**步骤7**。
- 否，执行**步骤8**。

步骤7 用户确认丢失块所在的文件是否已备份。

- 是，执行**步骤8**。
- 否，执行**步骤11**。

步骤8 以**root**用户登录HDFS客户端，用户密码为安装前用户自定义，请咨询系统管理员。执行如下命令：

- 安全模式：

  cd *客户端安装目录*

  source bigdata_env

  kinit hdfs

- 普通模式：

  su - omm

  cd *客户端安装目录*

  source bigdata_env

步骤9 在节点客户端执行**hdfs fsck / -delete**，删除丢失文件。如果丢失块所在的文件为有用文件，需要再次写入文件，恢复数据。

说明

删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。

步骤10 选择“运维 > 告警 > 告警”，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的“HDFS”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.79 ALM-14006 HDFS 文件数超过阈值

告警解释

系统每30秒周期性检测HDFS文件数，并把实际文件数和阈值相比较。当检测到HDFS文件数指标超出阈值范围时产生该告警。

平滑次数为1，HDFS文件数指标的值小于或等于阈值时，告警恢复；平滑次数大于1，HDFS文件数指标的值小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14006	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS文件数过多，磁盘存储不足可能造成数据入库失败。对HDFS系统性能产生影响。

可能原因

HDFS文件数超过阈值。

处理步骤

检查系统中文件数量。

步骤1 在FusionInsight Manager首页，查看当前的HDFS文件数。HDFS文件数可以通单击“集群 > *待操作集群的名称* > 服务 > HDFS”，单击图表区域右上角的下拉菜单，选择“定制 > 文件和块”，勾选“HDFS文件”和“HDFS块数”监控项查看。

步骤2 在“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”中查找“NameNode”下的GC_OPTS参数。

步骤3 配置文件对象数阈值：修改GC_OPTS参数中Xmx的值（Xmx内存值对应文件数阈值的公式为（y = 0.2007 x - 0.6312），其中x为内存数Xmx（GB），y为文件数（单位KW）。用户根据需要调整内存大小）。

步骤4 确认GC_PROFILE的值为custom，使GC_OPTS配置生效。单击“保存”，单击“更多 > 重启服务”重启服务。

步骤5 检查本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查系统中是否有不需要的文件。

步骤6 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**命令设置环境变量。

如果集群采用安全版本，要进行安全认证。

执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤7 执行**hdfs dfs -ls** *文件或目录路径*命令，检查该目录下的文件或目录是否是可以删除的无用文件。

- 是，执行**步骤8**。

- 否，执行**步骤9**。

步骤8 执行**hdfs dfs -rm -r** *文件或目录路径*命令。确认删除无用的文件后，等待文件在垃圾站中超过保留时间后（NameNode的配置参数“fs.trash.interval”指定了垃圾站中数据的保留时间），检查本告警是否清除。

说明

删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HDFS”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

NameNode JVM参数配置规则

NameNode JVM参数“GC_OPTS”默认值为：

-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=128M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -Djdk.tls.ephemeralDHKeySize=3072 -Djdk.tls.rejectClientInitiatedRenegotiation=true -Djava.io.tmpdir=${Bigdata_tmp_dir}

NameNode文件数量和NameNode使用的内存大小成比例关系，文件对象变化时请修改默认值中的“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”。参考值如下表所示。

表 10-88 NameNode JVM 配置

文件对象数量	参考值
10,000,000	“-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M”
20,000,000	“-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G”
50,000,000	“-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G”
100,000,000	“-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G”
200,000,000	“-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G”
300,000,000	“-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G”

10.13.80 ALM-14007 NameNode 堆内存使用率超过阈值

告警解释

系统每30秒周期性检测HDFS NameNode堆内存使用率，并把实际的HDFS NameNode堆内存使用率和阈值相比较。HDFS NameNode堆内存使用率指标默认提供一个阈值范围。当HDFS NameNode堆内存使用率超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

平滑次数为1，HDFS NameNode堆内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，HDFS NameNode堆内存使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

参数名称	参数含义
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS NameNode堆内存使用率过高，会影响HDFS的数据读写性能。

可能原因

HDFS NameNode配置的堆内存不足。

处理步骤

清除无用文件。

步骤1 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。

如果集群采用安全版本，要进行安全认证。

执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤2 执行**hdfs dfs -rm -r** *文件或目录路径*命令，确认删除无用的文件。

步骤3 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

查看NameNode JVM内存使用情况和当前配置。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤5 在“基本信息”区域，单击“NameNode(主)”，显示HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤6 在HDFS WebUI，单击“Overview”页签，查看Summary部分显示的HDFS中当前文件数量，目录数量和块数量信息。

步骤7 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”，在“搜索”中，输入“GC_OPTS”，确定当前“HDFS->NameNode”的“GC_OPTS”内存参数。

对系统进行调整。

步骤8 根据**步骤6**中的文件数据量和**步骤7**中NameNode配置的堆内存参数，检查当前配置的内存是否不合理。

- 是，执行**步骤9**。
- 否，执行**步骤11**。

说明

HDFS的文件对象数量（filesystem objects=files+blocks）和NameNode配置的JVM参数的对应关系建议如下：

- 文件对象数量达到10,000,000，则JVM参数建议配置为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
- 文件对象数量达到20,000,000，则JVM参数建议配置为：-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G
- 文件对象数量达到50,000,000，则JVM参数建议配置为：-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G
- 文件对象数量达到100,000,000，则JVM参数建议配置为：-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G
- 文件对象数量达到200,000,000，则JVM参数建议配置为：-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G
- 文件对象数量达到300,000,000，则JVM参数建议配置为：-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G

步骤9 按照文件对象数量和内存对应关系，对NameNode的堆内存参数进行修改，并单击“保存”，选择“概览 > 更多 > 重启服务”进行重启。

步骤10 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的如下节点信息。

- ZooKeeper
- HDFS

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.81 ALM-14008 DataNode 堆内存使用率超过阈值

告警解释

系统每30秒周期性检测HDFS DataNode堆内存使用率，并把实际的HDFS DataNode堆内存使用率和阈值相比较。HDFS DataNode堆内存使用率指标默认提供一个阈值范围。当HDFS DataNode堆内存使用率超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

平滑次数为1，HDFS DataNode堆内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，HDFS DataNode堆内存使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS DataNode堆内存使用率过高，会影响到HDFS的数据读写性能。

可能原因

HDFS DataNode配置的堆内存不足。

处理步骤

清除无用文件。

步骤1 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。

如果集群采用安全版本，要进行安全认证。

执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤2 执行**hdfs dfs -rm -r** *文件或目录路径*命令，确认删除无用的文件。

步骤3 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

查看DataNode JVM内存使用情况和当前配置。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤5 在“基本信息”区域，单击“NameNode(主)”，显示HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤6 在HDFS WebUI，单击“DataNodes”页签，查看所有告警DataNode节点的Block数量。

步骤7 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”。在“搜索”中，输入“GC_OPTS”，确定当前“HDFS->DataNode”的“GC_OPTS”内存参数。

对系统进行调整。

步骤8 根据**步骤6**中的Block数量和**步骤7**中DataNode配置的堆内存参数，检查当前配置的内存是否不合理。

- 是，执行**步骤9**。
- 否，执行**步骤11**。

说明

单个DataNode实例平均Block数量和DataNode内存的对应关系参考值如下：

- 单个DataNode实例平均Block数量达到2,000,000，DataNode的JVM参数参考值为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
- 单个DataNode实例平均Block数量达到5,000,000，DataNode的JVM参数参考值为：-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G

步骤9 按照Block数量和内存对应关系，对DataNode的堆内存参数进行修改，并单击“保存”，选择“概览 > 更多 > 重启服务”进行重启。

步骤10 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的“HDFS”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.82 ALM-14009 Dead DataNode 数量超过阈值

告警解释

系统每30秒周期性检测HDFS集群处于故障状态的DataNode数量，并把实际的故障状态的DataNode数量和阈值相比较。故障状态的DataNode数量指标默认提供一个阈值范围。当HDFS集群故障状态的DataNode数量超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

平滑次数为1，故障状态的DataNode数量小于或等于阈值时，告警恢复；平滑次数大于1，故障状态的DataNode数量小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

故障状态的DataNode节点无法提供HDFS服务。

可能原因

- DataNode故障或者负荷过高。
- NameNode和DataNode之间的网络断连或者繁忙。
- NameNode负荷过高。
- DataNode被删除后，没有重启NameNode。

处理步骤

查看DataNode是否故障。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤2 在“基本信息”区域，单击“NameNode(主)”，进入HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤3 在HDFS WebUI，单击“Datanodes”页签，在“In operation”区域，打开“Filter”下拉菜单，查看是否有“down”选项。

- 是，选择“down”，记录筛选出的DataNode节点的信息，执行**步骤4**。
- 否，执行**步骤8**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，在实例列表中，检查已记录的DataNode节点是否存在。

- 所有已记录的DataNode节点都存在时，执行**步骤5**。
- 所有已记录的DataNode节点都不存在时，执行**步骤6**。
- 部分已记录的DataNode节点存在时，执行**步骤7**。

步骤5 勾选对应的DataNode实例，选择“更多 > 重启实例”进行重启，重启结束后，查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

步骤6 勾选所有的NameNode实例，选择“更多 > 滚动重启实例”进行重启，重启结束后，查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤16**。

步骤7 勾选所有的NameNode实例，选择“更多 > 滚动重启实例”进行重启。重启完成后，勾选对应的DataNode实例，选择“更多 > 重启实例”进行重启，重启结束后，查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

查看NameNode和DataNode之间的网络情况。

步骤8 以**root**用户登录管理页面上存在且处于故障状态DataNode的业务平面IP节点，执行**ping** *NameNode的IP地址*命令以检查DataNode和NameNode之间的网络是否异常。

在FusionInsight Manager界面，单击“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，在实例列表中可查看处于故障状态DataNode的业务平面IP地址。

- 是，执行**步骤9**。
- 否，执行**步骤10**。

步骤9 修复网络故障，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤10**。

查看DataNode是否负荷过高。

步骤10 在FusionInsight Manager首页，单击“运维 > 告警 > 告警”，查看否存在“ALM-14008 HDFS DataNode内存使用率超过阈值”的告警。

- 是，执行**步骤11**。
- 否，执行**步骤13**。

步骤11 参考“ALM-14008 HDFS DataNode内存使用率超过阈值”的处理步骤，对该异常告警进行处理，查看是否消除该告警。

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 在告警列表中查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

查看NameNode是否负荷过高。

步骤13 在FusionInsight Manager首页，单击“运维 > 告警 > 告警”，查看是否存在“ALM-14007 HDFS NameNode内存使用率超过阈值”的告警。

- 是，执行**步骤14**。
- 否，执行**步骤16**。

步骤14 参考“ALM-14007 HDFS NameNode内存使用率超过阈值”的处理步骤，对该异常告警进行处理，查看是否消除告警。

- 是，执行**步骤15**。
- 否，执行**步骤16**。

步骤15 在告警列表中查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤16**。

收集故障信息。

步骤16 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤17 在“服务”中勾选待操作集群的“HDFS”。

步骤18 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤19 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.83 ALM-14010 NameService 服务异常

告警解释

系统每180秒周期性检测NameService服务状态，当检测到NameService服务不可用时产生该告警。

NameService服务恢复时，告警清除。

告警属性

告警ID	告警级别	是否自动清除
14010	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。

对系统的影响

无法为基于该NameService服务的HBase和MapReduce等上层部件提供服务。用户无法读写文件。

可能原因

- KrbServer服务异常。
- JournalNode节点故障。
- DataNode节点故障。
- 磁盘容量不足。
- NameNode节点进入安全模式。

处理步骤

检查KrbServer服务状态。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务”。

步骤2 查看KrbServer服务是否存在。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 单击“KrbServer”。

步骤4 单击“实例”。在KrbServer管理页面，选择故障实例，选择“更多 > 重启实例”。查看实例能否成功启动。

- 是，执行**步骤5**。
- 否，执行**步骤24**。

步骤5 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤6**。

检查JournalNode实例状态。

步骤6 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务”。

步骤7 选择“HDFS > 实例”。

步骤8 在实例页面中，查看JournalNode的“运行状态”是否为“良好”。

- 是，执行**步骤11**。
- 否，执行**步骤9**。

步骤9 选择故障的JournalNode，选择“更多 > 重启实例”。查看JournalNode能否成功启动。

- 是，执行**步骤10**。
- 否，执行**步骤24**。

步骤10 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

检查DataNode实例状态。

步骤11 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤12 单击“实例”查看所有DataNode节点的“运行状态”是否为“良好”。

- 是，执行**步骤15**。
- 否，执行**步骤13**。

步骤13 单击“实例”。在DataNode管理页面，选择故障DataNode，选择“更多 > 重启实例”。查看DataNode能否成功启动。

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤15**。

检查磁盘状态。

步骤15 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 主机”。

步骤16 在“磁盘”列，检查磁盘空间是否不足。

- 是，执行**步骤17**。
- 否，执行**步骤19**。

步骤17 对磁盘进行扩容。

步骤18 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤19**。

检查NameNode节点是否进入安全模式。

步骤19 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击服务异常的NameService的“NameNode(主)”，显示NameNode WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤20 在NameNode WebUI，查看是否显示如下信息：“Safe mode is ON.”

“Safe mode is ON.”表示安全模式已打开，后面的提示信息为告警信息，根据实际情况展现。

- 是，执行**步骤21**。
- 否，执行**步骤24**。

步骤21 以**root**用户登录客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。如果集群采用安全版本，要进行安全认证，执行**kinit hdfs**命令，按提示输入密码（向管理员获取密码）。如果集群采用非安全版本，需使用**omm**用户登录并执行命令，请确保**omm**用户具有客户端执行权限。

步骤22 执行**hdfs dfsadmin -safemode leave**。

步骤23 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤24**。

收集故障信息。

步骤24 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤25 在“服务”中勾选待操作集群的如下节点信息。

- ZooKeeper
- HDFS

步骤26 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤27 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.84 ALM-14011 DataNode 数据目录配置不合理

告警解释

DataNode的配置参数“dfs.datanode.data.dir”指定了DataNode的数据目录。当所配置的目录路径无法创建、与系统关键目录使用同一磁盘或多个目录使用同一磁盘时，系统即刻产生此告警。

当修改DataNode的数据目录合理后，重启该DataNode，告警清除。

告警属性

告警ID	告警级别	是否自动清除
14011	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

如果将DataNode数据目录挂载在根目录等系统关键目录，长时间运行后会将根目录写满，导致系统故障。

不合理的DataNode数据目录配置，会造成HDFS的性能下降。

可能原因

- DataNode数据目录创建失败。
- DataNode数据目录与系统关键目录（“/”或“/boot”）使用同一磁盘。
- DataNode数据目录中多个目录使用同一磁盘。

处理步骤

查看告警原因和产生告警的DataNode节点信息。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中单击此告警。

步骤2 通过“定位信息”的“主机名”，获取告警产生的DataNode节点的主机名。

删除DataNode数据目录中与磁盘规划不符的目录。

步骤3 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，在实例列表中单击产生告警的节点主机上的DataNode实例。

步骤4 单击“实例配置”，查看DataNode数据目录配置参数“dfs.datanode.data.dir”的值。

步骤5 查看所有的DataNode数据目录，是否有与磁盘规划不一致的目录。

- 是，执行**步骤6**。
- 否，执行**步骤9**。

步骤6 修改该DataNode节点的配置参数“dfs.datanode.data.dir”的值，删除错误的路径。

步骤7 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，重启该DataNode实例。

步骤8 检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

步骤9 以**root**用户登录到产生告警的DataNode的节点。

- 如果告警原因为“DataNode数据目录创建失败”，执行**步骤10**。
- 如果告警原因为“DataNode数据目录与系统关键目录（/或/boot）使用同一磁盘”，执行**步骤17**。
- 如果告警原因为“DataNode数据目录中多个目录使用同一磁盘”，执行**步骤21**。

检查DataNode数据目录是否创建失败。

步骤10 执行**su - omm**命令，切换到**omm**用户。

步骤11 使用**ls**命令查看DataNode数据目录中的每个目录是否存在。

- 是，执行**步骤26**。
- 否，执行**步骤12**。

步骤12 使用**mkdir** *数据目录*命令创建该目录，查看是否可以创建成功。

- 是，执行**步骤24**。
- 否，执行**步骤13**。

步骤13 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在告警“ALM-12017 磁盘容量不足”。

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 参考“ALM-12017 磁盘容量不足”对磁盘容量问题进行处理，查看“ALM-12017 磁盘容量不足”告警是否消除。

- 是，执行**步骤12**。
- 否，执行**步骤15**。

步骤15 查看**omm**用户对该目录的所有上层目录是否有“rwx”或者“x”权限。（例如“/tmp/abc/”，“tmp”目录有“x”权限，“abc”目录有“rwx”权限。）

- 是，执行**步骤24**。
- 否，执行**步骤16**。

步骤16 在**root**用户下，执行**chmod u+rwx** *path*或者**chmod u+x** *path*命令给这些路径添加**omm**用户的“rwx”或者“x”权限，然后执行**步骤12**。

检查DataNode数据目录是否与系统关键目录使用同一磁盘。

步骤17 分别使用**df**命令获取DataNode数据目录中的每个目录的磁盘挂载情况。

步骤18 查看命令结果的磁盘挂载目录是否为系统关键目录（“/”或“/boot”）。

- 是，执行**步骤19**。
- 否，执行**步骤24**。

步骤19 修改该DataNode节点的配置参数“dfs.datanode.data.dir”的值，删除与系统关键目录使用同一磁盘的目录。

步骤20 继续执行**步骤24**。

检查DataNode数据目录中是否多个目录使用同一磁盘。

步骤21 分别使用**df**命令获取DataNode数据目录中每个目录的磁盘挂载情况。记录命令结果的磁盘挂载目录。

步骤22 修改该DataNode节点的配置参数“dfs.datanode.data.dir”的值，对于其中磁盘挂载目录相同的DataNode目录，仅保留其中的一个目录，删除其他目录。

步骤23 继续执行**步骤24**。

重启DataNode，检查告警是否消除。

步骤24 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，重启该DataNode实例。

步骤25 检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤26**。

收集故障信息。

步骤26 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤27 在“服务”中勾选待操作集群的“HDFS”。

步骤28 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤29 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.85 ALM-14012 Journalnode 数据不同步

告警解释

在主NameNode节点上，系统每5分钟检测一次集群中所有JournalNode节点的数据同步性。如果有JournalNode节点的数据不同步，系统产生该告警。

当Journalnode数据同步5分钟后，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14012	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。

对系统的影响

当一个JournalNode节点工作状态异常时，其数据就会与其他JournalNode节点的数据不同步。如果超过一半的JournalNode节点的数据不同步时，NameNode将无法工作，导致HDFS服务不可用。

可能原因

- JournalNode实例不存在（被删除或被迁移）。
- JournalNode实例未启动或已停止。
- JournalNode实例运行状态异常。
- JournalNode节点的网络不可达。

处理步骤

查看JournalNode实例是否启动。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中单击此告警。

步骤2 查看“定位信息”，获取告警产生的JournalNode节点IP地址。

步骤3 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，在实例列表中查看告警节点上是否存在JournalNode实例。

- 是，执行**步骤5**。
- 否，执行**步骤4**。

步骤4 选择“运维 > 告警 > 告警”，在告警列表中单击此告警“操作”栏中的“清除”，在弹出窗口中单击“确定”，处理完毕。

步骤5 单击该JournalNode实例，查看其“配置状态”是否为“已同步”。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 勾选该JournalNode实例，单击“启动实例”，等待启动完成。

步骤7 等待5分钟后，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤15**。

查看JournalNode实例运行状态是否正常。

步骤8 查看该JournalNode实例的“运行状态”是否为“良好”。

- 是，执行**步骤11**。
- 否，执行**步骤9**。

步骤9 勾选该JournalNode实例，选择“更多 > 重启实例”，等待启动完成。

步骤10 等待5分钟后，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤15**。

查看JournalNode节点网络是否可达。

步骤11 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，查看主NameNode节点的业务IP地址。

步骤12 以**root**用户登录主NameNode节点。

步骤13 使用**ping**命令检查主NameNode与该JournalNode之间的网络状况，是否有超时或者网络不可达的情况。

ping *JournalNode的业务IP地址*

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 联系网络管理员处理网络故障，故障恢复后等待5分钟，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤15**。

收集故障信息。

步骤15 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤16 在“服务”中勾选待操作集群的“HDFS”。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.86 ALM-14013 NameNode FsImage 文件更新失败

告警解释

HDFS的元数据信息存储在NameNode数据目录（由配置项“dfs.namenode.name.dir”指定）中的FsImage文件中。备NameNode会周期将已有的FsImage和JournalNode中存储的Editlog合并生成新的FsImage，然后推送到主NameNode的数据目录。这个周期由HDFS的配置项“dfs.namenode.checkpoint.period”指定，默认为3600秒，即1个小时。如果主NameNode数据目录的FsImage没有更新，则说明HDFS元数据合并功能异常，需要修复。

在主NameNode节点上，系统每5分钟检测其上的FsImage文件的信息。如果在三个合并周期没有新的FsImage文件生成，则系统产生该告警。

当新的FsImage文件生成并成功推送到主NameNode，说明HDFS元数据合并功能恢复正常，告警自动恢复。

告警属性

告警ID	告警级别	是否自动清除
14013	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。

对系统的影响

如果主NameNode数据目录的FsImage没有更新，则说明HDFS元数据合并功能异常，需要修复。如不修复，HDFS在运行一段时间后，Editlog会一直增长。此时如果重启HDFS，由于要加载非常多的Editlog，会导致启动非常耗时。另外，该告警的产生也说明备NameNode功能异常，导致NameNode的HA机制失效。一旦主NameNode故障，则整个HDFS服务将不可用。

可能原因

- 备NameNode被停止。
- 备NameNode实例运行状态异常。
- 备NameNode合并新的FsImage失败。
- 备NameNode数据目录空间不足。
- 备NameNode推送FsImage到主NameNode失败。
- 主NameNode数据目录空间不足。

处理步骤

查看备NameNode是否被停止。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中单击此告警。

步骤2 在告警详情区域，查看“定位信息”，获取告警产生的主NameNode的主机名和所在的NameService名称。

步骤3 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，在实例列表中找到该NameService的备NameNode实例，查看其“配置状态”是否为“已同步”。

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 勾选该备NameNode实例，单击“启动实例”，等待启动完成。

步骤5 等待1个NameNode合并元数据的周期时间后，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

查看备NameNode实例运行状态是否正常。

步骤6 查看该备NameNode实例的“运行状态”是否为“良好”。

- 是，执行**步骤9**。
- 否，执行**步骤7**。

步骤7 勾选该备NameNode实例，单击“更多 > 重启实例”，等待启动完成。

步骤8 启动完成后，等待1个NameNode合并元数据的周期时间，然后查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤30**。

备NameNode合并新的FsImage是否失败。

步骤9 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”，搜索并获取“dfs.namenode.checkpoint.period”的值，该值即为NameNode合并元数据的周期。

步骤10 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，获取产生该告警的NameService的主、备NameNode节点的业务IP地址。

步骤11 单击“NameNode(*xx*,备)”，单击“实例配置”，获取配置项“dfs.namenode.name.dir”的值，该值即为备NameNode的FsImage存储目录。

步骤12 以**root**或**omm**用户登录备NameNode节点。

步骤13 进入到FsImage存储目录，查看最新的FsImage的生成时间。

cd *备NameNode存储目录*/**current**

stat -c %y $(ls -t | grep "fsimage_[0-9]*$" | head -1)

步骤14 执行**date**命令获取系统当前时间。

步骤15 计算最新FsImage的生成时间和当前时间的时间差，判断该时间差是否大于元数据合并周期的三倍。

- 是，执行**步骤16**。
- 否，执行**步骤20**。

步骤16 备NameNode合并元数据的功能异常。执行以下命令查看是否为存储空间不足造成。

进入到FsImage存储目录，查看最近一个的FsImage的大小（单位为MB）。

cd *备NameNode存储目录*/**current**

du -m $(ls -t | grep "fsimage_[0-9]*$" | head -1) | awk '{print $1}'

步骤17 执行命令查看备NameNode的磁盘剩余空间（单位为MB）。

df -m ./ | awk 'END{print $4}'

步骤18 对比FsImage的大小和目录剩余空间大小，看剩余空间是否还能存储一个FsImage文件。

- 是，执行**步骤7**。
- 否，执行**步骤19**。

步骤19 清理该目录所在磁盘的冗余文件，以便给元数据存放预留足够的空间。空间清理完毕后等待1个NameNode合并元数据的周期时间，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤20**。

查看备NameNode推送FsImage到主NameNode是否失败。

步骤20 以**root**用户登录备NameNode节点。

步骤21 执行**su - omm**命令切换到**omm**用户。

步骤22 使用如下命令查看备NameNode是否能将文件推送到主NameNode上。

tmpFile=/tmp/tmp_test_$(date +%s)

echo "test" > $tmpFile

scp $tmpFile *主NameNode的业务IP*:**/tmp**

- 是，执行**步骤24**。
- 否，执行**步骤23**。

步骤23 联系系统管理员，处理在**omm**用户下备NameNode无法推送数据到主NameNode的原因。故障恢复后等待1个NameNode合并元数据的周期时间，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤24**。

查看主NameNode数据目录空间是否不足。

步骤24 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，单击产生告警的NameService的主NameNode，单击“实例配置”，获取配置项“dfs.namenode.name.dir”的值，该值即为主NameNode的FsImage存储目录。

步骤25 以**root**或**omm**用户登录主NameNode节点。

步骤26 进入到FsImage存储目录，查看最近一个的FsImage的大小（单位为MB）。

cd *主NameNode存储目录*/**current**

du -m $(ls -t | grep "fsimage_[0-9]*$" | head -1) | awk '{print $1}'

步骤27 执行如下命令查看主NameNode的磁盘剩余空间（单位为MB）。

df -m ./ | awk 'END{print $4}'

步骤28 对比FsImage的大小和目录剩余空间大小，看剩余空间是否还能存储一个FsImage文件。

- 是，执行**步骤30**。
- 否，执行**步骤29**。

步骤29 清理该目录所在磁盘的冗余文件，以便给元数据存放预留足够的空间。空间清理完毕后等待1个NameNode合并元数据的周期时间，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤30**。

收集故障信息。

步骤30 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤31 在“服务”中勾选待操作集群的“NameNode”。

步骤32 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤33 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.87 ALM-14014 NameNode 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测NameNode进程的垃圾回收（GC）占用时间，当检测到NameNode进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14014	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NameNode进程的垃圾回收时间过长，可能影响该NameNode进程正常提供服务。

可能原因

该节点NameNode实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ALM-14014 NameNode进程垃圾回收（GC）时间超过阈值”所在行的下拉菜单，在“定位信息”中查看告警上报的角色名并确定实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > NameNode（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 垃圾回收”，勾选“NameNode垃圾回收（GC）时间”。查看NameNode每分钟的垃圾回收时间统计情况。

步骤3 查看NameNode每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置 > NameNode > 系统”。将“GC_OPTS”参数值根据实际情况调大。

说明

HDFS的文件对象数量（filesystem objects=files+blocks）和NameNode配置的JVM参数的对应关系建议如下：

- 文件对象数量达到10,000,000，则JVM参数建议配置为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
- 文件对象数量达到20,000,000，则JVM参数建议配置为：-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G
- 文件对象数量达到50,000,000，则JVM参数建议配置为：-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G
- 文件对象数量达到100,000,000，则JVM参数建议配置为：-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G
- 文件对象数量达到200,000,000，则JVM参数建议配置为：-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G
- 文件对象数量达到300,000,000，则JVM参数建议配置为：-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G

步骤5 保存配置，并重启该NameNode实例。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“NameNode”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.88 ALM-14015 DataNode 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测DataNode进程的垃圾回收（GC）占用时间，当检测到DataNode进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14015	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

DataNode进程的垃圾回收时间过长，可能影响该DataNode进程正常提供服务。

可能原因

该节点DataNode实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ALM-14015 DataNode进程垃圾回收（GC）时间超过阈值”所在行的下拉菜单，在“定位信息”中查看告警上报的角色名并确定实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > DataNode（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 垃圾回收”，勾选“DataNode垃圾回收（GC）时间”。查看DataNode每分钟的垃圾回收时间统计情况。

步骤3 查看DataNode每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置 > DataNode > 系统”。将“GC_OPTS”参数值根据实际情况调大。

说明

单个DataNode实例平均Block数量和DataNode内存的对应关系参考值如下：

- 单个DataNode实例平均Block数量达到2,000,000，DataNode的JVM参数参考值为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
- 单个DataNode实例平均Block数量达到5,000,000，DataNode的JVM参数参考值为：-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G

步骤5 保存配置，并重启该DataNode实例。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“DataNode”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.89 ALM-14016 DataNode 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测HDFS服务直接内存使用状态，当检测到DataNode实例直接内存使用率超出阈值（最大内存的90%）时，产生该告警。

直接内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14016	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

DataNode可用直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点DataNode实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ALM-14016 DataNode直接内存使用率超过阈值”所在行的下拉菜单，在“定位信息”中查看告警上报的角色名并确定实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > DataNode（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“DataNode内存使用详情”。查看直接内存使用情况。

步骤3 查看DataNode使用的直接内存是否已达到DataNode设定的最大直接内存的90%(默认阈值)。

- 是，执行**步骤4**。
- 否，执行**步骤8**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置 > DataNode > 系统”。查看“GC_OPTS”参数中是否存在“-XX:MaxDirectMemorySize”。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在“GC_OPTS”中把参数“-XX:MaxDirectMemorySize”删除。保存配置，并重启DataNode实例。

步骤6 查看告警信息，是否存在告警“ALM-14008 DataNode堆内存使用率超过阈值”。

- 是，参考“ALM-14008 DataNode堆内存使用率超过阈值”进行处理。
- 否，执行**步骤7**。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“DataNode”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.90 ALM-14017 NameNode 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测HDFS服务直接内存使用状态，当检测到NameNode实例直接内存使用率超出阈值（最大内存的90%）时，产生该告警。

直接内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14017	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NameNode可用直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点NameNode实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“ALM-14017 NameNode直接内存使用率超过阈值”所在行的下拉菜单，在“定位信息”中查看告警上报的角色名并确定实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > NameNode（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“NameNode内存使用详情”。查看直接内存使用情况。

步骤3 查看NameNode使用的直接内存是否已达到NameNode设定的最大直接内存的90%(默认阈值)。

- 是，执行**步骤4**。
- 否，执行**步骤8**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > NameNode > 全部配置”。查看“GC_OPTS”参数中是否存在“-XX:MaxDirectMemorySize”。

- 是，执行**步骤5**。

- 否，执行**步骤6**。

步骤5 在“GC_OPTS”中把参数“-XX:MaxDirectMemorySize”删除。保存配置，并重启NameNode实例。

步骤6 查看告警信息，是否存在告警“ALM-14007 NameNode堆内存使用率超过阈值”。

- 是，查看“ALM-14007 NameNode堆内存使用率超过阈值”进行处理。
- 否，执行**步骤7**。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“NameNode”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.91 ALM-14018 NameNode 非堆内存使用率超过阈值

告警解释

系统每30秒周期性检测HDFS NameNode非堆内存使用率，并把实际的HDFS NameNode非堆内存使用率和阈值相比较。HDFS NameNode非堆内存使用率指标默认提供一个阈值范围。当HDFS NameNode非堆内存使用率超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

当HDFS NameNode非堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14018	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS NameNode非堆内存使用率过高，会影响HDFS的数据读写性能。

可能原因

HDFS NameNode配置的非堆内存不足。

处理步骤

清除无用文件。

步骤1 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。

如果集群采用安全版本，要进行安全认证。

执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤2 执行**hdfs dfs -rm -r** *文件或目录路径*命令，确认删除无用的文件。

步骤3 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

查看NameNode JVM非堆内存使用情况和当前配置。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”，弹出“HDFS 服务状态”页面。

步骤5 在“基本信息”区域，单击“NameNode(主)”，显示HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤6 在HDFS WebUI，单击“Overview”页签，查看Summary部分显示的HDFS中当前文件数量，目录数量和块数量信息。

步骤7 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”，在“搜索”中，输入“GC_OPTS”，确定当前“HDFS->NameNode”的“GC_OPTS”非堆内存参数。

对系统进行调整。

步骤8 根据**步骤6**中的文件数据量和**步骤7**中NameNode配置的非堆参数，检查当前配置的非堆内存是否不合理。

- 是，执行**步骤9**。
- 否，执行**步骤12**。

说明

HDFS的文件对象数量（filesystem objects=files+blocks）和NameNode配置的JVM参数的对应关系建议如下：

- 文件对象数量达到10,000,000，则JVM参数建议配置为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
- 文件对象数量达到20,000,000，则JVM参数建议配置为：-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G
- 文件对象数量达到50,000,000，则JVM参数建议配置为：-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G
- 文件对象数量达到100,000,000，则JVM参数建议配置为：-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G
- 文件对象数量达到200,000,000，则JVM参数建议配置为：-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G
- 文件对象数量达到300,000,000，则JVM参数建议配置为：-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G

步骤9 按照文件对象数量和非堆内存对应关系，对NameNode的“GC_OPTS”参数进行修改。

步骤10 保存配置，选择“概览 > 更多 > 重启服务”。

步骤11 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤12**。

收集故障信息。

步骤12 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤13 在“服务”中勾选待操作集群的如下服务。

- ZooKeeper
- HDFS

步骤14 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤15 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.92 ALM-14019 DataNode 非堆内存使用率超过阈值

告警解释

系统每30秒周期性检测HDFS DataNode非堆内存使用率，并把实际的HDFS DataNode非堆内存使用率和阈值相比较。HDFS DataNode非堆内存使用率指标默认提供一个阈值范围。当HDFS DataNode非堆内存使用率超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

当HDFS DataNode非堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14019	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS DataNode非堆内存使用率过高，会影响HDFS的数据读写性能。

可能原因

HDFS DataNode配置的非堆内存不足。

处理步骤

清除无用文件。

步骤1 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。

如果集群采用安全版本，要进行安全认证。

执行**kinit hdfs**命令，按提示输入密码。向管理员获取密码。

步骤2 执行**hdfs dfs -rm -r** *文件或目录路径*命令，确认删除无用的文件。

步骤3 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

查看DataNode JVM内存使用情况和当前配置。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”。

步骤5 在“基本信息”区域，单击“NameNode(主)”，显示HDFS WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤6 在HDFS WebUI，单击“Datanodes”页签，查看所有告警DataNode节点的Block数量。

步骤7 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”，在“搜索”中，输入“GC_OPTS”，确定当前“HDFS->DataNode”的“GC_OPTS”内存参数。

对系统进行调整。

步骤8 根据**步骤6**中的Block数量和**步骤7**中DataNode配置的内存参数，检查当前配置的内存是否不合理。

- 是，执行**步骤9**。
- 否，执行**步骤12**。

说明

单个DataNode实例上的平均Block数量和DataNode内存的对应关系参考值如下：

- 单个DataNode实例平均Block数量达到2,000,000，DataNode的JVM参数参考值为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
- 单个DataNode实例平均Block数量达到5,000,000，DataNode的JVM参数参考值为：-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G

步骤9 按照Block数量和内存对应关系，对DataNode的内存参数“GC_OPTS”进行修改。

步骤10 保存配置，选择“概览 > 更多 > 重启服务”。

步骤11 检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤12**。

收集故障信息。

步骤12 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤13 在“服务”中勾选待操作集群的如下服务。

- ZooKeeper

- HDFS

步骤14 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤15 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.93 ALM-14020 HDFS 目录条目数量超过阈值

告警解释

系统每一个小时获取指定目录下直接子文件/目录的数量，判断其是否达到HDFS目录最大子文件/目录个数的百分比阈值（默认为“90%”），如果超过该阈值，则触发告警。

当发出告警的目录的子目录/文件数所占百分比低于阈值后，该告警将自动恢复。当监控开关关闭，所有目录对应的该告警都将自动恢复。当从监控列表中移除指定目录时，该目录对应的告警也会自动恢复。

说明

- HDFS目录的子文件/目录最大个数由参数“dfs.namenode.fs-limits.max-directory-items”指定，默认值为“1048576”。如果一个目录的子文件/目录数量超过该值，则无法再在该目录下创建新的子文件/目录。
- 要监控的目录列表由参数“dfs.namenode.directory-items.monitor”指定，默认值为“/tmp,/SparkJobHistory,/mr-history”。
- 监控开关由参数“dfs.namenode.directory-items.monitor.enabled”指定，默认值为“true”，即该检测默认开启。

告警属性

告警ID	告警级别	是否自动清除
14020	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
NameService名	产生告警的NameService名称。
目录名	产生告警的目录名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

当监控目录下的条目数超过系统阈值的90%触发该告警，但不影响对该目录继续增加条目。一旦超过系统最大阈值，继续增加条目会失败。

可能原因

监控目录的条目数超过系统阈值的90%。

处理步骤

检查系统中是否有不需要的文件。

步骤1 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**命令设置环境变量。

如果集群采用安全版本，要进行安全认证。

执行**kinit hdfs**命令，按提示输入密码（向管理员获取密码）。

步骤2 执行如下命令，检查发出告警的目录下的文件或目录是否是可以删除的无用文件。

hdfs dfs -ls *产生告警的目录路径*

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 执行如下命令。删除无用的文件。

hdfs dfs -rm -r -f *文件或目录路径*

说明

删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。

步骤4 等待1个小时，检查该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

检查系统阈值是否正确设置。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”，搜索“dfs.namenode.fs-limits.max-directory-items”参数，确定当前值配置是否合理。

- 是，执行**步骤9**。
- 否，执行**步骤6**。

步骤6 增大该参数值。

步骤7 保存配置，选择“概览 > 更多 > 重启服务”。

步骤8 等待1个小时，检查该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HDFS”，单击“确定”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.94 ALM-14021 NameNode RPC 处理平均时间超过阈值

告警解释

系统每30秒周期性检测NameNode的RPC处理平均时间，并把实际的NameNode的RPC处理平均时间和阈值（默认为100ms）相比较。当检测到NameNode的RPC处理平均时间连续多次（默认为10次）超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

如果平滑次数为1，NameNode的RPC处理平均时间小于或等于阈值时，告警恢复；如果平滑次数大于1，NameNode的RPC处理平均时间小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14021	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NameNode无法及时处理来自HDFS客户端、依赖于HDFS的上层服务、DataNode等的RPC请求，表现为访问HDFS服务的业务运行缓慢，严重时会导致HDFS服务不可用。

可能原因

- NameNode节点的CPU性能不足，导致NameNode无法及时处理消息。
- NameNode所设置的内存太小，频繁Full GC造成JVM卡顿。
- NameNode配置参数不合理，导致NameNode无法充分利用机器性能。

处理步骤

获取该告警的信息。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中找到该告警。

步骤2 单击该告警，查看下面的告警详情。从“定位信息”中的“主机名”信息可知发出该告警的NameNode节点主机名；从“定位信息”中的NameServiceName信息可知发出该告警的NameService名称。

查看阈值是否设置过低。

步骤3 查看依赖于HDFS的业务的运行状态是否正常运行。查看是否存在运行慢、执行任务超时的情况。

- 是，执行**步骤8**
- 否，执行**步骤4**

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击图表区域右上角的下拉菜单，选择“定制 > RPC”，在弹出的对话框中选择“主NameNode RPC处理平均时间”，单击“确定”。

步骤5 查看“主NameNode RPC处理平均时间”监控中，获取发出告警的NameService的当前的监控值。

步骤6 在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”，找到“主NameNode RPC处理平均时间”，单击default规则中“操作”栏中的“修改”，修改“阈值”为告警出现前后1天内监控值的峰值的150%。单击“确定”，保存新阈值。

步骤7 等待5分钟，查看该告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤8**

查看NameNode节点的CPU性能是否不足。

步骤8 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否有该NameNode节点的ID为“12016”“ALM-12016 CPU使用率超过阈值”告警。

- 是，执行**步骤9**
- 否，**步骤11**

步骤9 按照“ALM-12016 CPU使用率超过阈值”告警处理文档，处理该告警。

步骤10 处理完12016告警后，等待10分钟，查看该告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤11**

查看NameNode节点的内存是否设置过小。

步骤11 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否有该NameNode节点的ID为“14007”“ALM-14007 HDFS NameNode堆内存使用率超过阈值”告警。

- 是，执行**步骤12**
- 否，执行**步骤14**

步骤12 按照“ALM-14007 HDFS NameNode堆内存使用率超过阈值”告警处理文档，处理该告警。

步骤13 处理完14007告警后，等待10分钟，查看该告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤14**

查看该NameNode配置参数是否合理。

步骤14 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，搜索配置项“dfs.namenode.handler.count”，查看其值。如果值小于或等于128，则设置为128；如果大于128但小于192，则设置为192。

步骤15 搜索配置项“ipc.server.read.threadpool.size”，查看其值。如果值小于5，则设置为5。

步骤16 单击“保存”，单击“确定”。

步骤17 在HDFS的“实例”页面，先勾选发出该告警的NameService的备NameNode，在“更多”中单击“重启实例”，输入密码后单击“确定”，等待备NameNode启动完毕。

步骤18 在HDFS的“实例”页面，先勾选发出该告警的NameService的主NameNode，在“更多”中单击“重启实例”，输入密码后单击“确定”，等待主NameNode启动完毕。

步骤19 等待1小时，查看该告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤20**

收集故障信息。

步骤20 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤21 在“服务”中勾选待操作集群的如下节点信息。

- HDFS

步骤22 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤23 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.95 ALM-14022 NameNode RPC 队列平均时间超过阈值

告警解释

系统每30秒周期性检测NameNode的RPC队列平均时间，并把实际的NameNode的RPC队列平均时间和阈值（默认为200ms）相比较。当检测到NameNode的RPC队列平均时间连续多次（默认为10次）超出阈值范围时，产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”修改阈值。

如果平滑次数为1，NameNode的RPC队列平均时间小于或等于阈值时，告警恢复；如果平滑次数大于1，NameNode的RPC队列平均时间小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14022	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NameNode无法及时处理来自HDFS客户端、依赖于HDFS的上层服务、DataNode等的RPC请求，表现为访问HDFS服务的业务运行缓慢，严重时会导致HDFS服务不可用。

可能原因

- NameNode节点的CPU性能不足，导致NameNode无法及时处理消息。
- NameNode所设置的内存太小，频繁Full GC造成JVM卡顿。
- NameNode配置参数不合理，导致NameNode无法充分利用机器性能。
- HDFS的业务访问量太大，超过了NameNode的负载能力。

处理步骤

获取该告警的信息。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中找到该告警。

步骤2 单击该告警，查看下面的告警详情。从“产生时间”可知该告警的触发时间；从“定位信息”中的“主机名”信息可知发出该告警的NameNode节点主机名；从“定位信息”中的NameServiceName信息可知发出该告警的NameService名称。

查看是否阈值设置过低。

步骤3 查看依赖于HDFS的业务的运行状态是否正常运行。查看是否存在运行慢、执行任务超时的情况。

- 是，执行**步骤8**。
- 否，执行**步骤4**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击图表区域右上角的下拉菜单，单击“定制”，在弹出的对话框中选择“主NameNode RPC队列平均时间”，单击“确定”。

步骤5 查看“主NameNode RPC队列平均时间”监控中，获取发出告警的NameService的当前的监控值。

步骤6 在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS”，找到“主NameNode RPC队列平均时间”，单击default规则中“操作”栏中的“修改”，修改“阈值”为当前监控值的150%。单击“确定”，保存新阈值。

步骤7 等待1分钟，查看该告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤8**。

查看NameNode节点的CPU性能是否不足。

步骤8 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看该NameNode节点是否有“ALM-12016 CPU使用率超过阈值”告警。

- 是，执行**步骤9**。
- 否，执行**步骤11**。

步骤9 按照“ALM-12016 CPU使用率超过阈值”告警处理文档，处理该告警。

步骤10 处理完12016告警后，等待10分钟，查看14022告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤11**。

查看NameNode节点的内存是否设置过小。

步骤11 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否有该NameNode节点的“ALM-14007 HDFS NameNode堆内存使用率超过阈值”告警。

- 是，执行**步骤12**。
- 否，执行**步骤14**。

步骤12 按照“ALM-14007 HDFS NameNode堆内存使用率超过阈值”告警处理文档，处理该告警。

步骤13 处理完14007告警后，等待10分钟，查看14022告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤14**。

查看该NameNode配置参数是否合理。

步骤14 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，搜索配置项“dfs.namenode.handler.count”，查看其值。如果值小于或等于128，则设置为128；如果大于128但小于192，则设置为192。

步骤15 搜索配置项“ipc.server.read.threadpool.size”，查看其值。如果值小于5，则设置为5。

步骤16 单击“保存”，单击“确定”。

步骤17 在HDFS的“实例”页面，先勾选发出该告警的NameService的备NameNode，在“更多”中单击“重启实例”，输入密码后单击“确定”，等待备NameNode启动完毕。

步骤18 在HDFS的“实例”页面，先勾选发出该告警的NameService的主NameNode，在“更多”中单击“重启实例”，输入密码后单击“确定”，等待主NameNode启动完毕。

步骤19 等待1小时，查看该告警是否自动消除。

- 是，处理结束。

- 否，执行**步骤20**。

查看HDFS负载变化情况，适当降低HDFS负载。

步骤20 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击图表区域右上角的下拉菜单，选择“定制”，单击“RPC”，在弹出的对话框中选择“NameNode RPC队列平均时间”，单击“确定”。

步骤21 单击，进入监控详细信息界面。

步骤22 设置监控显示的时间段，从告警产生的时间的前5天开始，到告警产生时刻结束。单击“确定”按钮。

步骤23 在“NameNode RPC队列平均时间”监控中，查看该监控是否有开始急剧增加的时间点。

- 是，执行**步骤24**。
- 否，执行**步骤27**。

步骤24 确认并排查在该时间点，是否有新增任务大量访问HDFS，确认该任务是否可以调优，减少对HDFS的访问。

步骤25 如果在该时间点有执行Balancer，则可以停止Balancer，或指定节点执行Balancer任务，来降低对HDFS的负载。

步骤26 等待1小时，查看该告警是否自动消除。

- 是，处理结束。
- 否，执行**步骤27**。

收集故障信息。

步骤27 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤28 在“服务”勾选待操作集群的HDFS节点信息。

步骤29 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤30 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.96 ALM-14023 总副本预留磁盘空间所占比率超过阈值

告警解释

系统每30秒周期性检测总副本预留磁盘空间所占比率（总副本预留磁盘空间/（总副本预留磁盘空间+总剩余的磁盘空间））），并把实际的总副本预留磁盘空间所占比率和阈

值（默认为90%）相比较。当检测到总副本预留磁盘空间所占比率连续多次（平滑次数）高于阈值时，产生该告警。

如果平滑次数为1，总副本预留磁盘空间所占比率小于或等于阈值时，告警恢复；如果平滑次数大于1，总副本预留磁盘空间所占比率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14023	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS集群磁盘容量不足，会影响到HDFS的数据写入。如果DataNode的剩余空间都已经给副本预留，则写入HDFS数据失败。

可能原因

- 告警阈值配置不合理。
- HDFS集群配置的磁盘空间不足。
- HDFS的业务访问量太大，超过了已有DataNode的负载能力。

处理步骤

查看阈值设置是否合理

步骤1 在FusiongInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS > 磁盘 > 总副本预留磁盘空间所占比率”，查看该告警阈值设置是否合理（默认90%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 根据实际服务的使用情况，在“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HDFS > 磁盘 > 总副本预留磁盘空间所单击率”页面单击“修改”更改阈值。

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

查看是否有磁盘空间不足告警

步骤4 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”查看是否存在告警“ALM-14001 HDFS磁盘空间使用率超过阈值”或“ALM-14002 DataNode磁盘空间使用率超过阈值”。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 参考“ALM-14001 HDFS磁盘空间使用率超过阈值”或“ALM-14002 DataNode磁盘空间使用率超过阈值”进行处理，查看对应告警是否清除。

- 是，**步骤6**。
- 否，**步骤7**。

步骤6 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

对DataNode进行扩容

步骤7 对DataNode进行扩容。

步骤8 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HDFS”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后20分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.97 ALM-14024 租户空间使用率超过阈值

告警解释

系统每小时周期性检测租户所关联的每个目录的空间使用率（每个目录已使用的空间大小/每个目录分配的空间大小），并把每个目录实际的空间使用率和该目录设置的阈值相比较。当检测到租户所关联的目录空间使用率高于该目录设置的阈值时，产生该告警。

当上报告警的目录的空间使用率小于或等于该目录设置的阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14024	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名称。
租户名	产生告警的租户名称。
目录名	产生告警的目录名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

当监控的租户目录空间使用率超过用户自定义设置的阈值时触发该告警，但不影响对该目录继续写入文件。一旦超过该目录分配的最大存储空间，则HDFS写入数据会失败。

可能原因

- 告警阈值配置不合理。
- 租户分配的空间容量不合理

处理步骤

查看阈值设置是否合理

步骤1 查看告警定位信息，获取上报告警的租户名称，租户目录。

步骤2 在FusiongInsight Manager首页，在“租户资源”页面选择上报告警的租户名称，单击“资源”，查看上报告警的租户目录所对应的存储空间阈值配置设置是否合理（默认90%为合理值，用户可以根据自己的实际情况设置）。

- 是，执行**步骤5**。
- 否，执行**步骤3**。

步骤3 根据租户空间实际的使用情况，在“资源”页面单击“修改”修改或取消上报告警的租户目录所对应的存储空间阈值配置。

步骤4 等待1分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

查看租户分配的空间容量是否合理

步骤5 在FusionInsight Manager首页，在“租户资源”页面选择上报告警的租户名称，单击“资源”，查看上报告警的租户目录所对应的存储空间配额设置是否合理（根据该租户目录实际业务情况而定）。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 根据该租户目录实际业务情况，在“资源”页面单击“修改”修改上报告警的租户目录所对应的存储空间配额。

步骤7 等待1分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息

步骤8 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“HDFS”和Manager下的NodeAgent。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后20分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.98 ALM-14025 租户文件对象使用率超过阈值

告警解释

系统每小时周期性检测租户所关联的每个目录的文件对象使用率（每个目录已使用的文件对象个数/每个目录分配的文件对象个数），并把每个目录实际的文件对象使用率和该目录设置的阈值相比较。当检测到租户所关联的目录文件对象使用率高于该目录的阈值时，产生该告警。

当上报告警的目录的文件对象使用率小于或等于该目录设置的阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14025	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名称。
租户名	产生告警的租户名称。
目录名	产生告警的目录名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

当监控的租户目录下的文件对象使用率超过用户自定义设置的阈值时触发该告警，但不影响对该目录继续写入文件。一旦超过该目录分配的最大文件对象个数，则HDFS写入数据会失败。

可能原因

- 告警阈值配置不合理。
- 租户分配的文件目录数上限不合理

处理步骤

查看阈值设置是否合理

步骤1 查看告警定位信息，获取上报告警的租户名称，租户目录。

步骤2 在FusiongInsight Manager首页，单击“租户资源”页面选择上报告警的租户名称，单击“资源”，查看上报告警的租户目录所对应的文件数阈值配置设置是否合理（默认90%为合理值，用户可以根据自己的实际需求调节）。

- 是，执行**步骤5**。
- 否，执行**步骤3**。

步骤3 根据该租户该目录文件数的实际使用情况，在“资源”页面单击“修改”修改或取消上报告警的租户目录所对应的文件数阈值配置。

步骤4 等待1分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

查看租户分配的文件对象数是否合理

步骤5 在FusionInsight Manager首页，在“租户资源”页面选择上报告警的租户名称，单击“资源”，查看上报告警的租户目录所对应的文件目录数上限设置是否合理（根据该租户该目录实际业务情况而定）。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 根据租户该目录的实际业务情况，在“资源”页面单击“修改”修改或取消上报告警的租户目录所对应的文件目录数上限。

步骤7 等待1分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息

步骤8 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“HDFS”和Manager下的NodeAgent。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后20分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.99 ALM-14026 DataNode 块数超过阈值

告警解释

系统每30秒周期性检测每个DataNode上的块数，当检测到当前的DataNode节点上块数超过阈值时产生该告警。

如果平滑次数为1，DataNode节点上的块数小于或等于阈值时，告警恢复；如果平滑次数大于1，DataNode节点上的块数小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14026	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

上报DataNode块数超过阈值告警时，表示该DataNode节点上块数太多，继续写入可能会由于磁盘空间不足导致写入HDFS数据失败。

可能原因

- 告警阈值配置不合理。
- DataNode节点间数据倾斜。
- HDFS集群配置的磁盘空间不足。

处理步骤

修改阈值配置

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > HDFS > 配置 > 全部配置”，查找HDFS->DataNode下的GC_OPTS参数。

步骤2 配置DataNode块数阈值：修改GC_OPTS参数中Xmx的值（Xmx内存值对应节点块数阈值为每GB对应500000块数，用户根据需要调整内存值），确认GC_PROFILE的值为custom，保存配置。

步骤3 选择“集群 > *待操作集群的名称* > HDFS > 实例”勾选状态为“配置过期”的DataNode实例，选择“更多 > 重启实例”使GC_OPTS配置生效。

步骤4 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

查看是否有关联告警

步骤5 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”中查看是否存在告警“ALM-14002 DataNode磁盘空间使用率超过阈值”。

- 是，执行**步骤6**。
- 否，执行**步骤8**。

步骤6 参考“ALM-14002 DataNode磁盘空间使用率超过阈值”进行处理，查看对应告警是否清除。

- 是，执行**步骤7**。
- 否，执行**步骤8**。

步骤7 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

对DataNode进行扩容

步骤8 对DataNode进行扩容。

步骤9 在FusionInsight Manager首页，等待5分钟后，查看本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤10**。

收集故障信息

步骤10 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选待操作集群的“HDFS”。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后20分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

DataNode JVM参数配置规则

DataNode JVM参数“GC_OPTS”默认值为：

-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=128M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -Djdk.tls.ephemeralDHKeySize=2048

集群中每个DataNode实例平均保存的Blocks= HDFS Block * 3÷DataNode节点数，单个DataNode实例平均Block数量变化时请修改默认值中的“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”。参考值如下表所示。

表 10-89 DataNode JVM 配置

单个DataNode实例平均Block数量	参考值
2,000,000	“-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M”
5,000,000	“-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G”

Xmx内存值对应DataNode节点块数阈值，每GB对应500000块数，用户可根据需要调整内存值。

10.13.100 ALM-14027 DataNode 磁盘故障

告警解释

系统每60秒周期性检测DataNode节点上的磁盘状况，当检测到有磁盘出现故障时产生该告警。

当DataNode上故障磁盘都恢复正常后，手动清除该告警，并重启该DataNode。

告警属性

告警ID	告警级别	是否自动清除
14027	重要	否

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Failed Volumes	故障的磁盘列表。

对系统的影响

上报DataNode磁盘故障告警时，表示该DataNode节点上存在故障的磁盘分区，可能会导致已写入的文件丢失。

可能原因

- 硬盘故障。
- 磁盘权限设置不正确。

处理步骤

查看是否存在磁盘告警

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”查看是否存在告警“ALM-12014 设备分区丢失”或“ALM-12033 慢盘故障”。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 参考“ALM-12014 设备分区丢失”或“ALM-12033 慢盘故障”告警进行处理，查看对应告警是否清除。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 等待5分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

修改磁盘权限

步骤4 在“运维 > 告警 > 告警”页面，查看该告警的“定位信息”和“附加信息”，获取该告警上报的故障磁盘位置信息。

步骤5 以**root**用户登录上报告警的节点，进入故障磁盘所在目录，使用**ll**命令查看该故障磁盘的权限是否711，用户是否为**omm**。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 修改故障磁盘权限，如故障磁盘为data1，则执行以下命令：

chown omm:wheel data1

chmod 711 data1

步骤7 在Manager告警列表中，单击该告警“操作”列下面的“清除”，手动清除告警。然后选择“集群 > 服务 > HDFS > 实例”勾选该DataNode，选择“更多 > 重启实例”，等待5分钟，查看是否有新的告警上报。

- 否，处理完毕。
- 是，执行**步骤8**。

收集故障信息

步骤8 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“HDFS”和“OMS”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后20分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统不会自动清除此告警，需手工清除。

参考信息

无。

10.13.101 ALM-14028 待补齐的块数超过阈值

告警解释

系统每30秒周期性检测待补齐的块数量，并把待补齐的块数量和阈值相比较。需补齐的块数量指标默认提供一个阈值范围。当检测到丢失的块数量超出阈值范围时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群名称* > HDFS > 文件和块 > 需要复制副本的块总数（NameNode）”修改阈值。

平滑次数为1，待补齐的块数量小于或等于阈值时，告警恢复；平滑次数大于1，待补齐的块数量小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14028	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

HDFS存储数据丢失，HDFS可能会进入安全模式，无法提供写服务。丢失的块数据无法恢复。

可能原因

- DataNode实例异常。
- 数据被删除。
- 写入文件的副本数大于DataNode的节点数。

处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”查看是否存在告警“ALM-14003 丢失的HDFS块数量超过阈值”。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 按照“ALM-14003 丢失的HDFS块数量超过阈值”的处理方法处理，然后等待5分钟，检查告警是否清除。

- 是，结束。
- 否，执行**步骤3**。

步骤3 以root用户登录HDFS客户端，用户密码为安装前用户自定义，请咨询系统管理员。执行如下命令：

- 安全模式：

  cd *客户端安装目录*

  source bigdata_env

  kinit hdfs

- 普通模式：

  su - omm

  cd *客户端安装目录*

source bigdata_env

步骤4 执行命令**hdfs fsck / >> fsck.log**，获取当前集群的状况。

步骤5 使用命令统计当前待复制块数量M：

cat fsck.log | grep "Under-replicated"

步骤6 使用命令统计“/tmp/hadoop-yarn/staging/”目录下的待复制块数量N：

cat fsck.log | grep "Under replicated" | grep "/tmp/hadoop-yarn/staging/" | wc -l

说明

“/tmp/hadoop-yarn/staging/”目录为默认值，如果客户有修改，可以通过mapred-site.xml文件配置项“yarn.app.mapreduce.am.staging-dir”获取此路径。

步骤7 比对N是否占了M的大多数（N/M＞50%）。

- 是，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 执行命令来重新配置目录的文件副本数（文件副本数选择DataNode节点数或者默认文件副本数）：

hdfs dfs -setrep -w *文件副本数* **/tmp/hadoop-yarn/staging/**

说明

默认文件副本数通过如下方式获取：

登录Manager页面，选择“集群 > 服务 > HDFS > 配置 > 全部配置”，搜索dfs.replication参数，该参数的值即是默认文件副本数。

然后等待5分钟，检查告警是否清除。

- 是，结束。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HDFS”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.102 ALM-14029 单副本的块数超过阈值

告警解释

系统每4个小时周期性检测单副本块的数量，并把当前单副本的块数和阈值相比较。单副本的块数量指标默认提供一个阈值范围。当检测到单副本的块数量超出阈值范围时产生该告警。

待补齐的块数量小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
14029	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
NameService名	产生告警的NameService名称。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

单副本的数据在节点故障时容易丢失，单副本的文件过多会对HDFS文件系统的安全性造成影响。

可能原因

- DataNode节点故障。
- 磁盘故障。
- 单副本写入文件。

处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”查看是否存在告警“ALM-14003 丢失的HDFS块数量超过阈值”。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 按照“ALM-14003 丢失的HDFS块数量超过阈值”的处理方法处理，然后等待下个检测周期，检查告警是否清除。

- 是，结束。
- 否，执行**步骤3**。

步骤3 排查业务中是否写入过的单副本的文件。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 以**root**用户登录HDFS客户端，用户密码为安装前用户自定义，请咨询系统管理员。执行如下命令：

- 安全模式：

  cd *客户端安装目录*

  source bigdata_env

  kinit hdfs

- 普通模式：

  su - omm

  cd *客户端安装目录*

  source bigdata_env

步骤5 在客户端节点执行如下命令，增大单副本文件的副本数。

hdfs dfs -setrep -w *文件副本数 文件名或文件路径*

步骤6 等待下个检测周期，查看告警是否消除。

- 是，结束。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“HDFS”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.103 ALM-16000 连接到 HiveServer 的 session 数占最大允许数的百分比超过阈值

告警解释

系统每30秒周期性检测连接到HiveServer的Session数占HiveServer允许的最大session数的百分比，该指标可通过“集群 > *待操作集群的名称* > 服务 > Hive > 实例 > *具体的HiveServer实例*”查看。连接到HiveServer的session数占最大允许数的百分比指标默认提供一个阈值范围（90%），当检测到百分比指标超过阈值范围产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Hive > 连接到HiveServer的session数占最大允许session数的百分比”修改阈值。

平滑次数为1，百分比指标小于或等于阈值时，告警恢复；平滑次数大于1，百分比指标小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
16000	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

发生连接数告警时，表示连接到HiveServer的session数过多，将会导致无法建立新的连接。

可能原因

连接HiveServer的客户端过多。

处理步骤

增加Hive最大连接数配置。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置”，选择“全部配置”。

步骤2 然后查找“hive.server.session.control.maxconnections”，调大该配置项的数值。设该配置项的值为A，阈值为B，连接到HiveServer的session数为C，调整策略为A x B > C，连接到HiveServer的session数可在Hive的监控界面查看监控指标“HiveServer的session数统计”。

步骤3 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的“Hive”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.104 ALM-16001 Hive 数据仓库空间使用率超过阈值

告警解释

系统每30秒周期性检测Hive数据仓库空间使用率，该指标可在Hive服务监控界面查看，指标名称为“Hive已经使用的HDFS空间占可使用空间的百分比”。Hive数据仓库空间使用率指标默认提供一个阈值范围（85%），当检测到Hive数据仓库空间使用率超过阈值范围时产生该告警。

用户可通过“运维 >告警 >阈值设置 > *待操作集群的名称* > Hive > Hive已经使用的HDFS空间占可使用空间的百分比”修改阈值。

平滑次数为1，Hive数据仓库空间使用率小于或等于阈值时，告警恢复；平滑次数大于1，Hive数据仓库空间使用率小于或等于阈值的90%时，告警恢复。

说明

管理员可通过增加仓库容量或释放部分已使用空间的方式降低仓库空间使用率。

告警属性

告警ID	告警级别	是否自动清除
16001	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

系统可能无法正常写入数据，导致部分数据丢失。

可能原因

- Hive使用HDFS容量上限过小。
- HDFS空间不足。
- 部分数据节点瘫痪。

处理步骤

扩展系统配置。

步骤1 分析集群HDFS使用情况，增加HDFS分配给Hive使用的容量上限。

登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置”，选择“全部配置”，然后查找“hive.metastore.warehouse.size.percent”，调大该配置项。设配置项的值为A，HDFS总存储空间为B，阈值为C，Hive已经使用HDFS的空间大小为D。调整策略为A x B x C > D ，HDFS总存储空间可在HDFS NameNode页面查看，Hive已经使用HDFS的空间大小可在Hive的监控界面查看监控指标“Hive已经使用的HDFS空间大小”。

步骤2 检查该告警是否恢复。

- 是，操作结束。
- 否，执行**步骤3**。

对系统进行扩容。

步骤3 对系统进行扩容。

步骤4 检查该告警是否恢复。

- 是，操作结束。
- 否，执行**步骤5**。

检查数据节点是否正常。

步骤5 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”。

步骤6 查看是否有“ALM-12006 节点故障”、“ALM-12007 进程故障”、“ALM-14002 DataNode磁盘空间使用率超过阈值”告警。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 分别参考“ALM-12006 节点故障”、“ALM-12007 进程故障”、“ALM-14002 DataNode磁盘空间使用率超过阈值”的处理步骤处理告警。

步骤8 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“Hive”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.105 ALM-16002 Hive SQL 执行成功率低于阈值

告警解释

系统每30秒周期性检测执行的HQL成功百分比，HQL成功百分比由一个周期内Hive执行成功的HQL数/Hive执行HQL总数计算得到。该指标可通过“集群 > *待操作的集群名称* > 服务 > Hive > 实例 > *具体的HiveServer实例*”查看。执行的HQL成功百分比指标默认提供一个阈值范围（90%），当检测到百分比指标低于阈值范围产生该告警。在该告警的定位信息可查看产生该告警的主机名，该主机IP也是HiveServer节点IP。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Hive > 执行成功的HQL百分比”修改阈值。

当系统在一个检测周期检测到该指标高于阈值的110%时，恢复告警。

告警属性

告警ID	告警级别	是否自动清除
16002	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

系统执行业务能力过低，无法正常响应客户请求。

可能原因

- HQL命令语法错误。
- 执行Hive on HBase任务时HBase服务异常。
- 执行Hive on Spark任务时Spark服务异常。
- 依赖的基础服务HDFS、Yarn、ZooKeeper等异常。

处理步骤

检查HQL命令是否符合语法。

步骤1 在FusionInsight Manager界面选择“运维 > 告警”，查看告警详情，获取产生告警的节点信息。

步骤2 使用Hive客户端连接到产生该告警的HiveServer节点，查询Apache提供的HQL语法规范，确认输入的命令是否正确。详情请参见**https://cwiki.apache.org/confluence/display/hive/languagemanual**。

- 是，执行**步骤4**。
- 否，执行**步骤3**。

说明

若想查看执行错误语句的用户，可下载产生该告警的HiveServer节点的HiveServerAudit日志，下载的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟。打开日志文件查找“Result=FAIL”关键字筛选执行错误语句的日志信息，再根据日志信息中的“UserName”查看执行错误语句的用户。

步骤3 输入正确的HQL语句，观察命令是否正确执行。

- 是，执行**步骤12**。
- 否，执行**步骤4**。

检查HBase服务是否异常。

步骤4 与执行HQL命令的用户确认是否执行的是Hive on HBase任务。

- 是，执行**步骤5**。
- 否，执行**步骤8**。

步骤5 在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务”，在服务列表查看HBase服务状态是否正常。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 选择“运维 > 告警”，查看告警界面的HBase相关告警，参照对应告警帮助进行处理。

步骤7 输入正确的HQL语句，观察命令是否正确执行。

- 是，执行**步骤12**。
- 否，执行**步骤8**。

检查HDFS、Yarn、ZooKeeper等是否正常。

步骤8 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务”。

步骤9 在服务列表查看HDFS、Yarn、ZooKeeper等服务是否正常。

- 是，执行**步骤12**。
- 否，执行**步骤10**。

步骤10 查看告警界面的相关告警，参照对应告警帮助进行处理。

步骤11 输入正确的HQL语句，观察命令是否正确执行。

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 等待一分钟，查看本告警是否清除。

- 是，处理结束。
- 否，执行**步骤13**。

收集故障信息。

步骤13 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选待操作集群的如下节点信息。

- Mapreduce
- Hive

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.106 ALM-16003 Background 线程使用率超过阈值

告警解释

系统每30秒周期性检测Background线程使用率情况，默认阈值为90%。如果Hive使用的background线程池使用率超过阈值，则发出告警。

说明

MRS 3.X支持Hive多实例，若集群启用了多实例功能且安装了多个Hive服务，请根据“定位信息”的“服务名”值来确定具体产生告警的Hive服务。例如Hive1服务不可用，则“定位信息”中显示服务名=Hive1，处理步骤中的操作对象也应由Hive调整为Hive1。

告警属性

告警ID	告警级别	是否自动清除
16003	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

后台Background线程数过多，导致新提交的任务无法及时运行。

可能原因

Hive后台的background线程池使用率过大。

- HiveServer后台的background线程池执行的任务过多。
- HiveServer后台的background线程池的容量过小。

处理步骤

检查HiveServer background线程池执行任务数量

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > *具体的HiveServer实例*”，找到“Background线程数”与“Background线程使用率”监控信息。

步骤2 在Background线程数监控中，线程数目最近半小时时间内是否有异常偏高（默认队列数值为100，偏高数值>=90）。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 调整提交到background线程池的任务数（比如，取消一些后台性能低，耗时长的任务）。

步骤4 “Background线程数”和“Background线程数使用率”是否下降。

- 是，执行**步骤7**。
- 否，执行**步骤5**。

检查HiveServer background线程池容量。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > *具体的HiveServer实例*”，找到“Background线程数”与“Background线程使用率”监控信息。

步骤6 查看“${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/1_23_HiveServer/etc/hive-site.xml”文件中“hive.server2.async.exec.threads”数量，适当增大该数值（如：增大原数值的20%）。

步骤7 保存更新配置。

步骤8 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“Hive”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.107 ALM-16004 Hive 服务不可用

告警解释

系统每60秒周期性检测Hive服务状态。当Hive服务不可用时产生该告警。

当Hive服务恢复时，告警恢复。

说明

MRS 3.X支持Hive多实例，若集群启用了多实例功能且安装了多个Hive服务，请根据“定位信息”的“服务名”值来确定具体产生告警的Hive服务。例如Hive1服务不可用，则“定位信息”中显示服务名=Hive1，处理步骤中的操作对象也应由Hive调整为Hive1。

告警属性

告警ID	告警级别	是否自动清除
16004	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

系统无法提供数据加载，查询，提取服务。

可能原因

- Hive服务不可用可能与ZooKeeper、HDFS、Yarn和DBService等基础服务有关，也可能由Hive自身的进程故障引起。
  - ZooKeeper服务异常。
  - HDFS服务异常。

  - Yarn服务异常。
  - DBService服务异常。
  - Hive服务进程故障，如果告警由Hive进程故障引发，告警上报时间可能会延迟5分钟左右。
- Hive服务和基础服务间的网络通信中断。

处理步骤

检查HiveServer/MetaStore进程状态。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，在Hive实例列表中，查看所有HiveServer或MetaStore实例状态是否都呈现未知状态。

- 是，执行**步骤2**。
- 否，执行**步骤4**。

步骤2 在Hive实例列表上方，选择“更多 > 重启实例”，重启HiveServer/MetaStore进程。

步骤3 在告警列表中，查看“Hive服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤4**。

检查ZooKeeper服务状态。

步骤4 在FusionInsight Manager的告警列表中，查看是否有“进程故障”产生。

- 是，执行**步骤5**。
- 否，执行**步骤8**。

步骤5 在“进程故障”，查看“服务名”是否为“ZooKeeper”。

- 是，执行**步骤6**。
- 否，执行**步骤8**。

步骤6 参考“ALM-12007 进程故障”的处理步骤处理该故障。

步骤7 在告警列表中，查看“Hive服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

检查HDFS服务状态。

步骤8 在FusionInsight Manager的告警列表中，查看是否有“HDFS服务不可用”产生。

- 是，执行**步骤9**。
- 否，执行**步骤11**。

步骤9 参考“ALM-14000 HDFS服务不可用”的处理步骤处理该故障。

步骤10 在告警列表中，查看“Hive服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤11**。

检查Yarn服务状态。

步骤11 在FusionInsight Manager的告警列表中，查看是否有“Yarn服务不可用”产生。

- 是，执行**步骤12**。
- 否，执行**步骤14**。

步骤12 参考“ALM-18000 Yarn服务不可用”的处理步骤处理该故障。

步骤13 在告警列表中，查看“Hive服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤14**。

检查DBService服务状态。

步骤14 在FusionInsight Manager的告警列表中，查看是否有“DBService服务不可用”产生。

- 是，执行**步骤15**。
- 否，执行**步骤17**。

步骤15 参考“ALM-27001 DBService服务不可用”的处理步骤处理该故障。

步骤16 在告警列表中，查看“Hive服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤17**。

检查Hive与ZooKeeper、HDFS、Yarn和DBService之间的网络连接。

步骤17 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive”。

步骤18 单击“实例”。

显示HiveServer实例列表。

步骤19 单击“HiveServer”行的“主机名称”。

弹出HiveServer主机状态页面。

步骤20 记录“基本信息”下的IP地址。

步骤21 以**omm**用户通过**步骤20**获取的IP地址登录HiveServer所在的主机。

步骤22 执行**ping**命令，查看HiveServer所在主机与ZooKeeper、HDFS、Yarn和DBService服务所在主机的网络连接是否正常。（获取ZooKeeper、HDFS、Yarn和DBService服务所在主机的IP地址的方式和获取HiveServer IP地址的方式相同。）

- 是，执行**步骤25**。
- 否，执行**步骤23**。

步骤23 联系网络管理员恢复网络。

步骤24 在告警列表中，查看“Hive服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤25**。

收集故障信息。

步骤25 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤26 在“服务”中勾选待操作集群的如下节点信息。

- ZooKeeper
- HDFS
- Yarn
- DBService
- Hive

步骤27 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤28 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.108 ALM-16005 Hive 服务进程堆内存使用超出阈值

告警解释

系统每30秒周期性检测Hive堆内存使用率，并把实际的Hive堆内存使用率和阈值相比较。当Hive堆内存使用率超出阈值（默认为最大堆内存的95%）时产生该告警。

用户可通过“运维 >告警 > 阈值设置 > *待操作集群的名称* > Hive”修改阈值。

当Hive堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
16005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

参数名称	参数含义
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Hive堆内存使用率过高，会影响Hive任务运行的性能，甚至造成内存溢出导致Hive服务不可用。

可能原因

该节点Hive实例堆内存使用量过大，或分配的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“16005”的告警，查看“定位信息”中的角色名并确定实例的IP地址。

- 告警上报的角色是HiveServer，执行**步骤2**。
- 告警上报的角色是MetaStore，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的HiveServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“HiveServer内存使用率统计”，单击“确定”，查看HiveServer进程使用的堆内存是否已达到HiveServer进程设定的最大堆内存的阈值（默认95%）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的MetaStore，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“MetaStore内存使用率统计”，单击“确定”，查看MetaStore进程使用的堆内存是否已达到MetaStore进程设定的最大堆内存的阈值（默认95%）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置”，选择“全部配置”，选择“HiveServer/MetaStore > JVM”，将“HIVE_GC_OPTS/METASTORE_GC_OPTS”参数中“-Xmx”的值根据如下建议进行调整，并单击“保存”。

说明

1. HiveServer的GC参数配置建议
   - 当HiveServer进程使用的堆内存已达到HiveServer进程设定的堆内存的阈值时，将“-Xmx”值调整为默认值的2倍，比如：“-Xmx”默认设置为2G时，调整“-Xmx”的值为4G。在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Hive > CPU和内存 > HiveServer堆内存使用率统计 (HiveServer)”，可查看“阈值”。
   - 建议同时调节“-Xms”的值，使“-Xms”和“-Xmx”比值为1:2，这样可以避免JVM动态调整堆内存大小时影响性能。
2. MetaServer的GC参数配置建议
   - 当MetaStore进程使用的堆内存已达到MetaStore进程设定的堆内存的阈值时，将“-Xmx”值调整为默认值的2倍，比如：“-Xmx”默认设置为2G时，调整“-Xmx”的值为4G。在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Hive > CPU和内存 > MetaStore堆内存使用率统计 (MetaStore)”，可查看“阈值”。
   - 建议同时调节“-Xms”的值，使“-Xms”和“-Xmx”比值为1:2，这样可以避免JVM动态调整堆内存大小时影响性能。

步骤5 选择“更多 > 重启服务”重启服务。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Hive”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.109 ALM-16006 Hive 服务进程直接内存使用超出阈值

告警解释

系统每30秒周期性检测Hive直接内存使用率，并把实际的Hive直接内存使用率和阈值相比较。当Hive直接内存使用率超出阈值（默认为最大直接内存的95%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Hive”修改阈值。

当Hive直接内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
16006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Hive直接内存使用率过高，会影响Hive任务运行的性能，甚至造成内存溢出导致Hive服务不可用。

可能原因

该节点Hive实例直接内存使用量过大，或分配的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“16006”的告警，查看“定位信息”中的角色名并确定实例的IP地址。

- 告警上报的角色是HiveServer，执行**步骤2**。
- 告警上报的角色是MetaStore，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的HiveServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“HiveServer内存使用率统计”，单击“确定”，查看HiveServer进程使用的直接内存是否已达到HiveServer进程设定的最大直接内存的阈值（默认95%）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的MetaStore，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“MetaStore内存使用率统计”，单击“确定”，查看MetaStore进程使用的直接内存是否已达到MetaStore进程设定的最大直接内存的阈值（默认95%）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置”，选择“全部配置”， 选择“HiveServer/MetaStore > JVM”，将“HIVE_GC_OPTS/METASTORE_GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值根据如下建议进行调整，并单击“保存”。

说明

1. HiveServer的GC参数配置建议
   - 建议将“-XX:MaxDirectMemorySize”值设置为“-Xmx”值的1/8，比如：当“-Xmx”设置为8G时，“-XX:MaxDirectMemorySize”设置为1024M，“-Xmx”设置为4G时，“-XX:MaxDirectMemorySize”设置为512M。并且建议“-XX:MaxDirectMemorySize”值不小于512M。
2. MetaServer的GC参数配置建议
   - 建议将“-XX:MaxDirectMemorySize”值设置为“-Xmx”值的1/8，比如：当“-Xmx”设置为8G时，“-XX:MaxDirectMemorySize”设置为1024M，“-Xmx”设置为4G时，“-XX:MaxDirectMemorySize”设置为512M。并且建议“-XX:MaxDirectMemorySize”值不小于512M。

步骤5 选择“更多 > 重启服务”重启服务。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Hive”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.110 ALM-16007 Hive GC 时间超出阈值

告警解释

系统每60秒周期性检测Hive服务的GC时间，当检测到Hive服务的GC时间超出阈值(连续3次检测超过12秒)时产生该告警。用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Hive”修改阈值。当Hive GC时间小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
16007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC时间超出阈值，会影响到Hive数据的读写。

可能原因

该节点Hive实例内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“16007”的告警，查看“定位信息”中的角色名并确定实例的IP地址。

- 告警上报的角色是HiveServer，执行**步骤2**。
- 告警上报的角色是MetaStore，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的HiveServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > GC”，勾选“HiveServer的GC时间”，单击“确定”，查看GC时间是否大于12秒。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的MetaStore，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > GC”，勾选“MetaStore的GC时间”，单击“确定”，查看GC时间是否大于12秒。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

查看JVM的当前配置

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置”，选择“全部配置”，选择“HiveServer/MetaStore > JVM”，将“HIVE_GC_OPTS/METASTORE_GC_OPTS”参数中“-Xmx”的值根据如下建议进行调整，并单击“保存”。

说明

1. HiveServer的GC参数配置建议
   - 当Hive GC时间超出阈值时，将“-Xmx”值调整为默认值的2倍，比如：“-Xmx”默认设置为2G时，调整“-Xmx”的值为4G。
   - 建议同时调节“-Xms”的值，使“-Xms”和“-Xmx”比值为1:2，这样可以避免JVM动态调整堆内存大小时影响性能。
2. MetaServer的GC参数配置建议
   - 当Meta GC时间超出阈值时，将“-Xmx”值调整为默认值的2倍，比如：“-Xmx”默认设置为2G时，调整“-Xmx”的值为4G。
   - 建议同时调节“-Xms”的值，使“-Xms”和“-Xmx”比值为1:2，这样可以避免JVM动态调整堆内存大小时影响性能。

步骤5 选择“更多 > 重启服务”重启服务。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在主备集群的FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Hive”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.111 ALM-16008 Hive 服务进程非堆内存使用超出阈值

告警解释

系统每30秒周期性检测Hive非堆内存使用率，并把实际的Hive非堆内存使用率和阈值相比较。当Hive非堆内存使用率超出阈值（默认为最大非堆内存的95%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Hive”修改阈值。

当Hive非堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
16008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Hive非堆内存使用率过高，会影响Hive任务运行的性能，甚至造成内存溢出导致Hive服务不可用。

可能原因

该节点Hive实例非堆内存使用量过大，或分配的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“16008”的告警，查看“定位信息”中的角色名并确定实例的IP地址。

- 告警上报的角色是HiveServer，执行**步骤2**。
- 告警上报的角色是MetaStore，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的HiveServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“HiveServer内存使用率统计”，单击“确定”，查看HiveServer进程使用的非堆内存是否已达到HiveServer进程设定的最大非堆内存的阈值（默认95%）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，单击告警上报的MetaStore，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存”，勾选“MetaStore内存使用率统计”，单击“确定”，查看MetaStore进程使用的非堆内存是否已达到MetaStore进程设定的最大非堆内存的阈值（默认95%）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置”，选择“全部配置”， 选择“HiveServer/MetaStore > JVM”，将“HIVE_GC_OPTS/METASTORE_GC_OPTS”参数中“-XX:MaxMetaspaceSize”的值根据如下建议进行调整，并单击“保存”。

说明

1. HiveServer的GC参数配置建议
   - 建议将“-XX:MaxMetaspaceSize”值设置成为“-Xmx”大小的1/8，比如：“-Xmx”设置为2G时，“-XX:MaxMetaspaceSize”设置为256M；“-Xmx”设置为4G时，“-XX:MaxMetaspaceSize”设置为512M。
2. MetaServer的GC参数配置建议
   - 建议将“-XX:MaxMetaspaceSize”值设置成为“-Xmx”大小的1/8，比如：“-Xmx”设置为2G时，“-XX:MaxMetaspaceSize”设置为256M；“-Xmx”设置为4G时，“-XX:MaxMetaspaceSize”设置为512M。

步骤5 选择“更多 > 重启服务”重启服务。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Hive”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.112 ALM-16009 Map 数超过阈值

告警解释

系统每30秒周期性检测执行的HQL的Map数是否超过阈值，超过阈值发出告警。系统默认的平滑次数为3次，默认阈值为5000。

告警属性

告警ID	告警级别	是否自动清除
16009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Hive执行的HQL的Map数过高，一方面会导致HQL执行较慢，另一方面会大量占用资源。

可能原因

执行的HQL语句存在可以优化的可能。

处理步骤

检查HQL的Map个数。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Hive > 资源”，查看“HQL的Map数”图表，找出Map数过大的HQL语句（Map数>=5000）。

步骤2 找到对应的HQL语句，优化在监控上显示map数过大的HQL语句，再尝试执行。

步骤3 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的“Hive”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.113 ALM-16045 Hive 数据仓库被删除

告警解释

系统每60秒周期性检测Hive数据仓库情况，Hive数据仓库被删除告警。

告警属性

告警ID	告警级别	是否自动清除
16045	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Hive默认数据仓库被删除，会导致在默认数据仓库中创建库、创建表失败，影响业务正常使用。

可能原因

Hive定时查看默认数据仓库的状态，发现Hive默认数据仓库被删除。

处理步骤

检查Hive默认数据仓库。

步骤1 以**root**用户登录客户端所在节点。

步骤2 执行以下命令，检查“hdfs://hacluster/user/*{用户名}*/.Trash/Current/ ”目录下是否存在该warehouse目录。

hdfs dfs -ls hdfs://hacluster/user/*<用户名>*/.Trash/Current/

例如存在“user/hive/warehouse”：

```
host01:/opt/Bigdata/client # hdfs dfs -ls hdfs://hacluster/user/test/.Trash/Current/
Found 1 items
drwx------   - test hadoop          0 2019-06-17 19:53 hdfs://hacluster/user/test/.Trash/Current/user
```

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 默认数据仓库存在自动恢复机制，用户可等待默认数据仓库的恢复（5 ~10s）。如果未恢复，用户可执行以下命令，将warehouse重新复原。

hdfs dfs -mv hdfs://hacluster/user/*<用户名>*/.Trash/Current/user/hive/warehouse /user/hive/warehouse

步骤4 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤5**。

收集故障信息。

步骤5 收集客户端后台“/.Trash/Current/”目录下内容的相关信息。

步骤6 请联系运维人员，并发送已收集的故障信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.114 ALM-16046 Hive 数据仓库权限被修改

告警解释

系统每60秒周期性检测Hive数据仓库的权限是否被修改，如果修改发出告警。

告警属性

告警ID	告警级别	是否自动清除
16046	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Hive默认数据仓库的权限被修改，会影响当前用户，用户组，其他用户在默认数据仓库中创建库、创建表等操作的操作权限范围。会扩大或缩小权限。

可能原因

Hive定时查看默认数据仓库的状态，发现Hive默认数据仓库权限发生更改。

处理步骤

检查Hive默认数据仓库权限情况。

步骤1 以**root**用户登录客户端所在节点。

步骤2 请使用具有supergroup组权限的用户，根据当前集群情况恢复目录权限：

- 安全环境：执行命令**hdfs dfs -chmod 770 hdfs://hacluster/user/hive/warehouse**修复默认数据仓库权限。
- 非安全环境：执行命令**hdfs dfs -chmod 777 hdfs://hacluster/user/hive/warehouse**修复默认数据仓库权限。

步骤3 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤4**。

收集故障信息。

步骤4 收集客户端后台“hdfs://hacluster/user/hive/warehouse”目录下内容的相关信息。

步骤5 请联系运维人员，并发送已收集的故障信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.115 ALM-16047 HiveServer 已从 Zookeeper 注销

告警解释

系统每60秒周期性检测Hive服务，若Hive在Zookeeper上的注册信息丢失，或者Hive无法连接上Zookeeper，将会发出告警。

告警属性

告警ID	告警级别	是否自动清除
16047	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

当无法在Zookeeper上读取到Hive的配置，将会导致hiveserver不可用。

可能原因

- 网络故障。
- ZooKeeper实例状态异常。

处理步骤

重启相关实例。

步骤1 登录FusionInsight Manager，在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击告警“Hive解注Zookeeper”所在行的下拉菜单，在“定位信息”中查看告警上报的角色名并确定实例IP地址。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，勾选上报告警IP对应的实例，选择“更多 > 重启实例”。

步骤3 重启完成后，等待5分钟，查看告警是否消除。

- 是，处理完毕。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager首页，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的“Hive”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.116 ALM-16048 Tez 或者 Spark 库路径不存在

告警解释

系统每180秒周期性检测Tez和Spark库路径，不存在则产生该告警。

告警属性

告警ID	告警级别	是否自动清除
16048	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Tez或者Spark库路径不存在，会影响Hive on Tez，Hive on Spark的功能。

可能原因

Tez或者Spark在HDFS上库路径被删除。

处理步骤

检查Tez和Spark库路径。

步骤1 以**root**用户登录客户端所在节点。

步骤2 执行以下命令，检查“hdfs://hacluster/user/*{用户名}*/.Trash/Current/ ”目录下是否存在该tezlib或者sparklib目录。

hdfs dfs -ls hdfs://hacluster/user/*<用户名>*/.Trash/Current/

例如存在“/user/hive/tezlib/8.1.0.1/”和“/user/hive/sparklib/8.1.0.1/”：

```
host01:/opt/Bigdata/client # hdfs dfs -ls hdfs://hacluster/user/test/.Trash/Current/
Found 1 items
drwx------   - test hadoop          0 2019-06-17 19:53 hdfs://hacluster/user/test/.Trash/Current/user
```

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 执行以下命令，将tezlib和sparklib重新复原。

hdfs dfs -mv hdfs://hacluster/user/*<用户名>*/.Trash/Current/user/hive/tezlib/8.1.0.1/tez.tar.gz /user/hive/tezlib/8.1.0.1/tez.tar.gz

步骤4 查看本告警是否恢复。

- 是，操作结束。
- 否，执行**步骤5**。

收集故障信息。

步骤5 收集客户端后台“/.Trash/Current/”目录下内容的相关信息。

步骤6 请联系运维人员，并发送已收集的故障信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.117 ALM-17003 Oozie 服务不可用

告警解释

系统每5秒周期性检测Oozie服务状态，当Oozie或者Oozie所依赖的组件无法正常提供服务时，系统产生此告警。

当Oozie服务恢复可用状态时，告警自动消除。

告警属性

告警ID	告警级别	是否自动清除
17003	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Details	对告警信息的补充。

对系统的影响

无法使用Oozie服务提交作业。

可能原因

- DBService服务异常或者Oozie存储在DBService中的数据遭到破坏，导致Oozie服务不可用。
- HDFS服务异常或者Oozie存储在HDFS中的数据遭到破坏时，导致Oozie服务不可用。
- Yarn服务异常，导致Oozie服务不可用。
- Nodeagent进程故障，导致Oozie服务不可用。

处理步骤

查询Oozie服务健康状态码。

步骤1 在FusionInsight Manager中，选择“集群 > *待操作集群的名称* > 服务 > Oozie”，单击“oozie WebUI”的“oozie”（两个任选一个），进入Oozie WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

步骤2 在浏览器地址栏的URL地址后追加“/servicehealth”重新访问，“statusCode”对应的值即为当前Oozie的服务健康状态码。

例如，在浏览器中访问“https://10.10.0.117:20026/Oozie/oozie/130/oozie/servicehealth”，显示结果为：

```
{"beans":[{"name":"serviceStatus","statusCode":0}]}
```

如果无法查询出健康状态码或者浏览器一直无响应，可能是由于Oozie进程故障导致服务不可用，请参考**步骤13**进行处理。

步骤3 根据查询到的错误码执行相关处理步骤，请参考**表10-90**。

表 10-90 Oozie 服务健康状态码一览表

状态码	错误描述	错误原因	处理步骤
0	服务正常	无	无
18002	DBService服务异常	Oozie连接DBservice失败或者存储在DBService中的数据遭到破坏	请参考**步骤4**。
18003	HDFS服务异常	Oozie连接HDFS失败或者存储在HDFS中的数据遭到破坏	请参考**步骤7**。
18005	Mapreduce服务异常	Yarn服务异常	请参考**步骤11**。

检查DBService服务。

步骤4 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”，检查DBService服务当前状态是否正常。

- 是，执行**步骤6**。
- 否，执行**步骤5**。

步骤5 参考DBService服务的相关告警帮助进行处理，然后查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤18**。

步骤6 登录Oozie数据库检查数据是否完整。

1. 以**root**用户登录DBService主节点。

   在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > DBService > 实例”，即可查看DBService主节点IP地址信息。

2. 执行以下命令登录Oozie数据库。

   su - omm

   source ${BIGDATA_HOME}/FusionInsight_BASE_8.1.0.1/install/FusionInsight-dbservice-2.7.0/.dbservice_profile

   gsql -U *用户名* **-W** *Oozie数据库密码* **-p 20051 -d** *数据库名称*

3. 登录成功后，输入**\d**，检查数据表是否共有15张。

   Oozie服务默认有15张数据表，如果这些数据表被删除或者表结构被修改都可能导致Oozie服务不可用，请联系运维人员备份相关数据后进行恢复。

检查HDFS服务。

步骤7 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”，检查HDFS服务当前状态是否正常。

- 是，执行**步骤9**。
- 否，执行**步骤8**。

步骤8 参考HDFS服务的相关告警帮助进行处理，然后查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤18**。

步骤9 登录HDFS检查Oozie文件目录是否完整。

1. 下载并安装HDFS客户端。
2. 以**root**用户登录客户端所在节点，执行以下命令，检查“/user/oozie/share”路径是否存在。

   如果集群采用安全版本，要进行安全认证。

   kinit admin

   hdfs dfs -ls /user/oozie/share

- 是，执行**步骤18**。
- 否，执行**步骤10**。

步骤10 在Oozie客户端安装目录中手动将share目录上传至HDFS的“/user/oozie”路径下，检查告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤18**。

检查Yarn/Mapreduce服务。

步骤11 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”，检查Yarn、Mapreduce服务当前状态是否正常。

- 是，执行**步骤18**。
- 否，执行**步骤12**。

步骤12 参考Yarn、Mapreduce服务的相关告警帮助进行处理，然后查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤18**。

检查Oozie进程。

步骤13 以**root**用户分别登录Oozie服务两个节点。

在FusionInsight Manager界面单击“集群 > *待操作集群的名称* > 服务 > Oozie > 实例”，即可查看服务所在节点的IP地址信息。

步骤14 执行命令**ps -ef | grep oozie**，检查Oozie进程是否存在。

- 是，执行**步骤15**。
- 否，执行**步骤18**。

步骤15 分别检查和收集Oozie日志目录“/var/log/Bigdata/oozie”中的prestartDetail.log、oozie.log、catalina.out里的异常信息，确认非人为误操作导致的问题后，执行**步骤16**。

检查Nodeagent进程。

步骤16 以**root**用户分别登录Oozie服务两个节点。执行命令**ps -ef | grep nodeagent**，检查Nodeagent进程是否存在。

- 是，执行**步骤17**。
- 否，执行**步骤18**。

步骤17 执行**kill -9** *查询到的nodeagent进程ID*命令，等待10分钟后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤18**。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.118 ALM-17004 Oozie 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测Oozie服务堆内存使用状态，当检测到Oozie实例堆内存使用率超出阈值（最大内存的95%）时产生该告警。堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
17004	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

堆内存溢出可能导致服务崩溃。

可能原因

该节点Oozie实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Oozie堆内存使用率超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Oozie > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > 内存”中的“Oozie堆内存使用率”，单击“确定”。

步骤3 查看Oozie使用的堆内存是否已达到Oozie设定的阈值（默认值为最大堆内存的95%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务> Oozie > 配置”，选择“全部配置”。在搜索栏里搜索“GC_OPTS”参数，将“-Xmx”的值根据实际情况调大，并单击“保存”，单击“确定”。

说明

Oozie的GC参数配置建议：

建议“-Xms”和“-Xmx”设置成相同的值，这样可以避免JVM动态调整堆内存大小时影响性能。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Oozie”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.119 ALM-17005 Oozie 非堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Oozie服务非堆内存使用状态，当检测到Oozie实例非堆内存使用率超出阈值（最大内存的80%）时产生该告警。非堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
17005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

非堆内存溢出可能导致服务崩溃。

可能原因

该节点Oozie实例非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Oozie非堆内存使用率超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Oozie > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > 内存”中的“Oozie非堆内存使用率”，单击“确定”。

步骤3 查看Oozie使用的非堆内存是否已达到Oozie设定的阈值（默认值为最大非堆内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务> Oozie > 配置”，选择“全部配置”，在搜索栏里搜索“GC_OPTS”参数，查看参数中是否有“-XX: MaxMetaspaceSize”。如果是，将“-XX: MaxMetaspaceSize”的值根据实际情况调大。如果否，手动添加“-XX: MaxMetaspaceSize”并将值设置成为“-Xmx”大小的1/8。单击“保存”，单击“确定”。

说明

JDK1.8不再支持MaxPermSize。

Oozie的GC参数配置建议：

建议将“-XX:MaxMetaspaceSize”值设置成为“-Xmx”大小的1/8，比如：“-Xmx”设置为2G时，“-XX:MaxMetaspaceSize”设置为256M；“-Xmx”设置为4G时，“-XX:MaxMetaspaceSize”设置为512M。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Oozie”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.120 ALM-17006 Oozie 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测Oozie服务直接内存使用状态，当检测到Oozie实例直接内存使用率超出阈值（最大内存的80%）时，产生该告警。当Oozie直接内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
17006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

直接内存溢出可能导致服务崩溃。

可能原因

该节点Oozie实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Oozie直接内存使用率超过阈值 > 定位信息”检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Oozie > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > 内存”中的“Oozie直接内存使用率”，单击“确定”。

步骤3 查看Oozie使用的直接内存是否已达到Oozie设定的阈值（默认值为最大直接内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务> Oozie > 配置”，选择“全部配置”，在搜索栏里搜索“GC_OPTS”参数。将“-XX:MaxDirectMemorySize”的值根据实际情况调大，并单击“保存”，单击“确定”。

说明

Oozie的GC参数配置建议：

建议将“-XX:MaxDirectMemorySize”值设置为“-Xmx”值的1/4，比如：当“-Xmx”设置为4G时，“-XX:MaxDirectMemorySize”设置为1024M，“-Xmx”设置为2G时，“-XX:MaxDirectMemorySize”设置为512M。并且建议“-XX:MaxDirectMemorySize”值不小于512M。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Oozie”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.121 ALM-17007 Oozie 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测Oozie进程的垃圾回收（GC）占用时间，当检测到Oozie进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
17007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

导致Oozie提交任务响应变慢。

可能原因

该节点Oozie实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Oozie进程垃圾回收（GC）时间超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Oozie > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > GC”中的“Oozie垃圾回收（GC）总时间”，单击“确定”。

步骤3 查看Oozie每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务> Oozie > 配置”，选择“全部配置”，在搜索栏里搜索“GC_OPTS”参数。将“-Xmx”的值根据实际情况调大，并单击“保存”，单击“确定”进行保存。

说明

Oozie的GC参数配置建议：

建议“-Xms”和“-Xmx”设置成相同的值，这样可以避免JVM动态调整堆内存大小时影响性能。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Oozie”，单击“确定”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.122 ALM-18000 Yarn 服务不可用

告警解释

告警模块按60秒周期检测Yarn服务状态。当检测到Yarn服务不可用时产生该告警。

Yarn服务恢复时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

集群无法提供Yarn服务。用户无法执行新的application。已提交的application无法执行。

可能原因

- ZooKeeper服务异常。
- HDFS服务异常。
- Yarn集群中没有主ResourceManager实例。
- Yarn集群中的所有NodeManager节点异常。

处理步骤

检查ZooKeeper服务状态。

步骤1 在FusionInsight Manager的告警列表中，查看是否有告警“ALM-13000 ZooKeeper服务不可用”产生。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 参考“ALM-13000 ZooKeeper服务不可用”的处理步骤处理故障后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤3**。

检查HDFS服务状态。

步骤3 在FusionInsight Manager的告警列表中，查看是否有HDFS相关告警产生。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 选择“运维 > 告警 > 告警”，根据告警帮助处理HDFS相关告警后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

检查Yarn集群中的ResourceManager状态。

步骤5 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* >服务 > Yarn”。

步骤6 在“概览”中，检查Yarn集群中是否存在主ResourceManager实例。

- 是，执行**步骤7**。
- 否，执行**步骤10**。

检查Yarn集群中的NodeManager节点状态。

步骤7 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”。

步骤8 查看NodeManager的“运行状态”，检查是否有处于非健康状态的节点。

- 是，执行**步骤9**。
- 否，执行**步骤10**。

步骤9 按“ALM-18002 NodeManager心跳丢失”或“ALM-18003 NodeManager不健康”提供的步骤处理该故障，故障修复后检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤10**。

收集故障信息。

步骤10 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”勾选待操作集群的“Yarn”。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.123 ALM-18002 NodeManager 心跳丢失

告警解释

系统每30秒周期性检测丢失的NodeManager节点，并把丢失的节点数和阈值相比较。“丢失的节点数”指标默认提供一个阈值。当检测到“丢失的节点数”的值超出阈值时产生该告警。

用户可通过选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置”，修改yarn.nodemanager.lost.alarm.threshold的值来配置阈值（修改该参数不用重启Yarn，就可以生效）。

阈值默认为零，当丢失节点数超过该值时，触发告警，小于阈值时会自动消除告警。

告警属性

告警ID	告警级别	是否自动清除
18002	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Lost Host	丢失节点的主机列表。

对系统的影响

- 丢失的NodeManager节点无法提供Yarn服务。
- 容器减少，集群性能下降。

可能原因

- NodeManager没有经过退服操作，强制被删除。
- NodeManager所有实例被停止或者进程故障。
- NodeManager节点所在主机故障。
- NodeManager和ResourceManager之间的网络断连或者繁忙。

处理步骤

检查NodeManager状态。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”，在告警列表中找到当前告警，单击 获取告警详细信息，在“附加信息”中获取丢失状态的节点。

步骤2 确认处于丢失状态的节点是否是人为未经过退服操作，直接主动删除的主机。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 选择“集群 > *待操作集群的名称* > 服务 > Yarn”，进入“配置”页面，选择“全部配置”，搜索“yarn.nodemanager.lost.alarm.threshold”，修改值为未退服主动删除的主机个数。设置成功后检查告警是否清除。

- 是，处理完毕。
- 否，执行**步骤4**。

步骤4 手动清除此告警，后续删除主机前务必进行退服操作。

步骤5 在FusionInsight Manager界面，选择“集群 > 主机”，查看**步骤1**中获取的节点是否健康。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

步骤6 参考“ALM-12006 节点故障”的操作步骤进行处理，节点恢复正常后，查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

检查进程状态。

步骤7 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”，查看是否存在状态为非“良好”的NodeManager。

- 是，执行**步骤10**。
- 否，执行**步骤8**。

步骤8 确认此NodeManager实例是否被删除。

- 是，执行**步骤9**。
- 否，执行**步骤11**。

步骤9 重启ResourceManager的主备实例，然后检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

检查实例状态。

步骤10 选择处于非“良好”状态的NodeManager实例并重启该实例。检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

检查网络状态。

步骤11 登录管理节点，**ping**丢失的NodeManager节点的IP地址，检查网络是否断连或繁忙。

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 修复网络故障，然后查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

收集故障信息。

步骤13 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选待操作集群的“Yarn”。

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.124 ALM-18003 NodeManager 不健康

告警解释

系统每30秒周期性检测不健康NodeManager节点，并把不健康节点数和阈值相比较。“不健康的节点数”指标默认提供一个阈值。当检测到“不健康的节点数”的值超出阈值时产生该告警。

用户可通过选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置”，修改

“yarn.nodemanager.unhealthy.alarm.threshold”的值来配置阈值（修改该参数不用重启Yarn，就可以生效）。

阈值默认为零，当不健康节点数超过该值时，触发告警，小于阈值时会自动消除告警。

告警属性

告警ID	告警级别	是否自动清除
18003	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Unhealthy Host	不健康节点的主机列表。

对系统的影响

- 故障的NodeManager节点无法提供Yarn服务。
- 容器减少，集群性能下降。

可能原因

- NodeManager节点所在主机的硬盘空间不足。
- NodeManager节点本地目录**omm**用户无访问权限。

处理步骤

检查主机的硬盘空间。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”，在告警列表中找到当前告警，单击 获取告警详细信息，在“附加信息”中获取不健康状态的节点。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”，选择对应主机的NodeManager实例，选择“实例配置 > 全部配置”，搜索“yarn.nodemanager.local-dirs”和“yarn.nodemanager.log-dirs”对应的磁盘。

步骤3 选择“运维 > 告警 > 告警”，在告警列表中查看对应的磁盘是否存在“ALM-12017 磁盘容量不足”告警。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 参考“ALM-12017 磁盘容量不足”操作步骤进行处理，故障恢复后，查看本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

步骤5 选择“主机 > *待查看的主机名称*”，在主机的概览页面查看对应分区的磁盘使用情况。检查挂载磁盘使用空间百分比是否已经超过Yarn参数“yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage”所配置的值。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤6 将磁盘使用率降到该配置值以下，等待10-20分钟，然后检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

检查NodeManager节点本地目录的访问权限。

步骤7 获取**步骤2**中查看到的NodeManager目录，以**root**用户登录每个NodeManager节点，并进入获取到的目录。

步骤8 执行**ll**命令查看对应localdir的文件夹和containerlogs文件夹权限，确认权限是否是“755”，且“用户:属组”是否为“omm:ficommon”。

- 是，处理完毕。
- 否，执行**步骤9**。

步骤9 执行如下命令将文件夹权限修改为“755”，并将“用户:属组”修改为“omm:ficommon”。

chmod 755 *<folder_name>*

chown omm:ficommon *<folder_name>*

步骤10 等待10～20分钟，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的“Yarn”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.125 ALM-18008 ResourceManager 堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Yarn ResourceManager堆内存使用率，并把实际的Yarn ResourceManager堆内存使用率和阈值相比较。当Yarn ResourceManager堆内存使用率超出阈值（默认为最大堆内存的95%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn”修改阈值。

平滑次数为1，Yarn ResourceManager堆内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，Yarn ResourceManager堆内存使用率小于或等于阈值的95%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Yarn ResourceManager堆内存使用率过高，会影响Yarn任务提交和运行的性能，甚至造成内存溢出导致Yarn服务不可用。

可能原因

该节点Yarn ResourceManager实例堆内存使用量过大，或分配的堆内存不合理，导致使用量超过阈值。

处理步骤

检查堆内存使用量。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警 > Yarn ResourceManager堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > ResourceManager（对应上报告警实例主机名）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“ResourceManager内存使用率”。查看堆内存使用情况。

步骤3 查看ResourceManager使用的堆内存是否已达到ResourceManager设定的最大堆内存的95%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > ResourceManager > 系统”。将“GC_OPTS”参数根据实际情况调大，并单击“保存”，保存完成后重启角色实例。

说明

集群中的NodeManager实例数量和ResourceManager内存大小的对应关系参考如下：

- 集群中的NodeManager实例数据达到100，ResourceManager实例的JVM参数建议配置为：-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到200，ResourceManager实例的JVM参数建议配置为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到500，ResourceManager实例的JVM参数建议配置为：-Xms10G -Xmx10G -XX:NewSize=1G -XX:MaxNewSize=2G
- 集群中的NodeManager实例数据达到1000，ResourceManager实例的JVM参数建议配置为：-Xms20G -Xmx20G -XX:NewSize=1G -XX:MaxNewSize=2G
- 集群中的NodeManager实例数据达到2000，ResourceManager实例的JVM参数建议配置为：-Xms40G -Xmx40G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到3000，ResourceManager实例的JVM参数建议配置为：-Xms60G -Xmx60G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到4000，ResourceManager实例的JVM参数建议配置为：-Xms80G -Xmx80G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到5000，ResourceManager实例的JVM参数建议配置为：-Xms100G -Xmx100G -XX:NewSize=3G -XX:MaxNewSize=6G

步骤5 观察界面告警是否清除。

- 是，处理完毕。

- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- NodeAgent
- Yarn

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.126 ALM-18009 JobHistoryServer 堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Mapreduce JobHistoryServer堆内存使用率，并把实际的Mapreduce JobHistoryServer堆内存使用率和阈值相比较。当Mapreduce JobHistoryServer堆内存使用率超出阈值（默认为最大堆内存的95%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Mapreduce”修改阈值。

平滑次数为1，MapReduce JobHistoryServer堆内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，MapReduce JobHistoryServer堆内存使用率小于或等于阈值的95%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Mapreduce JobHistoryServer堆内存使用率过高，会影响Mapreduce 服务日志归档的性能，甚至造成内存溢出导致Mapreduce服务不可用。

可能原因

该节点Mapreduce JobHistoryServer实例堆内存使用量过大，或分配的堆内存不合理，导致使用量超过阈值。

处理步骤

检查内存使用量。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > MapReduce JobHistoryServer堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Mapreduce > 实例 > JobHistoryServer（对应上报告警实例主机名）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“JobHistoryServer堆内存使用百分比统计”。查看堆内存使用情况。

步骤3 查看JobHistoryServer使用的堆内存是否已达到JobHistoryServer设定的最大堆内存的95%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Mapreduce > 配置 > 全部配置 > JobHistoryServer > 系统”。将“GC_OPTS”参数根据实际情况调大，并单击“保存”，单击“确定”并进重启。

说明

历史任务数10000和JobHistoryServer内存的对应关系如下：

-Xms30G -Xmx30G -XX:NewSize=1G -XX:MaxNewSize=2G

步骤5 观察界面告警是否清除?

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- NodeAgent
- Mapreduce

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.127 ALM-18010 ResourceManager 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测ResourceManager进程的垃圾回收（GC）占用时间，当检测到ResourceManager进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18010	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

ResourceManager进程的垃圾回收时间过长，可能影响该ResourceManager进程正常提供服务。

可能原因

该节点ResourceManager实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18010 ResourceManager进程垃圾回收（GC）时间超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > ResourceManager（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 垃圾回收”，勾选“ResourceManager垃圾回收（GC）时间”。查看ResourceManager每分钟的垃圾回收时间统计情况。

步骤3 查看ResourceManager每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > ResourceManager > 系统”。将“GC_OPTS”参数根据实际情况调大。

说明

集群中的NodeManager实例数量和ResourceManager内存大小的对应关系参考如下：

- 集群中的NodeManager实例数据达到100，ResourceManager实例的JVM参数建议配置为：-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到200，ResourceManager实例的JVM参数建议配置为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到500，ResourceManager实例的JVM参数建议配置为：-Xms10G -Xmx10G -XX:NewSize=1G -XX:MaxNewSize=2G
- 集群中的NodeManager实例数据达到1000，ResourceManager实例的JVM参数建议配置为：-Xms20G -Xmx20G -XX:NewSize=1G -XX:MaxNewSize=2G
- 集群中的NodeManager实例数据达到2000，ResourceManager实例的JVM参数建议配置为：-Xms40G -Xmx40G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到3000，ResourceManager实例的JVM参数建议配置为：-Xms60G -Xmx60G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到4000，ResourceManager实例的JVM参数建议配置为：-Xms80G -Xmx80G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到5000，ResourceManager实例的JVM参数建议配置为：-Xms100G -Xmx100G -XX:NewSize=3G -XX:MaxNewSize=6G

步骤5 保存配置，并重启该ResourceManager实例。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“ResourceManager”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.128 ALM-18011 NodeManager 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测NodeManager进程的垃圾回收（GC）占用时间，当检测到NodeManager进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18011	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。

参数名称	参数含义
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NodeManager进程的垃圾回收时间过长，可能影响该NodeManager进程正常提供服务。

可能原因

该NodeManager节点实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18011 NodeManager进程垃圾回收（GC）时间超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > NodeManager（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 垃圾回收”，勾选“NodeManager垃圾回收（GC）时间”。查看NodeManager每分钟的垃圾回收时间统计情况。

步骤3 查看NodeManager每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > NodeManager > 系统”。将“GC_OPTS”参数根据实际情况调大。

说明

集群中的NodeManager实例数量和NodeManager内存大小的对应关系参考如下：

- 集群中的NodeManager实例数据达到100，NodeManager实例的JVM参数建议配置为：-Xms2G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到200，NodeManager实例的JVM参数建议配置为：-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到500以上，NodeManager实例的JVM参数建议配置为：-Xms8G -Xmx8G -XX:NewSize=1G -XX:MaxNewSize=2G

步骤5 保存配置，并重启NodeManager实例。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“NodeManager”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.129 ALM-18012 JobHistoryServer 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测JobHistoryServer进程的垃圾回收（GC）占用时间，当检测到JobHistoryServer进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18012	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JobHistoryServer进程的垃圾回收时间过长，可能影响该JobHistoryServer进程正常提供服务。

可能原因

该节点JobHistoryServer实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18012 JobHistoryServer进程垃圾回收（GC）时间超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > MapReduce > 实例 > JobHistoryServer（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 垃圾回收”，勾选“JobHistoryServer垃圾回收（GC）时间”。查看JobHistoryServer每分钟的垃圾回收时间统计情况。

步骤3 查看JobHistoryServer每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Mapreduce > 配置 > 全部配置 > JobHistoryServer > 系统”。将“GC_OPTS”参数根据实际情况调大。

说明

历史任务数10000和JobHistoryServer内存的对应关系如下：

-Xms30G -Xmx30G -XX:NewSize=1G -XX:MaxNewSize=2G

步骤5 保存配置，并重启JobHistoryServer实例。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“JobHistoryServer”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.130 ALM-18013 ResourceManager 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测ResourceManager服务直接内存使用状态，当检测到ResourceManager实例直接内存使用率超出阈值（最大内存的90%）时，产生该告警。

直接内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18013	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

ResourceManager可用直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点ResourceManager实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18013 ResourceManager直接内存使用率超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > ResourceManager（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“ResourceManager内存使用详情”。查看直接内存使用情况。

步骤3 查看ResourceManager使用的直接内存是否已达到ResourceManager设定的最大直接内存的90%(默认阈值)。

- 是，执行**步骤4**。
- 否，执行**步骤9**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > ResourceManager > 系统”。查看“GC_OPTS”参数中是否存在“-XX:MaxDirectMemorySize”。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 在“GC_OPTS”中把参数“-XX:MaxDirectMemorySize”删除。

步骤6 保存配置，并重启ResourceManager实例。

步骤7 查看告警信息，是否存在告警“ALM-18008 ResourceManager堆内存使用率超过阈值”。

- 是，查看“ALM-18008 ResourceManager堆内存使用率超过阈值”进行处理。
- 否，执行**步骤8**。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“ResourceManager”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.131 ALM-18014 NodeManager 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测Yarn服务直接内存使用状态，当检测到NodeManager实例直接内存使用率超出阈值（最大内存的90%）时，产生该告警。

直接内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18014	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NodeManager可用直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点NodeManager实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18014 NodeManager直接内存使用率超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > NodeManager（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“NodeManager内存使用率”。查看直接内存使用情况。

步骤3 查看NodeManager使用的直接内存是否已达到NodeManager设定的最大直接内存的90%(默认阈值)。

- 是，执行**步骤4**。
- 否，执行**步骤9**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > NodeManager > 系统”。查看“GC_OPTS”参数中是否存在“-XX:MaxDirectMemorySize”。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 在“GC_OPTS”中把参数“-XX:MaxDirectMemorySize”删除。

步骤6 保存配置，并重启NodeManager实例。

步骤7 查看告警信息，是否存在告警“ALM-18018 NodeManager堆内存使用率超过阈值”。

- 是，查看“ALM-18018 NodeManager堆内存使用率超过阈值”进行处理。
- 否，执行**步骤8**。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“NodeManager”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.132 ALM-18015 JobHistoryServer 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测MapReduce服务直接内存使用状态，当检测到JobHistoryServer实例直接内存使用率超出阈值（最大内存的90%，默认阈值）时，产生该告警。

直接内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18015	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

MapReduce可用直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点JobHistoryServer实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18015 JobHistory直接内存使用率超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > MapReduce > 实例 > JobHistoryServer（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“JobHistoryServer内存使用详情”。查看直接内存使用情况。

步骤3 查看MapReduce使用的直接内存是否已达到MapReduce设定的最大直接内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤9**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > MapReduce > 配置 > 全部配置 > JobHistoryServer > 系统”。查看“GC_OPTS”参数中是否存在“-XX:MaxDirectMemorySize”。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 在“GC_OPTS”中把参数“-XX:MaxDirectMemorySize”删除。

步骤6 保存配置，并重启JobHistoryServer实例。

步骤7 查看告警信息，是否存在告警“ALM-18009 JobHistoryServer堆内存使用率超过阈值”。

- 是，查看“ALM-18009 JobHistoryServer堆内存使用率超过阈值”进行处理。
- 否，执行**步骤8**。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“JobHistoryServer”，单击“确定”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.133 ALM-18016 ResourceManager 非堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Yarn ResourceManager非堆内存使用率，并把实际的Yarn ResourceManager非堆内存使用率和阈值相比较。当Yarn ResourceManager非堆内存使用率超出阈值（默认为最大非堆内存的90%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn”修改阈值。

当Yarn ResourceManager非堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18016	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Yarn ResourceManager非堆内存使用率过高，会影响Yarn任务提交和运行的性能，甚至造成内存溢出导致Yarn服务不可用。

可能原因

该节点Yarn ResourceManager实例非堆内存使用量过大，或分配的非堆内存不合理，导致使用量超过阈值。

处理步骤

检查非堆内存使用量。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警 > ALM-18016 Yarn ResourceManager非堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > ResourceManager（对应上报告警实例主机名）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“ResourceManager内存使用率”。查看非堆内存使用情况。

步骤3 查看ResourceManager使用的非堆内存是否已达到ResourceManager设定的最大非堆内存的90%。

- 是，执行**步骤4**。

- 否，执行**步骤6**。

步骤4 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > ResourceManager > 系统”。对ResourceManager 的内存参数“GC_OPTS”进行调整。保存配置，并重启ResourceManager实例。

说明

集群中的NodeManager实例数量和ResourceManager内存大小的对应关系参考如下：

- 集群中的NodeManager实例数据达到100，ResourceManager实例的JVM参数建议配置为：-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到200，ResourceManager实例的JVM参数建议配置为：-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到500，ResourceManager实例的JVM参数建议配置为：-Xms10G -Xmx10G -XX:NewSize=1G -XX:MaxNewSize=2G
- 集群中的NodeManager实例数据达到1000，ResourceManager实例的JVM参数建议配置为：-Xms20G -Xmx20G -XX:NewSize=1G -XX:MaxNewSize=2G
- 集群中的NodeManager实例数据达到2000，ResourceManager实例的JVM参数建议配置为：-Xms40G -Xmx40G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到3000，ResourceManager实例的JVM参数建议配置为：-Xms60G -Xmx60G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到4000，ResourceManager实例的JVM参数建议配置为：-Xms80G -Xmx80G -XX:NewSize=2G -XX:MaxNewSize=4G
- 集群中的NodeManager实例数据达到5000，ResourceManager实例的JVM参数建议配置为：-Xms100G -Xmx100G -XX:NewSize=3G -XX:MaxNewSize=6G

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- NodeAgent
- Yarn

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.134 ALM-18017 NodeManager 非堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Yarn NodeManager非堆内存使用率，并把实际的Yarn NodeManager非堆内存使用率和阈值相比较。当Yarn NodeManager非堆内存使用率超出阈值（默认为最大非堆内存的90%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn”修改阈值。

当Yarn NodeManager非堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18017	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Yarn NodeManager非堆内存使用率过高，会影响Yarn任务提交和运行的性能，甚至造成内存溢出导致Yarn服务不可用。

可能原因

该节点Yarn NodeManager实例非堆内存使用量过大，或分配的非堆内存不合理，导致使用量超过阈值。

处理步骤

检查非堆内存使用量。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警 > ALM-18017 Yarn NodeManager非堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > NodeManager（对应上报告警实例主机名）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“NodeManager内存使用率”。查看非堆内存使用情况。

步骤3 查看NodeManager使用的非堆内存是否已达到NodeManager设定的最大非堆内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > NodeManager > 系统”。对NodeManager的内存参数“GC_OPTS”进行调整，并单击“保存”，在弹出的对话框中单击“确定”并重启角色实例。

说明

集群中的NodeManager实例数量和NodeManager内存大小的对应关系参考如下：

- 集群中的NodeManager实例数据达到100，NodeManager实例的JVM参数建议配置为：-Xms2G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到200，NodeManager实例的JVM参数建议配置为：-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到500以上，NodeManager实例的JVM参数建议配置为：-Xms8G -Xmx8G -XX:NewSize=1G -XX:MaxNewSize=2G

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”下拉框中勾选待操作集群的如下节点信息，单击“确定”。

- NodeAgent
- Yarn

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.135 ALM-18018 NodeManager 堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Yarn服务堆内存使用状态，当检测到NodeManager实例堆内存使用率超出阈值（最大内存的95%）时产生该告警。

堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18018	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

NodeManager堆内存使用率过高，会影响Yarn任务提交和运行的性能，甚至可能会造成内存溢出导致Yarn服务崩溃。

可能原因

该节点NodeManager实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-18018 NodeManager堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例 > NodeManager（对应上报告警实例IP地址）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“NodeManager内存使用率”。查看堆内存使用情况。

步骤3 查看NodeManager使用的堆内存是否已达到NodeManager设定的最大堆内存的95%(默认阈值)。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置 > NodeManager > 系统”。将“GC_OPTS”参数的值根据实际情况调大。保存配置，并重启NodeManager实例。

说明

集群中的NodeManager实例数量和NodeManager内存大小的对应关系参考如下：

- 集群中的NodeManager实例数据达到100，NodeManager实例的JVM参数建议配置为：-Xms2G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到200，NodeManager实例的JVM参数建议配置为：-Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=1G
- 集群中的NodeManager实例数据达到500以上，NodeManager实例的JVM参数建议配置为：-Xms8G -Xmx8G -XX:NewSize=1G -XX:MaxNewSize=2G

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- NodeAgent
- Yarn

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.136 ALM-18019 JobHistoryServer 非堆内存使用率超过阈值

告警解释

系统每30秒周期性检测MapReduce JobHistoryServer非堆内存使用率，并把实际的MapReduce JobHistoryServer非堆内存使用率和阈值相比较。当MapReduce JobHistoryServer非堆内存使用率超出阈值（默认为最大非堆内存的90%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > MapReduce”修改阈值。

当MapReduce JobHistoryServer非堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18019	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

MapReduce JobHistoryServer非堆内存使用率过高，会影响MapReduce任务提交和运行的性能，甚至造成内存溢出导致MapReduce服务不可用。

可能原因

该节点MapReduce JobHistoryServer实例非堆内存使用量过大，或分配的非堆内存不合理，导致使用量超过阈值。

处理步骤

检查非堆内存使用量。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警 > ALM-18019 MapReduce JobHistoryServer非堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > MapReduce > 实例 > JobHistoryServer（对应上报告警实例主机名）”，单击图表区域右上角的下拉菜单，选择“定制 > 资源”，勾选“JobHistoryServer非堆内存使用百分比统计”。查看非堆内存使用情况。

步骤3 查看JobHistoryServer使用的非堆内存是否已达到JobHistoryServer设定的最大非堆内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > MapReduce > 配置 > 全部配置 > JobHistoryServer > 系统”。对NodeManager 的内存参数“GC_OPTS”进行调整，并单击“保存”，单击“确定”进行重启。

说明

历史任务数10000和JobHistoryServer内存的对应关系如下：

-Xms30G -Xmx30G -XX:NewSize=1G -XX:MaxNewSize=2G

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- NodeAgent
- MapReduce

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.137 ALM-18020 Yarn 任务执行超时

告警解释

系统每15分钟周期性检测提交到Yarn上的Mapreduce和Spark应用任务（JDBC常驻任务除外），当检测到任务执行时间超过用户指定的超时时间时，产生该告警，但任务仍继续正常执行。其中，Mapreduce的客户端超时参数为“mapreduce.application.timeout.alarm”，Spark的客户端超时参数为“spark.application.timeout.alarm”（单位：毫秒）。

当该任务结束或者任务被终止后，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
18020	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
应用名	产生告警的应用名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

任务执行超时后的运行时间内，该告警一直存在，但任务仍继续正常执行，没有任何影响。

可能原因

- 指定的超时时间少于所需执行时间。
- 任务运行的队列资源不足。
- 任务数据倾斜，导致一些任务处理的数据量大，执行时间长。

处理步骤

检查超时时间是否正确设置。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，弹出告警页面。

步骤2 在告警页面，选中“告警ID”为“18020”的告警，在该页面的告警详情里查看“定位信息”，查看超时任务的名称和超时时间。

步骤3 根据任务名称和超时时间，选择“集群 > *待操作集群的名称* > 服务 > Yarn > ResourceManager(主)”，登录Yarn的原生页面。在原生页面找到该任务，查看该任务的“StartTime”，根据系统当前时间计算任务已执行的时间。查看已执行的时间是否大于超时时间。

- 是，执行**步骤5**。
- 否，执行**步骤10**。

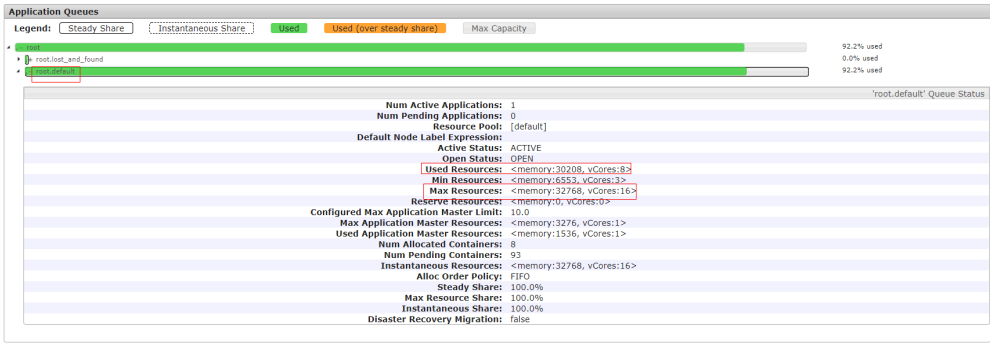
步骤4 请根据业务合理评估任务的预期执行时间，并与任务的超时时间对比。若超时时间设置过小，请设置客户端的超时时间（“mapreduce.application.timeout.alarm”或

“spark.application.timeout.alarm”）为任务的预期执行时间。重新运行任务后，查看是否不再上报告警。

- 是，处理完毕。
- 否，执行**步骤5**。

检查队列资源是否不足。

步骤5 在原生页面找到该任务，查看该任务的“Queue”中的队列名。单击原生页面左侧“Scheduler”,在“Applications Queues”页框中查找对应的队列名，并下拉展开队列的详细信息，如图所示：

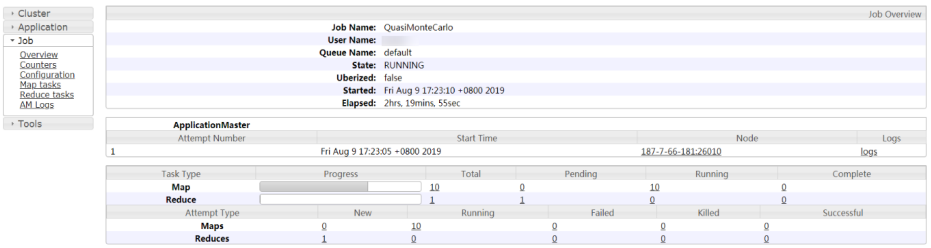


步骤6 查看队列详情中“Used Resources”是否近似等于“Max Resources”，即任务提交的队列中资源已经使用完毕，若队列资源不足，请在FusionInsight Manager的“租户资源 > 动态资源计划 > 资源分布策略”中调大队列的“最大资源”。重新运行任务后，查看是否不再上报告警。

- 是，处理完毕。
- 否，执行**步骤7**。

检查任务是否发生数据倾斜。

步骤7 在Yarn的原生页面，选择“*任务ID*（如application_1565337919723_0002）> Tracking URL:ApplicationMaster > job_1565337919723_0002”，进入如下页面：



步骤8 选择左侧“Job > Map tasks”或者“Job > Reduce tasks”，查看每个Map或者每个Reduce任务的执行时间是否相差很大，如果相差很大，说明任务数据发生了倾斜，需要对任务数据进行均衡。

步骤9 按照如上原因进行处理后，重新执行任务，观察本告警是否还出现。

- 是，执行**步骤10**。
- 否，处理完毕。

收集故障信息

步骤10 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选待操作集群的“Yarn”。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.138 ALM-18021 Mapreduce 服务不可用

告警解释

告警模块按60秒周期检测Mapreduce服务状态。当检测到Mapreduce服务不可用时产生该告警。

Mapreduce服务恢复时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
18021	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

集群无法提供Mapreduce服务，如无法通过Mapreduce查看任务日志，无法提供Mapreduce服务的日志归档功能等。

可能原因

- JobHistoryServer实例异常。
- KrbServer服务异常。
- ZooKeeper服务异常。
- HDFS服务异常。
- Yarn服务异常。

处理步骤

检查Mapreduce服务JobHistoryServer实例状态。

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Mapreduce > 实例”。

步骤2 查看**JobHistoryServer**的“运行状态”，检查**JobHistoryServer**是否处于良好状态。

- 是，执行**步骤11**。
- 否，执行**步骤3**。

检查KrbServer服务状态。

步骤3 在FusionInsight Manager的告警列表中，查看是否有“ALM-25500 KrbServer服务不可用”告警产生。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 参考“ALM-25500 KrbServer服务不可用”的处理步骤处理故障后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

检查Zookeeper服务状态。

步骤5 在FusionInsight Manager的告警列表中，查看是否有“ALM-13000 ZooKeeper服务不可用”告警产生。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤6 参考“ALM-13000 ZooKeeper服务不可用”的处理步骤处理故障后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

检查HDFS服务状态。

步骤7 在FusionInsight Manager的告警列表中，查看是否有“ALM-14000 HDFS服务不可用”告警产生。

- 是，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 参考“ALM-14000 HDFS服务不可用”的处理步骤处理故障后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

检查Yarn服务状态。

步骤9 在FusionInsight Manager的告警列表中，查看是否有“ALM-18000 Yarn服务不可用”告警产生。

- 是，执行**步骤10**。
- 否，执行**步骤11**。

步骤10 参考“ALM-18000 Yarn服务不可用”的处理步骤处理故障后，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤12 在“服务”中勾选待操作集群的“Mapreduce”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.139 ALM-18022 Yarn 队列资源不足

告警解释

告警模块按60秒周期检测Yarn队列资源，当队列可用资源或队列AM（ApplicationMaster）可用资源不足时，产生该告警。

当可用资源充足时，该告警自动消除。

告警属性

告警ID	告警级别	是否自动清除
18022	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
队列名	产生告警的队列名。
队列指标名	产生告警的队列指标名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

- 应用任务结束时间变长。
- 新应用提交后长时间无法运行。

可能原因

- NodeManager节点资源过小。
- 队列最大资源容量设置过小。
- AM最大资源百分比设置过小。

处理步骤

检查告警详情。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”，弹出告警页面。

步骤2 查看“Yarn队列资源不足”告警详情中的“定位信息”，查看“定位信息”是否为“队列名=root;队列指标名=Memory”或“队列名=root;队列指标名=vCores”。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 出现该定位信息表示Yarn集群内存或CPU不足，登录NodeManager节点，分别使用命令**free -g**和**cat /proc/cpuinfo**，查询节点可用内存和可用CPU，据此在FusionInsight Manager界面增大Yarn NodeManager的资源参数“yarn.nodemanager.resource.memory-mb”和“yarn.nodemanager.resource.cpu-vcores”的值，然后重启NodeManager实例。查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤4**。

步骤4 查看“定位信息”为“队列名=<租户队列名>;队列指标名=Memory”或“队列名=<租户队列名>;队列指标名=vCores”，然后查看“附加信息”是否包含“available Memory =”或“available vCores =”。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 出现该附加信息表示该租户队列内存或者CPU不足，选择“租户资源 > 动态资源计划 > 资源分布策略”，调大“最大资源容量”的值，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤6**。

步骤6 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置 > 全部配置”，输入搜索关键字“threshold”，单击“ResourceManager”，调整如下参数阈值：

如果“附加信息”中包含“available Memory =”，调整“yarn.queue.memory.alarm.threshold”的阈值使其小于“附加信息”中的“available Memory =”的值。

如果“附加信息”中包含“available vCores =”，调整“yarn.queue.vcore.alarm.threshold”的阈值使其小于“附加信息”中的“available vCores =”的值。

等待5分钟，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤9**。

步骤7 查看“附加信息”包含“available AmMemory =”或“available AmvCores =”，表示该租户队列的ApplicationMaster内存和CPU不足，选择“租户资源 > 动态资源计划 > 队列配置”，增大“AM最大资源百分比”，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤8**。

步骤8 选择“集群 > *待操作集群的名称* >服务 > Yarn > 配置 > 全部配置”，输入搜索关键字“threshold”，单击“ResourceManager”：调整如下参数阈值：

如果“附加信息”包含“available AmMemory =”，调整“yarn.am.memory.alarm.threshold”的阈值使其小于“附加信息”中的“available AmMemory =”的值。

如果“附加信息”包含“available AmvCores =”，调整“yarn.am.vcore.alarm.threshold”的阈值使其小于“附加信息”中的“available AmvCores =”的值。

等待5分钟，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在主集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“Yarn”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.140 ALM-18023 Yarn 任务挂起数超过阈值

告警解释

告警模块按60秒周期检测Yarn队列上pending的应用的数量，当root队列上处于pending状态的应用的数量超过60时，触发该告警。

告警属性

告警ID	告警级别	是否自动清除
18023	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
队列名	产生告警的队列名。
队列指标名	产生告警的队列指标名。

对系统的影响

- 应用任务结束时间变长。
- 新应用提交后长时间无法运行。

可能原因

- NodeManager节点资源过小。
- 队列最大资源容量设置过小，AM最大资源百分比设置过小。
- 监控阈值设置过小。

处理步骤

检查NodeManager节点资源

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > ResourceManager(主)”，进入ResourceManager的WebUI页面。

步骤2 单击“Scheduler”，在“Application Queues”中查看root队列是否资源用满。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 对Yarn服务的NodeManager实例进行扩容。扩容后，查看告警是否消除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查队列最大资源容量和AM最大资源百分比

步骤4 查看pending任务对应的队列的资源是否用满。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager界面，选择 “租户资源 > 动态资源计划”，根据实际需要，适当增加相应的队列资源。查看告警是否消除。

- 是，处理完毕。
- 否，执行**步骤6**

调整监控阈值

步骤6 在FusionInsight Manager界面，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn > 任务 > 正在挂起的任务”，根据实际需要，适当增加该告警的监控阈值。

步骤7 等待5分钟，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“Yarn”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.141 ALM-18024 Yarn 任务挂起内存量超阈值

告警解释

告警模块按60秒周期检测Yarn当前挂起的内存量大小，当Yarn上面挂起的内存量大小超过阈值时，触发该告警。挂起的内存量表示当前所有提交的Yarn应用还没有满足的内存量总和。

告警属性

告警ID	告警级别	是否自动清除
18024	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
队列名	产生告警的队列名。
队列指标名	产生告警的队列指标名。

对系统的影响

- 应用任务结束时间变长。
- 新应用提交后长时间无法运行。

可能原因

- NodeManager节点资源过小。
- 队列最大资源容量设置过小，AM最大资源百分比设置过小。
- 监控阈值设置过小。

处理步骤

检查NodeManager节点资源

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > ResourceManager(主)”，进入ResourceManager的WebUI页面。

步骤2 单击“Scheduler”，在“Application Queues”中查看root队列是否资源用满。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 对Yarn服务的NodeManager实例进行扩容。扩容后，查看告警是否消除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查队列最大资源容量和AM最大资源百分比

步骤4 查看pending任务对应的队列的资源是否用满。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager界面，选择 “租户资源 > 动态资源计划”，根据实际需要，适当增加相应的队列资源。查看告警是否消除。

- 是，处理完毕。
- 否，执行**步骤6**。

调整监控阈值

步骤6 在FusionInsight Manager界面，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn > CPU和内存 > 挂起的内存量”，根据实际需要，适当增加该告警的监控阈值。

步骤7 等待5分钟，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“Yarn”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.142 ALM-18025 Yarn 被终止的任务数超过阈值

告警解释

告警模块按60秒周期检测Yarn root队列上被终止的应用的数量，当root队列上该监控周期内新增的被终止的应用的数量超过50，且连续发生3次以上时，触发该告警。

告警属性

告警ID	告警级别	是否自动清除
18025	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

大量应用任务被强制终止。

可能原因

- 人为强制终止大量任务。
- 系统出于某种错误终止任务。

处理步骤

检查告警详情。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”，打开告警页面。

步骤2 查看“Yarn被终止的任务数超过阈值”告警详情中的“附加信息”，确认监控阈值是否设置过小。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn > 其它 > root队列下被杀死的任务数”，修改该监控的阈值。执行**步骤6**。

步骤4 选择“集群 > *待操作集群的名称* > 服务 > Yarn > ResourceManager(主)”，进入ResourceManager的WebUI页面。

步骤5 单击“Applications”下的“KILLED”，单击最上面的任务。查看“Diagnostics”对应的描述信息，根据定位的任务被终止的详情（例如：被某用户终止）处理相关问题。

步骤6 等待3分钟，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Yarn”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.143 ALM-18026 Yarn 上运行失败的任务数超过阈值

告警解释

告警模块按60秒周期检测Yarn root队列上失败的应用的数量，当root队列上该监控周期内新增的运行失败的应用的数量超过50时，且连续发生3次以上，触发该告警。

告警属性

告警ID	告警级别	是否自动清除
18026	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

- 大量应用任务运行失败。
- 运行失败的任务需要重新提交。

可能原因

任务出于某种错误运行失败。

处理步骤

检查告警详情。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”，打开告警页面。

步骤2 查看“Yarn上运行失败的任务数超过阈值”告警详情中的“附加信息”，确认监控阈值是否设置过小。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Yarn > 其它 > root队列下失败的任务数”，修改该监控的阈值。执行**步骤6**。

步骤4 选择“集群 > *待操作集群的名称* > 服务 > Yarn > ResourceManager(主)”，进入ResourceManager的WebUI页面。

步骤5 单击“Applications”下的“FAILED”，单击最上面的任务。查看“Diagnostics”对应的描述信息，根据定位的任务失败原因，处理相关问题。

步骤6 等待3分钟，查看该告警是否消除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Yarn”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.144 ALM-19000 HBase 服务不可用

告警解释

告警模块按120秒周期检测HBase服务状态。当HBase服务不可用时产生该告警。

HBase服务恢复时，告警清除。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

无法进行数据读写和创建表等操作。

可能原因

- ZooKeeper服务异常。
- HDFS服务异常。
- HBase服务异常。
- 网络异常。

处理步骤

检查ZooKeeper服务状态。

步骤1 在FusionInsight Manager的服务列表中，查看ZooKeeper运行状态是否为“良好”。

- 是，执行**步骤5**。
- 否，执行**步骤2**。

步骤2 在告警列表中，查看是否有“ALM-13000 ZooKeeper服务不可用”告警产生。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 参考“ALM-13000 ZooKeeper服务不可用”的处理步骤处理该故障。

步骤4 等待几分钟后检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

检查HDFS服务状态。

步骤5 在告警列表中，查看是否有“ALM-14000 HDFS服务不可用”告警产生。

- 是，执行**步骤6**。
- 否，执行**步骤8**。

步骤6 参考“ALM-14000 HDFS服务不可用”的处理步骤处理该故障。

步骤7 等待几分钟后检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

步骤8 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS”，查看HDFS“安全模式”是否为“ON”。

- 是，执行**步骤9**。
- 否，执行**步骤12**。

步骤9 以**root**用户登录HDFS客户端。执行**cd**命令进入客户端安装目录，然后执行**source bigdata_env**。

如果集群采用安全版本，要进行安全认证。预先向管理员获取hdfs用户的密码，执行**kinit hdfs**命令，按提示输入密码。

步骤10 执行以下命令手动退出安全模式。

hdfs dfsadmin -safemode leave

步骤11 等待几分钟后检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤12**。

检查HBase服务状态。

步骤12 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HBase”。

步骤13 查看2个HMaster的状态是否为一“主”一“备”。

- 是，执行**步骤15**。
- 否，执行**步骤14**。

步骤14 单击“实例”，选择非主状态的HMaster实例，单击“更多 > 重启实例”重启HMaster，再次查看2个HMaster的状态是否为一“主”一“备”。

- 是，执行**步骤15**。
- 否，执行**步骤21**。

步骤15 选择“集群 > *待操作集群的名称* > 服务 > HBase > HMaster(主)”，进入HMaster的WebUI页面。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

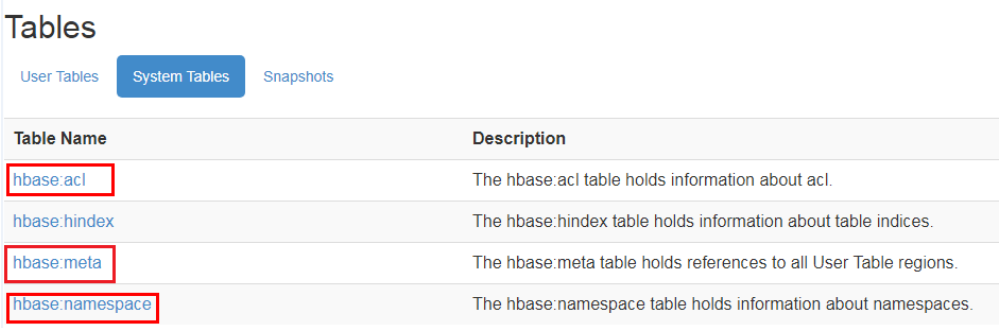
步骤16 查看Region Servers下是否存在至少一个RegionServer。

- 是，执行**步骤17**。
- 否，执行**步骤21**。

步骤17 查看“Tables > System Tables”，如**图10-34**，查看该标签的“Table Name”列下是否存在“hbase:meta”、“hbase:namespace”和“hbase:acl”。

- 是，执行**步骤18**。
- 否，执行**步骤19**。

图 10-34 HBase 系统表



步骤18 如**图10-34**，分别单击“hbase:meta”、“hbase:namespace”和“hbase:acl”超链接，查看所有页面是否能正常打开。如果页面能正常打开，说明表都正常。

- 是，执行**步骤19**。
- 否，执行**步骤23**。

说明

由于普通模式下的HBase默认未开启ACL权限控制，只有在手动开启ACL权限控制后才会存在“hbase:acl”表，需要检查该表，否则不需要检查该表。

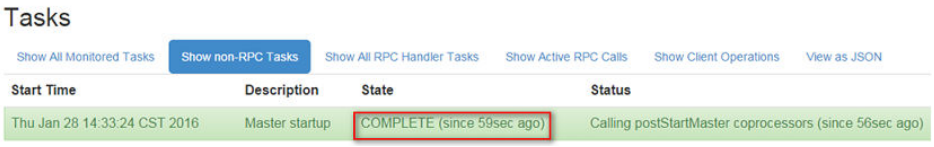
步骤19 查看HMaster的启动状态。

如**图10-35**在“Tasks” 下有“RUNNING”的状态表示HMaster正在启动，“State”列有HMaster处于“RUNNING”状态的时间。如**图10-36**中的“COMPLETE”状态表示HMaster启动完成。

查看HMaster是否持续了很长一段时间处于“RUNNING”状态。

图 10-35 HMaster 正在启动的状态

图 10-36 HMaster 启动完成的状态



- 是，执行**步骤20**。
- 否，执行**步骤21**。

步骤20 查看HMaster页面是否有hbase:meta长时间处于“Region in Transition”的状态。

图 10-37 Region 处于 Region in Transition 的状态

- 是，执行**步骤21**。
- 否，执行**步骤22**。

步骤21 确认在不影响业务的情况下，登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HBase > 更多 > 重启服务”，输入密码，单击“确定”。

- 是，执行**步骤22**。
- 否，执行**步骤23**。

步骤22 等待几分钟后检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤23**。

检查HMaster和依赖组件之间的网络连接。

步骤23 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > HBase”。

步骤24 单击“实例”，显示HMaster实例列表，记录“HMaster(主)”行的“管理IP”。

步骤25 以**omm**用户通过**步骤24**获取的IP地址登录主HMaster节点。

步骤26 执行**ping**命令，查看主HMaster节点和依赖组件所在主机的网络连接是否正常。（依赖组件包括ZooKeeper、HDFS和Yarn等，获取依赖组件所在主机的IP地址的方式和获取主HMaster的IP地址的方式相同。）

- 是，执行**步骤29**。
- 否，执行**步骤27**。

步骤27 联系网络管理员恢复网络。

步骤28 在告警列表中，查看“HBase服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤29**。

收集故障信息。

步骤29 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤30 在“服务”中勾选待操作集群的如下节点信息。

- ZooKeeper
- HDFS
- HBase

步骤31 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤32 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.145 ALM-19006 HBase 容灾同步失败

告警解释

告警模块每30s检查一次HBase容灾数据的同步状态，当同步容灾数据到备集群失败时，发送该告警。

当容灾数据同步成功后，告警清除。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19006	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

无法同步集群中HBase的数据到备集群，导致主备集群数据不一致。

可能原因

- 备集群HBase服务异常。
- 网络异常。

处理步骤

观察告警是否自动修复。

步骤1 在主集群的FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤2 在告警列表中单击该告警，从完整的告警信息中的“产生时间”处获得告警的产生时间，查看告警是否持续超过5分钟。

- 是，执行**步骤4**。
- 否，执行**步骤3**。

步骤3 等待5分钟后检查本告警是否自动恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查备集群HBase服务状态。

步骤4 登录主集群FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤5 在告警列表中单击该告警，从完整的告警信息中的“定位信息”处获得“主机名”。

步骤6 以**omm**用户进入主集群HBase客户端所在节点。

如果集群采用了安全版本，要进行安全认证，然后使用**hbase**用户进入**hbase shell**界面。

cd /opt/Bigdata/client

source ./bigdata_env

kinit *hbaseuser*

步骤7 执行**status 'replication', 'source'**命令查看故障节点的容灾同步状态。

节点的容灾同步状态如下：

```
10-10-10-153:
SOURCE: PeerID=abc, SizeOfLogQueue=0, ShippedBatches=2, ShippedOps=2, ShippedBytes=320,
LogReadInBytes=1636, LogEditsRead=5, LogEditsFiltered=3, SizeOfLogToReplicate=0,
TimeForLogToReplicate=0, ShippedHFiles=0, SizeOfHFileRefsQueue=0, AgeOfLastShippedOp=0,
TimeStampsOfLastShippedOp=Mon Jul 18 09:53:28 CST 2016, Replication Lag=0,
FailedReplicationAttempts=0
SOURCE: PeerID=abc1, SizeOfLogQueue=0, ShippedBatches=1, ShippedOps=1, ShippedBytes=160,
LogReadInBytes=1636, LogEditsRead=5, LogEditsFiltered=3, SizeOfLogToReplicate=0,
TimeForLogToReplicate=0, ShippedHFiles=0, SizeOfHFileRefsQueue=0, AgeOfLastShippedOp=16788,
TimeStampsOfLastShippedOp=Sat Jul 16 13:19:00 CST 2016, Replication Lag=16788,
FailedReplicationAttempts=5
```

步骤8 找到“FailedReplicationAttempts”的值大于0的记录所对应的“PeerID”值。

如上步骤中，故障节点“10-10-10-153”同步数据到“PeerID”为“abc1”的备集群失败。

步骤9 继续执行**list_peers**命令，查找该“PeerID”对应的集群和HBase实例。

```
PEER_ID CLUSTER_KEY STATE TABLE_CFS
abc1 10.10.10.110,10.10.10.119,10.10.10.133:2181:/hbase2 ENABLED
abc 10.10.10.110,10.10.10.119,10.10.10.133:2181:/hbase ENABLED
```

如上所示，**/hbase2**表示数据是同步到备集群的HBase2实例。

步骤10 在备集群FusionInsight Manager的服务列表中，查看通过**步骤9**获取的HBase实例运行状态是否为“良好”。

- 是，执行**步骤14**。
- 否，执行**步骤11**。

步骤11 在告警列表中，查看是否有“ALM-19000 HBase服务不可用”告警产生。

- 是，执行**步骤12**。
- 否，执行**步骤14**。

步骤12 参考“ALM-19000 HBase服务不可用”的处理步骤处理该故障。

步骤13 等待几分钟后检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤14**。

检查主备集群RegionServer之间的网络连接。

步骤14 登录主集群FusionInsight Manager界面，选择“运维 > 告警 > 告警”。

步骤15 在告警列表中单击该告警，从完整的告警信息中“定位信息”处获得“主机名”。

步骤16 以**omm**用户通过**步骤15**获取的IP地址登录故障RegionServer节点。

步骤17 执行**ping**命令，查看故障RegionServer节点和备集群RegionServer所在主机的网络连接是否正常。

- 是，执行**步骤20**。
- 否，执行**步骤18**。

步骤18 联系网络管理员恢复网络。

步骤19 网络恢复后，在告警列表中，查看本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤20**。

收集故障信息。

步骤20 在主备集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤21 在“服务”中勾选待操作集群的“HBase”。

步骤22 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤23 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.146 ALM-19007 HBase GC 时间超出阈值

告警解释

系统每60秒周期性检测HBase服务的老年代GC时间，当检测到HBase服务的老年代GC时间超出阈值（默认连续3次检测超过5秒）时产生该告警。 在FusionInsight Manager首页，用户可通过选择“运维 > 告警 > 阈值设置> HBase > GC > GC中回收old区所花时长”修改阈值。 当HBase服务的老年代GC时间小于或等于阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

老年代GC时间超出阈值，会影响到HBase数据的读写。

可能原因

该节点HBase实例内存使用率过大，或配置的堆内存不合理，或HBase存在大量的IO操作，导致进程GC频繁。

处理步骤

检查GC时间

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19007”的告警，查看“定位信息”中的角色名并确定实例的IP地址。

- 告警上报的角色是HMaster，执行**步骤2**。
- 告警上报的角色是RegionServer，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，单击告警上报的HMaster，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > GC > HMaster的GC时间”，单击“确定”，查看该图表中“GC中回收old区所花时长”监控项的值是否连续3个检测周期大于阈值（默认阈值为5秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，单击告警上报的RegionServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > GC > RegionServer的GC时间”，单击“确定”，查看该图表中“GC中回收old区所花时长”监控项的值是否连续3个检测周期大于阈值（默认阈值为5秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

查看JVM的当前配置

步骤4 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”。 在搜索框中输入“GC_OPTS”，确定当前告警角色HMaster(HBase->HMaster)，RegionServer(HBase->RegionServer)的“GC_OPTS”内存参数。将GC_OPTS参数中的“-Xmx”和“-XX:CMSInitiatingOccupancyFraction”的值参考以下说明进行调整。

说明

1. HMaster的GC参数配置建议：
   - 建议“-Xms”和“-Xmx”设置成相同的值，这样可以避免JVM动态调整堆内存大小时影响性能。
   - 调整“-XX:NewSize”大小时，建议把其设置成和“-XX:MaxNewSize”相同，均为“-Xmx”大小的1/8。
   - 当HBase集群规模越大、Region数量越多时，可以适当调大HMaster的GC_OPTS参数，配置建议如下：Region总数小于10万个，“-Xmx”设置为4G；超过10万个，“-Xmx”设置为不小于6G；超过10万时，每增加35000个Region，增加2G的“-Xmx”，整体的“-Xmx”的大小不超过32G。
2. RegionServer的GC参数配置建议：
   - 建议“-Xms”和“-Xmx”设置成相同的值，这样可以避免JVM动态调整堆内存大小时影响性能。
   - 调整“-XX:NewSize”大小的时候，建议把其设置为“-Xmx”大小的1/8。
   - RegionServer需要的内存一般比HMaster要大。在内存充足的情况下，堆内存可以相对设置大一些。
   - 根据机器的内存大小设置“-Xmx”大小：机器内存>200G，“-Xmx”设置为32G；128G<机器内存<200G，“-Xmx”设置为16G；机器内存<128G，“-Xmx”设置为8G。“-Xmx”配置为32G，可支持单RegionServer节点2000个Region，200个热点Region。
   - “XX:CMSInitiatingOccupancyFraction”建议设置为“100 * (hfile.block.cache.size +hbase.regionserver.global.memstore.size)”，最大值不超过85。

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在主备集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“HBase”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.147 ALM-19008 HBase 服务进程堆内存使用率超出阈值

告警解释

系统每30秒周期性检测HBase服务堆内存使用状态，当检测到HBase服务堆内存使用率超出阈值（最大内存的90%）时产生该告警。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

HBase可用内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点HBase服务堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19008”的告警，查看“定位信息”中的角色名并确定实例的IP地址。

- 告警上报的角色是HMaster，执行**步骤2**。
- 告警上报的角色是RegionServer，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，单击告警上报的HMaster，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > HMaster堆内存使用率与直接内存使用率统计”，单击“确定”，查看HBase服务进程使用的堆内存是否已达到HBase服务进程设定的最大堆内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，单击告警上报的RegionServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > RegionServer堆内存使用率与直接内存使用率统计”，单击“确定”，查看HBase服务进程使用的堆内存是否已达到HBase服务进程设定的最大堆内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”， 选择“HMaster/RegionServer > 系统”，将“GC_OPTS”参数中“-Xmx”的值参考以下说明进行调大。

说明

1. HMaster的GC参数配置建议
   - 建议“-Xms”和“-Xmx”设置成相同的值，这样可以避免JVM动态调整堆内存大小时影响性能。
   - 调整“-XX:NewSize”大小的时候，建议把其设置成和“-XX:MaxNewSize”相同，均为“-Xmx”大小的1/8。
   - 当HBase集群规模越大、Region数量越多时，可以适当调大HMaster的GC_OPTS参数，配置建议如下：Region总数小于10万个，“-Xmx”设置为4G；超过10万个，“-Xmx”设置为不小于6G；超过10万时，每增加35000个Region，增加2G的“-Xmx”，整体的“-Xmx”的大小不超过32G。
2. RegionServer的GC参数配置建议
   - 建议“-Xms”和“-Xmx”设置成相同的值，这样可以避免JVM动态调整堆内存大小时影响性能。
   - 调整“-XX:NewSize”大小的时候，建议把其设置为“-Xmx”大小的1/8。
   - RegionServer需要的内存一般比HMaster要大。在内存充足的情况下，堆内存可以相对设置大一些。
   - 根据机器的内存大小设置“-Xmx”大小：机器内存>200G，“-Xmx”设置为32G；128G<机器内存<200G，“-Xmx”设置为16G；机器内存<128G，“-Xmx”设置为8G。“-Xmx”配置为32G，可支持单RegionServer节点2000个Region，200个热点Region。

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”勾选待操作集群的“HBase”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.148 ALM-19009 HBase 服务进程直接内存使用率超出阈值

告警解释

系统每30秒周期性检测HBase服务直接内存使用状态，当检测到HBase服务直接内存使用率超出阈值（最大内存的90%）时产生该告警。

直接内存使用率小于阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

HBase可用的直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点HBase服务直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19009”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

- 告警上报的角色是HMaster，执行**步骤2**。
- 告警上报的角色是RegionServer，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，单击告警上报的HMaster，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > HMaster堆内存使用率与直接内存使用率统计”，单击“确定”，查看HBase服务进程使用的直接内存是否已达到HBase服务进程设定的最大直接内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤8**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，单击告警上报的RegionServer，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > RegionServer堆内存使用率与直接内存使用率统计”，单击“确定”，查看HBase服务进程使用的直接内存是否已达到HBase服务进程设定的最大直接内存的90%。

- 是，执行**步骤4**。
- 否，执行**步骤8**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”， 选择“HMaster/RegionServer > 系统”，查看“GC_OPTS”参数中是否存在“XX:MaxDirectMemorySize”。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”， 选择“HMaster/RegionServer > 系统”，在“GC_OPTS”中把参数“XX:MaxDirectMemorySize”删除。

步骤6 查看告警信息，是否产生“ALM-19008 HBase服务进程堆内存使用率超出阈值”告警。

- 是，参考“ALM-19008 HBase服务进程堆内存使用率超出阈值”处理告警。
- 否，执行**步骤8**。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“HBase”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.149 ALM-19011 RegionServer 的 Region 数量超出阈值

告警解释

系统每30秒周期性检测每个HBase服务实例中每个RegionServer的Region数。该指标可以在HBase服务监控界面和RegionServer角色监控界面查看，当检测到某个RegionServer上的Region数超出阈值（默认连续20次超过默认阈值2000）时产生该告警。用户可通过“运维 > 告警 > 阈值设置> 服务 > HBase”修改阈值。当Region数小于或等于阈值时，告警消除。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19011	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

RegionServer的Region数超出阈值，会影响HBase的数据读写性能。

可能原因

- RegionServer的Region分布不均衡。
- HBase集群规模过小。

处理步骤

查看告警定位信息

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19011”的告警，查看“定位信息”中产生该告警的服务实例和主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，查看RegionServer上Region分布是否均衡。

说明

admin用户默认不具备其他组件的管理权限，如果访问组件原生界面时出现因权限不足而打不开页面或内容显示不全时，可手动创建具备对应组件管理权限的用户进行登录。

- 是，执行**步骤9**。
- 否，执行**步骤3**。

图 10-38 HBase 的 WebUI

负载均衡

步骤3 以**root**用户登录HBase客户端所在节点。进入客户端安装目录，设置环境变量：

cd *客户端安装目录*

source bigdata_env

如果集群采用安全版本，要进行安全认证。执行**kinit hbase**命令，按提示输入密码（向管理员获取密码）。

步骤4 执行以下命令进入hbase shell，查看目前负载均衡功能是否打开：

hbase shell

balancer_enabled

- 是，执行**步骤6**。
- 否，执行**步骤5**。

步骤5 在hbase shell，中执行命令打开负载均衡功能，并执行命令查看确认成功打开：

balance_switch true

balancer_enabled

步骤6 执行**balancer**命令手动触发负载均衡。

说明

建议打开和手动触发负载均衡操作在业务低峰期进行。

步骤7 FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，刷新页面查看Region分布是否均衡。

- 是，执行**步骤8**。
- 否，执行**步骤21**。

步骤8 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

清理无用HBase表

说明

在清理过程中，请谨慎操作，确保删除数据的准确性。

步骤9 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，查看该HBase服务实例上存储的表并记录可删除的无用表。

步骤10 在hbase shell中，执行**disable**和**drop**命令，确认删除无用表，以减少Region数：

disable '*待删除表名*'

drop '*待删除表名*'

步骤11 在hbase shell中，执行命令查看目前负载均衡功能是否打开：

balancer_enabled

- 是，执行**步骤13**。
- 否，执行**步骤12**。

步骤12 在hbase shell中，执行命令打开负载均衡功能并确认成功打开：

balance_switch true

balancer_enabled

步骤13 在hbase shell中，执行**balancer**命令手动触发负载均衡。

步骤14 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击产生该告警的HBase服务实例，单击“HMaster(主)”，打开该HBase实例的WebUI，刷新页面查看Region分布是否均衡。

- 是，执行**步骤15**。
- 否，执行**步骤21**。

步骤15 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤16**。

调整阈值

步骤16 在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > HBase > 单个RegionServer的Region数目”，选中目前应用的规则，单击“修改”查看目前的阈值设置是否合理。

- 如果过小，则根据集群实际情况，增大阈值，执行**步骤17**。
- 如果阈值设置合理，则执行**步骤18**。

步骤17 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤18**。

系统扩容

步骤18 对HBase集群扩容，增加节点，并在节点上增加RegionServer实例，然后按照“负载均衡”小节中，打开负载均衡功能并手动触发。

步骤19 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务”，单击产生该告警的HBase服务实例，单击“HMaster(主)”，打开该HBase实例的WebUI，刷新页面查看Region分布是否均衡。

- 是，执行**步骤20**。
- 否，执行**步骤21**。

步骤20 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤21**。

收集故障信息

步骤21 在主备集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤22 在“服务”中勾选待操作集群的“HBase”。

步骤23 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤24 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.150 ALM-19012 HBase 系统表目录或文件丢失

告警解释

系统按120秒周期性检测HBase在HDFS上的如下目录和文件是否存在，当检测到文件或者目录不存在时，上报该告警。当文件或目录都恢复后，告警恢复。

检查内容：

- 命名空间hbase在HDFS上的目录。

- hbase.version文件。
- hbase:meta表在HDFS上的目录、.tableinfo和.regioninfo文件。
- hbase:namespace表在HDFS上的目录、.tableinfo和.regioninfo文件。
- hbase:hindex表在HDFS上的目录、.tableinfo和.regioninfo文件。
- hbase:acl表在HDFS上的目录、.tableinfo和.regioninfo文件(该表在普通模式集群默认不存在)。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19012	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

HBase服务重启/启动失败。

可能原因

HDFS上的文件或者目录缺失。

处理步骤

检查告警原因

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19012”的告警，查看“告警原因”中的是否提示未知异常。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 在FusionInsight Manager首页，选择“运维 > 备份恢复 > 备份管理”，查看任务名称为“default”的备份任务或者其他执行成功的用户自己配置的HBase元数据备份任务是否有执行成功的记录。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 使用最近一次备份的元数据，对HBase服务的元数据进行恢复操作。

收集故障信息

步骤4 在主备集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的有问题的HBase服务。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.151 ALM-19013 region 处在 RIT 状态的时长超过阈值

告警解释

系统按300秒周期性检测HBase上的region处在RIT状态的数量。当检测到处在RIT状态的region时长超过阈值时长（连续两次超过阈值），上报该告警。当处在超时状态的region都恢复后，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如HBase1服务不可用，则“定位信息”中显示服务名=HBase1，处理步骤中的操作对象也应由HBase调整为HBase1。

告警属性

告警ID	告警级别	是否自动清除
19013	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

表的部分数据丢失或不可用。

可能原因

- Compaction永久阻塞。
- HDFS文件异常。

处理步骤

检查告警原因

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19013”的告警，查看“定位信息”中的主机名及角色名。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HBase ”，单击图表区域右上角的下拉菜单，选择“定制 > 服务 > 处在RIT状态的region数”，单击“确定”，查看该图表中“处在RIT状态达到阈值时长的region数”监控项是否在连续3个检测周期内检测到值。（默认阈值为60秒）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 选择“集群 > *待操作集群的名称* > 服务 > HBase > HMaster（主） > Tables” ，查看是否只是某一个表的region RIT状态超时。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 在客户端执行**hbase hbck**是否报错“No table descriptor file under hdfs://hacluster/hbase/data/default/table”。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 以**root**用户登录客户端。执行如下命令：

cd *客户端安装目录*

source bigdata_env

如为安全模式集群，请执行**kinit hbase**

登录HMaster WebUI，在导航栏选择“Procedure & Locks”，在Procedures查看是否有处于Waiting状态的process id。如果有，需要执行以下命令将procedure lock释放：

hbase hbck -j *客户端安装目录*/**HBase/hbase/tools/hbase-hbck2-*.jar bypass -o** *pid*

查看State是否处于Bypass状态，如果界面上的procedures一直处于RUNNABLE(Bypass)状态，需要进行主备切换。执行**assigns**命令使region重新上线。

hbase hbck -j *客户端安装目录*/**HBase/hbase/tools/hbase-hbck2-*.jar assigns -o** *regionName*

步骤6 在客户端执行**hbase hbck**，查看否报错“No table descriptor file under hdfs://hacluster/hbase/data/default/table”。

- 是，执行**步骤7**。
- 否，处理完毕。

收集故障信息

步骤7 在主备集群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的有问题的HBase服务。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.152 ALM-19014 在 ZooKeeper 上的容量配额使用率严重超过阈值

告警解释

系统每120秒周期性检测HBase服务的ZNode使用情况，当检测到HBase服务的ZNode容量使用率超出紧急告警的阈值（默认90%）时产生该告警。

当ZNode的容量使用率小于严重告警的阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如“定位信息”中显示“服务名=HBase-1”，处理步骤中的操作对象也应由HBase调整为HBase-1。

告警属性

告警ID	告警级别	是否自动清除
19014	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Threshold	产生告警的阈值。

对系统的影响

产生该告警表示HBase服务的ZNode的容量使用率已经严重超过规定的阈值，会导致HBase服务的写入请求失败。

可能原因

- HBase配置了容灾并且容灾存在数据同步失败或者同步速度慢。
- HBase集群存在大量的WAL文件在进行split。

处理步骤

检查ZNode容量配置和使用量

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19014”的告警，查看“附加信息”中的阈值。

步骤2 以**root**用户登录HBase客户端。执行以下命令进入客户端安装目录：

cd *客户端安装目录*

然后执行以下命令设置环境变量：

source bigdata_env

如果集群采用安全版本，要执行以下命令进行安全认证：

kinit hbase

按提示输入密码（向管理员获取密码）。

步骤3 执行**hbase zkcli**命令进入ZooKeeper客户端，然后执行命令**listquota /hbase**查看对应HBase服务的ZNode容量配额，其中命令中的ZNode根目录为对应HBase服务的参数

“zookeeper.znode.parent”所指定。下图标注所示即为当前HBase服务根ZNode的容量配置。

```
[zk: 189-185-229-159:24002,189-185-229-114:24002,189-185-229-251:24002(CONNECTED) 145] listquota /hbase
absolute path is /zookeeper/quota/hbase
Output quota for /hbase count=1500000,bytes=10240
Output stat for /hbase count=42,bytes=1601
```

步骤4 执行命令**getusage /hbase/splitWAL**查看该ZNode的容量使用情况，查看返回结果的“**Data size**”跟ZNode容量配额的比值是否接近告警的阈值。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“12007”、“19000”或者“19013”且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤6 执行命令**getusage /hbase/replication**查看该ZNode的容量使用情况，查看返回结果的“**Data size**”跟ZNode容量配额的比值是否接近告警的阈值。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“19006”且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 等待5分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HBase”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.153 ALM-19015 在 ZooKeeper 上的数量配额使用率超过阈值

告警解释

系统每120秒周期性检测HBase服务的ZNode使用情况，当检测到HBase服务的ZNode数量使用率超出告警的阈值（默认75%）时产生该告警。

当ZNode的数量使用率小于告警的阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如“定位信息”中显示服务名=HBase-1，处理步骤中的操作对象也应由HBase调整为HBase-1。

告警属性

告警ID	告警级别	是否自动清除
19015	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Threshold	产生告警的阈值。

对系统的影响

产生该告警表示HBase服务的ZNode的数量使用率已经超过规定的阈值，如果不及时处理，可能会导致问题级别升级为紧急，影响数据写入。

可能原因

- HBase配置了容灾并且容灾存在数据同步失败或者同步速度慢；
- HBase集群存在大量的WAL文件在进行split。

处理步骤

检查ZNode数量配额和使用量

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19015”的告警，查看“附加信息”中的阈值。

步骤2 以**root**用户登录HBase客户端。执行以下命令进入客户端安装目录：

cd *客户端安装目录*

然后执行以下命令设置环境变量：

source bigdata_env

如果集群采用安全版本，要执行以下命令进行安全认证：

kinit hbase

按提示输入密码（向管理员获取密码）。

步骤3 执行**hbase zkcli**命令进入ZooKeeper客户端，然后执行命令**listquota /hbase**查看对应HBase服务的ZNode数量配额，其中命令中的ZNode根目录为对应HBase服务的参数“zookeeper.znode.parent”所指定。下图标注所示即为当前HBase服务根ZNode的数量配额。

```
[zk: 189-185-229-159:24002,189-185-229-114:24002,189-185-229-251:24002(CONNECTED) 7] listquota /hbase
absolute path is /zookeeper/quota/hbase
Output quota for /hbase count=1500000,bytes=10240
Output stat for /hbase count=59,bytes=1902
```

步骤4 执行命令**getusage /hbase/splitWAL**查看该ZNode的数量使用情况，查看返回结果的“**Node count**”跟ZNode数量配额的比值是否接近告警的阈值。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“12007”、“19000”或者“19013”且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤6 执行命令**getusage /hbase/replication**查看该ZNode的数量使用情况，查看返回结果的“**Node count**”跟ZNode数量配额的比值是否接近告警的阈值。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“19006”并且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HBase”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.154 ALM-19016 在 ZooKeeper 上的数量配额使用率严重超过阈值

告警解释

系统每120秒周期性检测HBase服务的ZNode使用情况，当检测到HBase服务的ZNode数量使用率超出紧急告警的阈值（默认90%）时产生该告警。

当ZNode的数量使用率小于严重告警的阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如“定位信息”中显示服务名=HBase-1，处理步骤中的操作对象也应由HBase调整为HBase-1。

告警属性

告警ID	告警级别	是否自动清除
19016	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Threshold	产生告警的阈值。

对系统的影响

产生该告警表示HBase服务的ZNode的数量使用率已经严重超过规定的阈值，会导致HBase服务的写入请求失败。

可能原因

- HBase配置了容灾并且容灾存在数据同步失败或者同步速度慢；
- HBase集群存在大量的WAL文件在进行split。

处理步骤

检查ZNode数量配置和使用量

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19016”的告警，查看“附加信息”中的阈值。

步骤2 以**root**用户登录HBase客户端。执行以下命令进入客户端安装目录：

cd *客户端安装目录*

然后执行以下命令设置环境变量：

source bigdata_env

如果集群采用安全版本，要执行以下命令进行安全认证：

kinit hbase

按提示输入密码（向管理员获取密码）。

步骤3 执行**hbase zkcli**命令进入ZooKeeper客户端，然后执行命令**listquota /hbase**查看对应HBase服务的ZNode容量配额，其中命令中的ZNode根目录为对应HBase服务的参数“zookeeper.znode.parent”所指定。下图标注所示即为当前HBase服务根ZNode的容量配置。

```
[zk: 189-185-229-159:24002,189-185-229-114:24002,189-185-229-251:24002(CONNECTED) 7] listquota /hbase
absolute path is /zookeeper/quota/hbase
Output quota for /hbase count=1500000,bytes=10240
Output stat for /hbase count=59,bytes=1902
```

步骤4 执行命令**getusage /hbase/splitWAL**查看该ZNode的数量使用情况，查看返回结果的“**Node count**”跟ZNode数量配额的比值是否接近告警的阈值。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“12007”、“19000”或者“19013”且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤6 执行命令**getusage /hbase/replication**查看该ZNode的数量使用情况，查看返回结果的“**Node count**”跟ZNode数量配额的比值是否接近告警的阈值。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“19006”并且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HBase”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.155 ALM-19017 在 ZooKeeper 上的容量配额使用率超过阈值

告警解释

系统每120秒周期性检测HBase服务的ZNode使用情况，当检测到HBase服务的ZNode容量使用率超出告警的阈值（默认75%）时产生该告警。

当ZNode的容量使用率小于告警的阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如“定位信息”中显示服务名=HBase-1，处理步骤中的操作对象也应由HBase调整为HBase-1。

告警属性

告警ID	告警级别	是否自动清除
19017	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Threshold	产生告警的阈值。

对系统的影响

产生该告警表示HBase服务的ZNode的容量使用率已经超过规定的阈值，如果不及时处理，可能会导致问题级别升级为紧急，影响数据写入。

可能原因

- HBase配置了容灾并且容灾存在数据同步失败或者同步速度慢；
- HBase集群存在大量的WAL文件在进行split。

处理步骤

检查ZNode容量配置和使用量

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“19017”的告警，查看“附加信息”中的阈值。

步骤2 以**root**用户登录HBase客户端。执行以下命令进入客户端安装目录：

cd *客户端安装目录*

然后执行以下命令设置环境变量：

source bigdata_env

如果集群采用安全版本，要执行以下命令进行安全认证：

kinit hbase

按提示输入密码（向管理员获取密码）。

步骤3 执行**hbase zkcli**命令进入ZooKeeper客户端，然后执行命令**listquota /hbase**查看对应HBase服务的ZNode容量配额，其中命令中的ZNode根目录为对应HBase服务的参数“zookeeper.znode.parent”所指定。下图标注所示即为当前HBase服务根ZNode的容量配置。

```
[zk: 189-185-229-159:24002,189-185-229-114:24002,189-185-229-251:24002(CONNECTED) 145] listquota /hbase
absolute path is /zookeeper/quota/hbase
Output quota for /hbase count=1500000,bytes=10240
Output stat for /hbase count=42,bytes=1601
```

步骤4 执行命令**getusage /hbase/splitWAL**查看该ZNode的容量使用情况，查看返回结果的“**Data size**”跟ZNode容量配额的比值是否接近告警的阈值。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 在FusionInsight Manager首页，查看是否存在“告警ID”为“12007”、“19000”或者“19013”且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤7**。

步骤6 执行命令**getusage /hbase/replication**查看该ZNode的容量使用情况，查看返回结果的“**Data size**”跟ZNode容量配额的比值是否接近告警的阈值。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在“告警ID”为“19006”并且“定位信息”中的“服务名”为当前HBase服务的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“HBase”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.156 ALM-19018 HBase 合并队列超出阈值

告警解释

系统每300秒周期性检测HBase服务的compaction队列长度，当检测到HBase服务的compaction队列长度超过告警的阈值（默认100）时产生该告警。当compaction队列长度小于告警的阈值时，告警恢复。

说明

若集群启用了多实例功能且安装了多个HBase服务，请根据“定位信息”的“服务名”值来确定具体产生告警的HBase服务。例如“定位信息”中显示服务名=HBase-1，处理步骤中的操作对象也应由HBase调整为HBase-1。

告警属性

告警ID	告警级别	是否自动清除
19018	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

产生该告警表示HBase服务的compaction队列长度已经超过规定的阈值，如果不及时处理，可能会导致集群性能下降，影响数据读写。

可能原因

- HBase regionserver数太少。
- HBase 单个regionserver上region数过多。
- HBase regionserver堆大小较小。
- 资源不足。
- 相关参数配置不合理。

处理步骤

检查相关配置是否合理

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，查看是否存在告警ID为“19011”的告警。

- 是，单击对应告警右侧的“查看帮助”并按照帮助文档进行处理，执行**步骤3**。
- 否，执行**步骤2**。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > HBase > 配置 > 全部配置，搜索“hbase.hstore.compaction.min”，“hbase.hstore.compaction.max”，“hbase.hstore.compactionThreshold”，

“hbase.regionserver.thread.compaction.small”和“hbase.regionserver.thread.compaction.throttle”，适当调大其值。

步骤3 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤4**。

收集故障信息

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的“HBase”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.157 ALM-19019 HBase 容灾等待同步的 HFile 文件数量超过阈值

告警解释

系统每30秒周期性检测每个HBase服务实例RegionServer等待同步的HFile文件数量。该指标可以在RegionServer角色监控界面查看，当检测到某个RegionServer上的等待同步HFile文件数量超出阈值（默认连续20次超过默认阈值128）时产生该告警。用户可通过“运维 > 告警 > 阈值设置 > *待操作集群* > HBase”来修改阈值。当等待同步的HFile文件数量小于或等于阈值时，告警消除。

告警属性

告警ID	告警级别	是否自动清除
19019	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

RegionServer等待同步的HFile文件数量超出阈值，会影响HBase使用的ZNode超出阈值，影响HBase服务状态。

可能原因

- 网络异常。
- RegionServer的Region分布不均匀。
- 备集群HBase服务规模过小。

处理步骤

查看告警定位信息

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选择“告警ID”为“19019”的告警，查看“定位信息”中产生该告警的服务实例和主机名。

检查主备集群RegionServer之间的网络连接。

步骤2 执行**ping**命令，查看故障RegionServer节点和备集群RegionServer所在主机的网络连接是否正常。

- 是，执行**步骤5**
- 否，执行**步骤3**

步骤3 联系网络管理员恢复网络。

步骤4 网络恢复后，在告警列表中，查看本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

检查主集群RegionServer的Region分布情况

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，查看Region Servers上Region分布是否均衡。

步骤6 以**omm**用户登录故障RegionServer节点。

步骤7 进入客户端安装目录，设置环境变量。

cd *客户端安装目录*

source bigdata_env

如果集群采用安全版本，要进行安全认证。执行**kinit hbase**命令，按提示输入密码（向管理员获取密码）。

步骤8 执行以下命令查看目前负载均衡功能是否打开。

hbase shell

balancer_enabled

- 是，执行**步骤10**。
- 否，执行**步骤9**。

步骤9 在hbase shell中执行命令打开负载均衡功能，并执行命令查看确认成功打开。

balance_switch true

balancer_enabled

步骤10 执行**balancer**命令手动触发负载均衡。

说明

建议打开和手动触发负载均衡操作在业务低峰期进行。

步骤11 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤12**。

检查备集群HBase服务规模

步骤12 对HBase集群扩容，增加节点，并在节点上增加RegionServer实例。然后执行步骤**步骤6** -步骤**步骤10**，打开负载均衡功能并手动触发。

步骤13 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，刷新页面查看Region分布是否均衡。

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤15**。

收集故障信息

步骤15 在主备群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤16 在“服务”中勾选待操作集群的“HBase”。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.158 ALM-19020 HBase 容灾等待同步的 wal 文件数量超过阈值

告警解释

系统每30秒周期性检测每个HBase服务实例RegionServer等待同步的wal文件数量。该指标可以在RegionServer角色监控界面查看，当检测到某个RegionServer上的等待同步wal文件数量超出阈值（默认连续20次超过默认阈值128）时产生该告警。用户可通过“运维 > 告警 > 阈值设置> *待操作集群* > HBase”修改阈值。当等待同步的wal文件数量小于或等于阈值时，告警消除。

告警属性

告警ID	告警级别	是否自动清除
19020	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

RegionServer等待同步的wal文件数量超出阈值，会影响HBase使用的ZNode超出阈值，影响HBase服务状态。

可能原因

- 网络异常。
- RegionServer的Region分布不均匀。

- 备集群HBase服务规模过小。

处理步骤

查看告警定位信息

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选择“告警ID”为“19020”的告警，查看“定位信息”中产生该告警的服务实例和主机名。

检查主备集群RegionServer之间的网络连接。

步骤2 执行**ping**命令，查看故障RegionServer节点和备集群RegionServer所在主机的网络连接是否正常。

- 是，执行**步骤5**
- 否，执行**步骤3**

步骤3 联系网络管理员恢复网络。

步骤4 网络恢复后，在告警列表中，查看本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

检查主集群RegionServer的Region分布情况

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，查看Region Servers上Region分布是否均衡。

步骤6 以**omm**用户登录故障RegionServer节点。

步骤7 进入客户端安装目录，设置环境变量。

cd *客户端安装目录*

source bigdata_env

如果集群采用安全版本，要进行安全认证。执行**kinit hbase**命令，按提示输入密码（向管理员获取密码）。

步骤8 执行以下命令查看目前负载均衡功能是否打开。

hbase shell

balancer_enabled

- 是，执行**步骤10**。
- 否，执行**步骤9**。

步骤9 在hbase shell中执行命令打开负载均衡功能，并执行命令查看确认成功打开。

balance_switch true

balancer_enabled

步骤10 执行**balancer**命令手动触发负载均衡。

说明

建议打开和手动触发负载均衡操作在业务低峰期进行。

步骤11 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤12**。

检查备集群HBase服务规模

步骤12 对HBase集群扩容，增加节点，并在节点上增加RegionServer实例。然后执行步骤**步骤6**-步骤**步骤10**，打开负载均衡功能并手动触发。

步骤13 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > HBase”，单击“HMaster(主)”，打开该HBase实例的WebUI，刷新页面查看Region分布是否均衡。

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 观察该告警是否清除。

- 是，处理完毕。
- 否，执行**步骤15**。

收集故障信息

步骤15 在主备群的FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤16 在“服务”中勾选待操作集群的“HBase”。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.159 ALM-20002 Hue 服务不可用

告警解释

系统按60秒周期性检测Hue服务状态。当Hue服务不可用时产生该告警。

当Hue服务恢复时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
20002	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

系统无法提供数据加载，查询，提取服务。

可能原因

- Hue服务所依赖内部服务KrbServer故障。
- Hue服务所依赖内部服务DBService故障。
- 与DBService连接的网络异常。

处理步骤

检查KrbServer服务是否正常。

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”，在服务列表中查看“KrbServer”的“运行状态”是否为“良好”。

- 是，执行**步骤4**。
- 否，执行**步骤2**。

步骤2 手动重启KrbServer服务。

步骤3 等待几分钟。检查“Hue服务不可用”告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查DBService是否正常

步骤4 登录FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”。

步骤5 在服务列表中查看DBService服务运行状态是否为“良好”。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 重启DBService服务。

说明

重启服务需要输入FusionInsight Manager管理员密码。

步骤7 等待几分钟。检查“Hue服务不可用”告警是否恢复。

- 是，操作结束。
- 否，执行**步骤8**。

检查与DBService连接的网络是否正常

步骤8 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Hue > 实例”，记录主Hue的IP地址。

步骤9 登录主Hue的IP地址。

步骤10 执行**ping**命令，查看主Hue所在主机与DBService服务所在主机的网络连接是否正常。（获取DBService服务IP地址的方式和获取主Hue IP地址的方式相同。）

- 是，执行**步骤13**。
- 否，执行**步骤11**。

步骤11 联系网络管理员恢复网络。

步骤12 等待几分钟。检查“Hue服务不可用”告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

收集故障信息

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”框中勾选如下节点信息。

- Hue
- Controller

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤16 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Hue”。

步骤17 选择“更多 > 重启服务”，单击“确定”。

步骤18 检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤19**。

步骤19 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.160 ALM-24000 Flume 服务不可用

告警解释

告警模块按180秒周期检测Flume服务状态，当检测到Flume服务异常时，系统产生此告警。

当系统检测到Flume服务恢复正常，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

当Flume服务不可用时，Flume不能正常工作，数据传输业务中断。

可能原因

Flume实例全部故障。

处理步骤

步骤1 以**omm**用户登录Flume实例所在节点，执行**ps -ef|grep "flume.role=server"**命令查看当前节点是否存在flume进程。

- 是，执行**步骤3**。
- 否，重启Flume故障实例或Flume服务，执行**步骤2**。

步骤2 在告警列表中查看“Flume服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤3**。

收集故障信息。

步骤3 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤4 在“服务”框中勾选待操作集群的“Flume”。

步骤5 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤6 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.161 ALM-24001 Flume Agent 异常

告警解释

Flume Agent监控模块对Flume Agent状态进行监控，当Flume Agent进程故障（每5秒检测一次）或Flume Agent启动失败时（即时上报告警），系统产生此告警。

当检测到Flume Agent进程故障恢复，Flume Agent启动成功，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24001	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
AgentId	产生告警的Agent id。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

产生告警的Flume Agent实例无法正常启动，定义在该实例下的数据传输任务暂时中断，对于实时数据传输，会丢失实时数据。

可能原因

- JAVA_HOME目录不存在或JAVA权限异常。
- Flume Agent目录权限异常。
- Flume Agent启动失败。

处理步骤

检查JAVA_HOME目录是否存在或JAVA权限是否正确

步骤1 以**root**用户登录故障节点IP所在主机。

步骤2 执行以下命令获取发生告警的Flume客户端安装目录。（AgentId可以在告警的“定位信息”中获取）

ps -ef|grep *AgentId* **| grep -v grep | awk -F 'conf-file ' '{print $2}' | awk -F 'fusioninsight' '{print $1}'**

步骤3 使用“**su -** *Flume安装*用户”命令切换到Flume安装用户，执行**cd** *Flume客户端安装目录*/**fusioninsight-flume-1.9.0/conf/**命令，进入Flume的配置目录。

步骤4 执行**cat ENV_VARS | grep JAVA_HOME**命令。

步骤5 检查JAVA_HOME目录是否存在，若步骤**步骤4**执行结果返回不为空，且**ll $JAVA_HOME/**不为空，则JAVA_HOME目录存在。

- 是，执行**步骤7**。
- 否，执行**步骤6**。

步骤6 指定正确的JAVA_HOME目录。

步骤7 执行**$JAVA_HOME/bin/java -version**命令检查Flume Agent运行用户是否有JAVA可执行权限，若可以查到java版本，这说明JAVA权限满足，否则不满足。

- 是，执行**步骤9**。
- 否，执行**步骤8**。

说明

JAVA_HOME为安装Flume客户端时export导出的环境变量，也可以进入到*Flume客户端安装目录*/**fusioninsight-flume-1.9.0/conf**目录下，执行**cat ENV_VARS | grep JAVA_HOME**命令来查看变量的值。

步骤8 执行**chmod 750 $JAVA_HOME/bin/java**命令赋予Flume Agent运行用户JAVA可执行权限。

检查Flume Agent的目录权限。

步骤9 以**root**用户登录故障节点IP所在主机。

步骤10 执行以下命令，进入Flume Agent的安装目录。

cd *Flume客户端安装目录*/**fusioninsight-flume-1.9.0/conf/**

步骤11 执行**ls -al * -R**命令，检查所有文件的所有者是否均是Flume Agent运行用户。

- 是，执行**步骤12**。
- 否，使用**chown**命令修改文件所有者为Flume Agent运行用户。

检查Flume Agent配置。

步骤12 执行**cat properties.properties | grep spooldir**以及**cat properties.properties | grep TAILDIR**命令，确认Flume Source是否是spooldir类型或TAILDIR类型，若任意一个命令有返回值，则为spooldir类型或TAILDIR类型。

- 是，执行**步骤13**。
- 否，执行**步骤17**。

步骤13 查看数据监控目录是否存在。

- 是，执行**步骤15**。
- 否，执行**步骤14**。

说明

查看spooldir监控目录，执行命令：**cat properties.properties | grep spoolDir**

```
[root@[redacted] /tmp/test/fusioninsight-flume-1.9.0/conf]# cat properties.properties | grep spoolDir
client.sources.aa1.spoolDir = /opt/liuxingcheng/flumeclient/sourcedata/flumesourcedata1
[root@[redacted] /tmp/test/fusioninsight-flume-1.9.0/conf]#
```

查看TAILDIR监控目录，执行命令：**cat properties.properties | grep parentDir**

```
[root@[redacted] /tmp/test/fusioninsight-flume-1.9.0/conf]# cat properties.properties | grep parentDir
server.sources.AAAA.filegroups.f1.parentDir = /tmp/flumetest/taildir_data
[root@[redacted] /tmp/test/fusioninsight-flume-1.9.0/conf]#
```

步骤14 指定服务器上用户自定义已经存在的数据监控目录。

步骤15 查看Flume Agent运行用户对步骤**步骤13**所指定的监控目录是否有可读可写可执行权限。

- 是，执行**步骤17**。
- 否，执行**步骤16**。

说明

使用Flume运行用户进入监控目录，若可以创建文件，这说明Flume运行用户是否对该监控目录具有可读可写可执行权限。

步骤16 执行“**chmod 777** *Flume监控目录*”命令赋予Flume Agent运行用户对步骤**步骤13**监控目录的可读可写可执行权限。

步骤17 确认Flume Sink对接组件是否处于安全模式。

- 是，执行**步骤18**。
- 否，执行**步骤23**。

说明

若用户业务配置文件properties.properties的sink为hdfs sink、hbase sink，当配置文件中包含有keytab时，则Flume Sink对接组件处于安全模式。

若用户业务配置文件properties.properties的sink为kafka sink，当配置参数*.security.protocol的值为SASL_PLAINTEXT或为SASL_SSL时，则Flume Sink对接的Kafka处于安全模式。

步骤18 使用“**ll** *ketab路径命令*”查看配置文件“*.kerberosKeytab”参数所指的keytab认证路径是否存在。

- 是，执行**步骤20**。
- 否，执行**步骤19**。

说明

ketab路径查看方式：**cat properties.properties | grep keytab**

```
[root@[redacted] /tmp/test/fusioninsight-flume-1.9.0/conf]# cat properties.properties | grep keytab
lient.sinks.CCCC.kerberosKeytab = /opt/huawei/Bigdata/FusionInsight_Porter_8.0.0/1_11_Flume/etc/user.keytab
[root@[redacted] /tmp/test/fusioninsight-flume-1.9.0/conf]#
```

步骤19 将步骤**步骤18**中kerberosKeytab参数的值指定为用户自定的keytab路径，执行**步骤21**。

步骤20 执行步骤**步骤18**查看Flume Agent运行用户是否有访问keytab认证文件的权限，若返回为keytab路径，则表示有权限，否则无权限。

- 是，执行**步骤22**。
- 否，执行**步骤21**。

步骤21 执行“**chmod 755** *ketab文件*”赋予步骤**步骤19**中所指定的keytab文件的可读权限，并重启Flume进程。

步骤22 查看告警列表中该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤23**。

收集故障信息。

步骤23 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤24 在“服务”框中勾选待操作集群的“Flume”。

步骤25 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤26 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.162 ALM-24003 Flume Client 连接中断

告警解释

告警模块对Flume Server的连接端口状态进行监控。当Flume Client连接到Flume Server的某个端口，Client端连续3分钟未与Server端连接时，系统产生此告警。

当Flume Server收到Flume Client连接消息，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24003	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
客户端IP	Flume客户端IP地址。
客户端名称	Flume客户端的Agent名称。
sink名称	Flume Agent的sink名称。

对系统的影响

产生告警的Flume Client无法与Flume Server端进行通信，Flume Client端的数据无法传输到Flume Server端。

可能原因

- Flume Client端与Flume Server端网络故障。
- Flume Client端进程故障。
- Flume Client端配置错误。

处理步骤

检查Flume Client与Flume Server的网络状况。

步骤1 以**root**用户登录到告警定位参数中描述的Flume ClientIP所在主机。

步骤2 执行**ping** *Flume Server IP地址*命令，检查Flume Client到Flume Server的网络是否正常。

- 是，执行**步骤3**。
- 否，执行**步骤11**。

检查Flume Client端进程故障。

步骤3 以**root**用户登录到告警定位参数中描述的Flume ClientIP所在主机。

步骤4 执行**ps -ef|grep flume |grep client**命令，查看是否存在Flume Client进程。

- 是，执行**步骤5**。
- 否，执行**步骤11**。

检查Flume Client端的配置。

步骤5 以**root**用户登录到告警定位参数中描述的Flume ClientIP所在主机。

步骤6 执行**cd** *Flume客户端安装目录*/**fusioninsight-flume-1.9.0/conf/**命令，进入Flume的配置目录。

步骤7 执行**cat properties.properties**命令，查看当前的Flume Client配置文件。

步骤8 根据Flume Agent的配置说明检查“properties.properties”的配置是否有误。

- 是，执行**步骤9**。

- 否，执行**步骤11**。

步骤9 修改“properties.properties”配置文件。

查看告警是否已清除。

步骤10 查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤12 在“服务”框中勾选待操作集群的“Flume”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤14 使用传输工具，收集Flume Client端“/var/log/Bigdata/flume-client”下的日志。

步骤15 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.163 ALM-24004 Flume 读取数据异常

告警解释

告警模块对Flume Source的状态进行监控，当Source读取不到数据的时长超过阈值时，系统即时上报告警。

默认阈值为0，表示不开启。用户可通过conf目录下的配置文件properties.properties修改阈值：修改对应source的“NoDatatime”参数。

当Source读取到数据，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24004	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
主机名	产生告警的主机名。
AgentId	产生告警的Agent id。
部件类型	产生告警的元素类型。
部件名	产生告警的元素名称。

对系统的影响

如果数据源有数据，Flume Source持续读取不到数据，数据采集会停止。

可能原因

- Flume Source故障，导致数据无法发送。
- 网络故障，导致数据无法发送。

处理步骤

检查Flume Source是否故障。

步骤1 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“type = spooldir”关键字确认Flume Source是否是spooldir类型。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 查看设置的spoolDir监控目录，是否所有的文件均已传输完毕。

- 是，处理完毕。
- 否，执行**步骤5**。

说明

spooDir的监控目录为用户自定义配置文件properties.properties中.spoolDir的参数值。若监控目录文件已传输完毕，则该监控目录下的所有文件以.COMPLETED后缀结尾。

步骤3 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“org.apache.flume.source.kafka.KafkaSource”关键字确认Flume Source是否是Kafka类型。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 查看Kafka Source配置的topic数据是否已经消费完毕。

- 是，处理完毕。
- 否，执行**步骤5**。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”。

步骤6 单击进入故障节点的Flume实例页面，查看监控指标“Source速度指标”，检查告警中的Source速度是否为0。

- 是，执行**步骤11**。
- 否，执行步骤**步骤7**。

检查Flume Source配置的IP所在节点与故障节点的网络状态。

步骤7 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“type = avro”关键字确认Flume Source是否是avro类型。

- 是，执行**步骤8**。
- 否，执行**步骤11**。

步骤8 以**root**用户登录故障节点所在主机，执行**ping** *Flume Source配置的IP地址*命令查看对端主机是否可以ping通。

- 是，执行**步骤11**。
- 否，执行**步骤9**。

步骤9 联系网络管理员恢复网络。

步骤10 等待一段时间后，在告警列表中，查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤11**。

收集故障信息。

步骤11 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤12 在“服务”框中勾选待操作集群的“Flume”。

步骤13 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤14 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.164 ALM-24005 Flume 传输数据异常

告警解释

告警模块对Flume Channel的容量状态进行监控，当Channel满的时长超过阈值，或Source向Channel放数据失败的次数超过阈值后，系统即时上报告警。

默认阈值为10，用户可通过conf目录下的配置文件properties.properties修改阈值：修改对应channel的“channelfullcount”参数。

当Flume Channel空间被释放，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
主机名	产生告警的主机名。
AgentId	产生告警的Agent id。
部件类型	产生告警的元素类型。
部件名	产生告警的元素名称。

对系统的影响

Flume Channel的磁盘空间使用量有继续增长的趋势，将会使数据导入到指定目的地的时间增长，当Flume Channel的磁盘空间使用量达到100%时会导致Flume Agent进程暂停工作。

可能原因

- Flume Sink故障，导致数据无法发送。
- 网络故障，导致数据无法发送。

处理步骤

检查Flume Sink是否故障。

步骤1 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“type = hdfs”关键字确认Flume Sink是否是HDFS类型。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 在FusionInsight Manager的告警列表中查看是否有“HDFS服务不可用”告警产生，服务列表中HDFS是否已停止。

- 是，如果有告警参考“ALM-14000 HDFS服务不可用”的处理步骤处理该故障；如果HDFS已停止，启动HDFS服务，执行**步骤7**。

- 否，执行**步骤7**。

步骤3 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“type = hbase”关键字确认Flume Sink是否是HBase类型。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 在FusionInsight Manager的告警列表中，查看是否有“HBase服务不可用”告警产生，服务列表中HBase是否已停止。

- 是，如果有告警参考“ALM-19000 HBase服务不可用”的处理步骤处理该故障，如果HBase已停止，启动HBase服务。执行**步骤7**。
- 否，执行**步骤7**。

步骤5 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“org.apache.flume.sink.kafka.KafkaSink”关键字确认Flume Sink是否是Kafka类型。

- 是，执行**步骤6**。
- 否，执行**步骤9**。

步骤6 在FusionInsight Manager的告警列表中，查看是否有“Kafka服务不可用”告警产生，服务列表中Kafka是否已停止。

- 是，如果有告警参考“ALM-38000 Kafka服务不可用”的处理步骤处理该故障；如果Kafka已停止，启动Kafka服务，执行**步骤7**。
- 否，执行**步骤7**。

步骤7 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”。

步骤8 单击进入故障节点的Flume实例页面，查看指标“Sink速度指标”，检查其速度是否为0。

- 是，执行**步骤13**。
- 否，执行步骤**步骤9**。

检查Flume Sink配置的IP所在节点与故障节点的网络状态。

步骤9 本地打开用户自定义配置文件properties.properties，搜索配置文件中是否有“type = avro”关键字确认Flume Sink是否是avro类型。

- 是，执行**10**。
- 否，执行**步骤13**。

步骤10 以**root**用户登录故障节点所在主机，执行**ping** *Flume Sink配置的IP地址*命令查看对端主机是否可以ping通。

- 是，执行**步骤13**。
- 否，执行**步骤11**。

步骤11 联系网络管理员恢复网络。

步骤12 等待一段时间后，在告警列表中，查看告警是否清除。

- 是，处理完毕。
- 否， 执行**步骤13**。

收集故障信息。

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”框中勾选待操作集群的“Flume”。

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.165 ALM-24006 Flume Server 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测Flume服务堆内存使用状态，当连续10次检测到Flume实例堆内存使用率超出阈值（最大内存的95%）时产生该告警，堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件

对系统的影响

堆内存溢出可能导致服务崩溃。

可能原因

该节点Flume实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Flume堆内存使用率超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > Agent > Flume堆内存使用率”，单击“确定”。

步骤3 查看Flume使用的堆内存是否已达到Flume设定的阈值（默认值为最大堆内存的95%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置”，选择“全部配置”，选择“Flume > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并单击“保存”，单击“确定”。

说明

出现此告警时，说明当前flume server设置的堆内存无法满足当前数据传输所需的堆内存，建议堆内存调整为： channel capacity * 最大单条数据大小*通道个数，但xmx参数值不能超过节点剩余内存。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Flume”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.166 ALM-24007 Flume Server 直接内存使用率超过阈值

告警解释

系统每60秒周期性检测Flume服务直接内存使用状态，当连续5次检测到Flume实例直接内存使用率超出阈值（最大内存的80%）时，产生该告警。当Flume直接内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

直接内存溢出可能导致服务崩溃。

可能原因

节点Flume实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Flume直接内存使用率超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > Agent > Flume直接内存使用率”，单击“确定”。

步骤3 查看Flume使用的直接内存是否已达到Flume设定的阈值（默认值为最大直接内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置”，选择“全部配置”，选择“Flume > 系统”。将“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值根据实际情况调大，并单击“保存”，单击“确定”。

说明

出现此告警时，说明当前flume server实例设置直接内存大小无法满足当前业务使用场景，建议调整“-XX:MaxDirectMemorySize”的值为当前直接内存使用量的两倍（或根据实际情况进行调整）。

步骤5 重新启动受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Flume”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.167 ALM-24008 Flume Server 非堆内存使用率超过阈值

告警解释

系统每60秒周期性检测Flume服务非堆内存使用状态，当连续5次检测到Flume实例非堆内存使用率超出阈值（最大内存的80%）时产生该告警，非堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

非堆内存溢出可能导致服务崩溃。

可能原因

该节点Flume实例非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Flume非堆内存使用率超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > Agent > Flume非堆内存使用率”，单击“确定”。

步骤3 查看Flume使用的非堆内存是否已达到Flume设定的阈值（默认值为最大非堆内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置”，选择“全部配置”，选择“Flume > 系统”。将“GC_OPTS”参数中“-XX:MaxPermSize”的值根据实际情况调大，并单击“保存”，单击“确定”。

说明

出现此告警时，说明当前flume server实例设置非堆内存大小无法满足当前业务使用场景，建议调整“-XX:MaxPermSize”的值为当前非堆内存使用量的两倍（或根据实际情况进行调整）。

步骤5 重启受影响的服务或实例观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Flume”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.168 ALM-24009 Flume Server 垃圾回收(GC)时间超过阈值

告警解释

系统每60秒周期性检测Flume进程的垃圾回收（GC）占用时间，当连续5次检测到Flume进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

导致Flume数据传输效率低下。

可能原因

该节点Flume实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Flume进程垃圾回收（GC）时间超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > Agent > Flume垃圾回收（GC）总时间”，单击“确定”。

步骤3 查看Flume每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置”，选择“全部配置”，选择“Flume > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并单击“保存”，单击“确定”。

说明

出现此告警时，说明当前flume server设置的堆内存无法满足当前数据传输所需的堆内存，建议堆内存调整为： channel capacity * 最大单条数据大小*通道个数，但xmx参数值不能超过节点剩余内存。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Flume”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.169 ALM-24010 Flume 证书文件非法或已损坏

告警解释

Flume每隔一个小时，检查当前Flume证书文件是否合法（证书是否存在，证书格式是否正确），如果证书文件非法或已损坏，产生该告警。证书文件恢复合法时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24010	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Flume证书文件已经非法或损坏，功能受限，Flume客户端将无法访问Flume服务端。

可能原因

Flume证书文件非法或损坏。

处理步骤

查看告警信息。

步骤1 登录FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-24010 Flume证书文件非法或已损坏 > 定位信息”。查看告警上报的实例的IP地址。

检查系统中证书文件是否有效，重新生成证书文件。

步骤2 以**root**用户登录告警所在节点主机，并执行**su - omm**切换用户。

步骤3 执行以下命令进入Flume服务证书目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/conf

步骤4 执行命令**ls -l**，查看“flume_sChat.crt”文件是否存在。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 执行命令**openssl x509 -in flume_sChat.crt -text -noout**，查看是否正常显示证书具体信息。

- 是，执行**步骤9**。
- 否，执行**步骤6**。

步骤6 执行以下命令进入Flume脚本目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/bin

步骤7 执行以下令重新生成证书，等待一个小时，观察此告警是否被清除。

sh geneJKS.sh -f *Flume角色服务端的自定义证书密码* **-g** *Flume角色客户端的自定义证书密码*

- 是，执行**步骤8**。
- 否，执行**步骤9**。

说明

Flume角色服务端、客户端的自定义证书密码需满足以下复杂度要求：

- 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
- 至少8位，最多64位。
- 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤8 查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤9**。
- 否，处理完毕。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的Flume。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.170 ALM-24011 Flume 证书文件即将过期

告警解释

Flume每隔一个小时，检查当前Flume证书文件是否即将过期，如果剩余有效期小于或等于30天，产生该告警。证书文件剩余有效期大于30天，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24011	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Flume证书文件即将失效，对系统目前运行无影响。

可能原因

Flume证书文件即将到期。

处理步骤

查看告警信息。

步骤1 登录FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-24011 Flume证书文件即将过期 > 定位信息”。查看告警上报的实例的IP地址。

检查系统中合法证书文件的有效期，重新生成证书文件。

步骤2 以**root**用户登录告警所在节点主机，并执行**su - omm**切换用户。

步骤3 执行以下命令进入Flume服务证书目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/conf

步骤4 执行以下命令查看Flume用户证书的生效时间与失效时间。

openssl x509 -noout -text -in flume_sChat.crt

步骤5 根据需要，选择业务空闲期，执行**步骤6**~**步骤7**更新证书。

步骤6 执行以下命令进入Flume脚本目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/bin

步骤7 执行命令重新生成证书，等待1小时，观察此告警是否被清除。

sh geneJKS.sh -f *Flume角色服务端的自定义证书密码* **-g** *Flume角色客户端的自定义证书密码*

- 是，执行**步骤9**。
- 否，执行**步骤8**。

  说明

  Flume角色服务端、客户端的自定义证书密码需满足以下复杂度要求：
  - 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
  - 至少8位，最多64位。
  - 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤8 使用**omm**用户在Flume实例产生告警的节点，重复执行**步骤6**~**步骤7**，等待1小时，观察此告警是否被清除。

- 是，执行**步骤9**。
- 否，执行**步骤10**。

步骤9 查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤10**。
- 否，处理完毕。

收集故障信息。

步骤10 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选待操作集群的Flume。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.171 ALM-24012 Flume 证书文件已过期

告警解释

Flume每隔一个小时，检查当前系统中的证书文件是否已过期。如果服务端证书已过期，产生该告警。服务的证书文件恢复到有效期内，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24012	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

Flume证书文件已过期，功能受限，Flume客户端将无法访问Flume服务端。

可能原因

Flume证书文件已过期。

处理步骤

查看告警信息。

步骤1 登录FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-24012 Flume证书文件已过期 > 定位信息”。查看告警上报的实例的IP地址。

检查系统中合法证书文件的有效期，重新生成证书文件。

步骤2 以**root**用户登录告警所在节点主机，并执行**su - omm**切换用户。

步骤3 执行以下命令进入Flume服务证书目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/conf

步骤4 执行以下命令查看HA用户证书的生效时间与失效时间，查看目前时间是否在有效期内。

openssl x509 -noout -text -in flume_sChat.crt

- 是，执行**步骤9**。
- 否，执行**步骤5**。

步骤5 执行以下命令进入Flume脚本目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/bin

步骤6 执行以下命令重新生成证书，等待1小时，观察此告警是否被清除。

sh geneJKS.sh -f *Flume角色服务端的自定义证书密码* **-g** *Flume角色客户端的自定义证书密码*

- 是，执行**步骤8**。
- 否，执行**步骤7**。

说明

Flume角色服务端、客户端的自定义证书密码需满足以下复杂度要求：

- 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
- 至少8位，最多64位。
- 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤7 使用**omm**用户在Flume实例产生告警的节点，重复执行**步骤5**~**步骤6**，等待1小时，观察此告警是否被清除。

- 是，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤9**。
- 否，处理完毕。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的Flume。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.172 ALM-24013 Flume MonitorServer 证书文件非法或已损坏

告警解释

MonitorServer每隔一个小时，检查当前MonitorServer证书文件是否合法（证书是否存在，证书格式是否正确），如果证书文件非法或已损坏，产生该告警。证书文件恢复合法，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24013	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

MonitorServer证书文件已经非法或损坏，功能受限，Flume客户端将无法访问Flume服务端。

可能原因

MonitorServer证书文件非法或损坏。

处理步骤

查看告警信息。

步骤1 登录FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-24013 MonitorServer证书文件非法或已损坏 > 定位信息”。查看告警上报的实例的IP地址。

检查系统中证书文件是否有效，重新生成证书文件。

步骤2 以**root**用户登录告警所在节点主机，并执行**su - omm**切换用户。

步骤3 执行以下命令进入MonitorServer证书目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/conf

步骤4 执行命令**ls -l**，查看**ms_sChat.crt**文件是否存在。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 执行命令**openssl x509 -in ms_sChat.crt -text -noout**，查看是否正常显示证书具体信息。

- 是，执行**步骤9**。
- 否，执行**步骤6**。

步骤6 执行以下命令进入Flume脚本目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/bin

步骤7 执行以下命令重新生成证书，等待一个小时，观察此告警是否被清除。

sh geneJKS.sh -m *服务端的自定义MonitorServer证书密码* **-n** *客户端的自定义MonitorServer证书密码*

- 是，执行**步骤8**。
- 否，执行**步骤9**。

  说明

  服务端、客户端的自定义MonitorServer证书密码需满足以下复杂度要求：
  - 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
  - 至少8位，最多64位。
  - 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤8 查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤9**。
- 否，处理完毕。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的MonitorServer。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.173 ALM-24014 Flume MonitorServer 证书文件即将过期

告警解释

MonitorServer每隔一个小时，检查当前MonitorServer证书文件是否即将过期，如果剩余有效期小于或等于30天，产生该告警。剩余有效期大于30天，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24014	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

MonitorServer证书文件即将失效，对系统目前运行无影响。

可能原因

MonitorServer证书文件即将到期。

处理步骤

查看告警信息。

步骤1 登录FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-24014 MonitorServer证书文件即将过期 > 定位信息”。查看告警上报的实例的IP地址。

检查系统中合法证书文件的有效期，重新生成证书文件。

步骤2 以**root**用户登录告警所在节点主机，并执行**su - omm**切换用户。

步骤3 执行命令进入MonitorServer证书目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/conf

步骤4 执行命令查看用户证书的生效时间与失效时间。

openssl x509 -noout -text -in ms_sChat.crt

步骤5 根据需要，选择业务空闲期，执行**步骤6**~**步骤7**更新证书。

步骤6 执行以下命令进入Flume脚本目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/bin

步骤7 执行以下命令重新生成证书，等待1小时，观察此告警是否被清除。

sh geneJKS.sh -m *服务端的自定义MonitorServer证书密码* **-n** *客户端的自定义MonitorServer证书密码*

- 是，执行**步骤9**。
- 否，执行**步骤8**。

  说明

  服务端、客户端的自定义MonitorServer证书密码需满足以下复杂度要求：
  - 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
  - 至少8位，最多64位。
  - 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤8 使用**omm**用户在Flume实例产生告警的节点，重复执行**步骤6**~**步骤7**，等待1小时，观察此告警是否被清除。

- 是，执行**步骤9**。
- 否，执行**步骤10**。

步骤9 查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤10**。
- 否，处理完毕。

收集故障信息。

步骤10 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选待操作集群的MonitorServer。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.174 ALM-24015 Flume MonitorServer 证书文件已过期

告警解释

MonitorServer每隔一个小时健康检查时，检查当前系统中的证书文件是否已过期。如果服务端证书已过期，产生该告警。服务端证书恢复的有效期内，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
24015	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

MonitorServer证书文件已过期，功能受限，Flume客户端将无法访问Flume服务端。

可能原因

MonitorServer证书文件已过期。

处理步骤

查看告警信息。

步骤1 登录FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-24015 MonitorServer证书文件已过期 > 定位信息”。查看告警上报的实例的IP地址。

检查系统中合法证书文件的有效期，重新生成证书文件。

步骤2 以**root**用户登录告警所在节点主机，并执行**su - omm**切换用户。

步骤3 执行以下命令进入MonitorServer证书目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/conf

步骤4 执行以下命令查看用户证书的生效时间与失效时间，查看目前时间是否在有效期内。

openssl x509 -noout -text -in ms_sChat.crt

- 是，执行**步骤9**。
- 否，执行**步骤5**。

步骤5 执行以下命令进入Flume脚本目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_*/install/FusionInsight-Flume-*/flume/bin

步骤6 执行以下命令重新生成证书，等待1小时，观察此告警是否被清除。

sh geneJKS.sh -m *服务端的自定义MonitorServer证书密码* **-n** *客户端的自定义MonitorServer证书密码*

- 是，执行**步骤8**。
- 否，执行**步骤7**。

说明

服务端、客户端的自定义MonitorServer证书密码需满足以下复杂度要求：

- 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
- 至少8位，最多64位。
- 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤7 使用**omm**用户在Flume实例产生告警的节点，重复执行**步骤5**~**步骤6**，等待1小时，观察此告警是否被清除。

- 是，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 查看系统在定时检查时是否会再次产生此告警。

- 是，执行**步骤9**。
- 否，处理完毕。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的MonitorServer。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.175 ALM-25000 LdapServer 服务不可用

告警解释

系统按30秒周期性检测LdapServer的服务状态，当检测到两个LdapServer服务均异常时产生该告警。

当检测到一个或两个LdapServer服务恢复时告警恢复。

告警属性

告警ID	告警级别	是否自动清除
25000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

告警发生时，不能对集群中的KrbServer和LdapServer用户进行任何操作。例如，无法在FusionInsight Manager页面添加、删除或修改任何用户、用户组或角色，也无法修改用户密码。集群中原有的用户验证不受影响。

可能原因

- LdapServer服务所在节点故障。
- LdapServer进程故障。

处理步骤

检查LdapServer服务的两个SlapdServer实例所在节点是否故障。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 实例”。进入LdapServer实例页面获取两个SlapdServer实例所在节点的主机名。

步骤2 选择“运维 > 告警 > 告警”，在告警列表中查看是否有“节点故障”告警产生。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 查看告警信息里的主机名是否和**步骤1**主机名一致。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 按“ALM-12006 节点故障”提供的步骤处理该告警。

步骤5 在告警列表中查看“LdapServer服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤10**。

检查LdapServer进程是否正常。

步骤6 选择“运维 > 告警 > 告警”，在告警列表中查看是否有“进程故障”告警产生。

- 是，执行**步骤7**。
- 否，执行**步骤10**。

步骤7 查看告警信息中的服务名和主机名是否和LdapServer服务名和主机名一致。

- 是，执行**步骤8**。
- 否，执行**步骤10**。

步骤8 按“ALM-12007 进程故障”提供的步骤处理该告警。

步骤9 在告警列表中查看“LdapServer服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤10**。

收集故障信息。

步骤10 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”中勾选待操作集群的“LdapServer”。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.176 ALM-25004 LdapServer 数据同步异常

告警解释

系统按30秒周期性检测LdapServer数据，如果连续12次检测，Manager的主备LdapServer的数据内容都不一致，产生该告警，当两者的数据一致时，对应告警恢复。

系统按30秒周期性检测LdapServer数据，如果连续12次检测，集群中的LdapServer的数据与Manager的LdapServer数据都不一致，产生该告警，当两者的数据一致时，对应告警恢复。

告警属性

告警ID	告警级别	是否自动清除
25004	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机节点信息。

对系统的影响

LdapServer数据不一致时，有可能是Manager上的LdapServer数据损坏，也有可能是集群上的LdapServer数据损坏，此时数据损坏的LdapServer进程将无法对外提供服务，影响Manager和集群的认证功能。

可能原因

- LdapServer进程所在的节点网络故障。
- LdapServer进程异常。
- OS重启导致的LdapServer数据损坏。

处理步骤

检查LdapServer所在的节点网络是否故障。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。记录该告警定位信息中的“主机名”的IP地址为IP1（若出现多个告警，则分别记录其中的IP地址为IP1、IP2、IP3等）。

步骤2 联系运维人员，登录IP1节点，在这个节点上使用**ping**命令检查该节点与主OMS节点的管理平面IP是否可达。

- 是，执行**步骤4**。
- 否，执行**步骤3**。

步骤3 联系网络管理员恢复网络，然后查看“LdapServer数据同步异常”告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查LdapServer进程是否正常。

步骤4 在FusionInsight Manager的“告警”页面，查看是否有LdapServer的“OLdap资源异常”告警产生。

- 是，执行**步骤5**。
- 否，执行**步骤7**。

步骤5 按照“ALM-12004 OLdap资源异常”提供的步骤处理该告警。

步骤6 在告警列表中查看“LdapServer数据同步异常”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

步骤7 在FusionInsight Manager的“告警”页面，查看是否有LdapServer的“进程故障”告警产生。

- 是，执行**步骤8**。
- 否，执行**步骤10**。

步骤8 按照“ALM-12007 进程故障”提供的步骤处理该告警。

步骤9 在告警列表中查看“LdapServer数据同步异常”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤10**。

检查是否存在因为OS重启导致LdapServer数据损坏。

步骤10 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。记录该告警定位信息中的“主机名”的IP地址为IP1（若出现多个告警，则分别记录其中的IP地址为IP1，IP2，IP3等）。选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 配置”，记录LdapServer的端口号PORT（若告警定位信息中的IP地址为备管理节点IP地址，选择“系统 > OMS > oldap > 修改配置”，记录LdapServer服务侦听端口号）。

步骤11 以**omm**用户登录IP1节点。

步骤12 执行以下命令，观察查询出来的内容是否提示有error错误信息。

ldapsearch -H ldaps://*IP1*:*PORT* -LLL -x -D cn=root,dc=hadoop,dc=com -W -b ou=Peoples,dc=hadoop,dc=com

执行命令后需输入**LDAP**管理员密码，请联系系统管理员获取。

- 是，执行**步骤13**。
- 否，执行**步骤15**。

步骤13 使用告警出现日期之前的备份文件进行LdapServer恢复和OMS恢复。

说明

必须使用同一时间点的OMS和LdapServer备份数据进行恢复，否则可能造成业务和操作失败。当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作，否则会丢失从备份时刻到恢复时刻之间的Manager数据。

步骤14 在告警列表中查看“LdapServer数据同步异常”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤15**。

收集故障信息。

步骤15 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤16 在“服务”中勾选待操作集群的“LdapServer”和“OmsLdapServer”。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.177 ALM-25005 Nscd 服务异常

告警解释

系统每60秒周期性检测nscd服务的状态，如果连续4次（3分钟）查询不到nscd进程或者无法获取ldapserver中的用户时，产生该告警。

当进程恢复且可以获取ldapserver中的用户时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
25005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。

参数名称	参数含义
服务名	产生告警的服务名称。
主机名	产生告警的主机节点信息。

对系统的影响

nscd服务不可用时，可能会影响该节点从LdapServer上同步数据，此时，使用**id**命令可能会获取不到ldap中的数据，影响上层业务。

可能原因

- nscd服务未启动。
- 网络故障，无法访问ldap服务器。
- Name Service服务异常。
- OS执行命令慢导致无法查询用户。

处理步骤

检查nscd服务是否启动。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。记录该告警定位信息中的“主机名”的IP地址为IP1（若出现多个告警，则分别记录其中的IP地址为IP1、IP2、IP3等）。

步骤2 联系运维人员，以**root**用户登录IP1节点，在该节点上执行**ps -ef | grep nscd**命令，查看是否有/usr/sbin/nscd进程启动。

- 是，执行**步骤5**。
- 否，执行**步骤3**。

步骤3 以**root**用户执行**service nscd restart**命令，重启nscd服务，执行**ps -ef | grep nscd**命令，查看服务是否启动。

- 是，执行**步骤4**。
- 否，执行**步骤15**。

步骤4 5分钟后，以**root**用户再次执行命令，查看服务是否存在。

- 是，执行**步骤11**。
- 否，执行**步骤15**。

检查网络是否故障，无法访问ldap服务器。

步骤5 用**root**用户登录故障节点，在这个节点上使用**ping**命令检查该节点与LdapServer节点的网络是否畅通。

- 是，执行**步骤6**。
- 否，请联系网络管理员，解决网络故障。

检查Name Service服务是否异常。

步骤6 用**root**用户登录故障节点，执行**cat /etc/nsswitch.conf**命令，查看NameService配置中的“passwd”、“group”、“services”、“netgroup”、“aliases”五项配置是否正确。

正确配置请参照：“passwd: compat ldap”、“group: compat ldap”、“services: files ldap”、“netgroup: files ldap”、“aliases: files ldap”。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 用**root**用户登录故障节点，执行**cat /etc/nscd.conf**命令，查看配置文件中“enable-cache passwd”、“positive-time-to-live passwd”、“enable-cache group”、“positive-time-to-live group”四项配置是否正确。

正确配置请参照：“enable-cache passwd yes ”、“positive-time-to-live passwd 600”、“enable-cache group yes”、“positive-time-to-live group 3600”。

- 是，执行**步骤8**。
- 否，执行**步骤10**。

步骤8 用**root**用户执行**/usr/sbin/nscd -i group**和**/usr/sbin/nscd -i passwd**命令，等待2分钟，继续执行**id admin**和**id backup/manager**命令，查看是否能查询到结果。

- 是，执行**步骤11**。
- 否，执行**步骤15**。

步骤9 以**root**用户执行**vi /etc/nsswitch.conf**命令，将**步骤6**中的五项配置项改成正确配置，保存后执行**service nscd restart**命令重启nscd服务，等待2分钟，执行**id admin**和**id backup/manager**命令，查看是否能查询到结果。

- 是，执行**步骤11**。
- 否，执行**步骤15**。

步骤10 以**root**用户执行**vi /etc/nscd.conf**命令，将**步骤7**中的四项配置项改成正确配置，保存后执行**service nscd restart**命令重启nscd服务，等待2分钟，执行**id admin**和**id backup/manager**命令，查看是否能查询到结果。

- 是，执行**步骤11**。
- 否，执行**步骤15**。

步骤11 登录FusionInsight Manager界面，等待5分钟，然后查看“Nscd服务异常”告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤12**。

检查操作系统执行命令是否卡顿。

步骤12 用**root**用户登录故障节点，执行命令**id admin**，观察命令返回结果时长，观察执行命令是否缓慢（超过3s即可认为执行命令慢）。

- 是，执行**步骤13**。
- 否，执行**步骤15**。

步骤13 执行命令**cat /var/log/messages**，查看nscd是否频繁重启或者存在Can't contact LDAP server的异常信息。

nscd异常信息样例：

```
Feb 11 11:44:42 10-120-205-33 nscd: nss_ldap: failed to bind to LDAP server ldaps://10.120.205.55:21780: Can't contact LDAP server
Feb 11 11:44:43 10-120-205-33 ntpq: nss_ldap: failed to bind to LDAP server ldaps://10.120.205.55:21780: Can't contact LDAP server
Feb 11 11:44:44 10-120-205-33 ntpq: nss_ldap: failed to bind to LDAP server ldaps://10.120.205.92:21780: Can't contact LDAP server
```

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 执行命令**vi $BIGDATA_HOME/tmp/random_ldap_ip_order**，修改末尾数字，若原本为奇数则改为偶数，若原本为偶数则修改为奇数；

执行命令**vi /etc/ldap.conf**进入编辑模式，按“Insert”键开始编辑，然后将URI配置项的前两个IP进行调换。

修改完成后按“Esc”键退出编辑模式，并输入**:wq**保存退出。

执行命令**service nscd restart**，重启nscd服务，等待5分钟，再次执行**id admin**命令，观察返回结果时长，观察执行命令是否缓慢。

- 是，执行**步骤15**。
- 否，登录其他故障节点执行**步骤12**至**步骤14**；排查“/etc/ldap.conf”修改前URI中第一个ldapserver节点，是否故障，例如业务IP不可达、网络延时过长或者部署其他异常的软件。

收集故障信息。

步骤15 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤16 在“服务”中勾选待操作集群的“LdapClient”。

步骤17 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤18 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.178 ALM-25006 Sssd 服务异常

告警解释

系统每60秒周期性检测sssd服务的状态，如果连续4次（3分钟）查询不到sssd进程或者无法获取LdapServer中的用户时，产生该告警。

当进程恢复且可以获取LdapServer中的用户时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
25006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
主机名	产生告警的主机节点信息。

对系统的影响

sssd服务不可用时，可能会影响该节点从LdapServer上同步数据，此时，使用**id**命令可能会获取不到ldap中的数据，影响上层业务。

可能原因

- sssd服务未启动或启动错误。
- 网络故障，无法访问Ldap服务器。
- Name Service服务异常。
- OS执行命令慢导致无法查询用户。

处理步骤

检查sssd服务是否启动或启动错误。

步骤1 在FusionInsight Manager界面，选择“运维 > 告警 > 告警”。记录该告警定位信息中的“主机名”的IP地址为IP1（若出现多个告警，则分别记录其中的IP地址为IP1、IP2、IP3等）。

步骤2 联系运维人员，以**root**用户登录IP1节点，在该节点执行**ps -ef | grep sssd**命令，查看是否有/usr/sbin/sssd进程启动。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 查看**步骤2**中查询的sssd进程是否有三个子进程。

- 是，执行**步骤5**。
- 否，执行**步骤4**。

步骤4 以**root**用户执行**service sssd restart**命令重启sssd服务，执行**ps -ef | grep sssd**命令，查看sssd进程是否正常。

正常状态为：存在/usr/sbin/sssd进程和三个子进程/usr/libexec/sssd/sssd_be、/usr/libexec/sssd/sssd_nss、/usr/libexec/sssd/sssd_pam。

- 是，执行**步骤9**。
- 否，执行**步骤13**。

检查网络是否故障，无法访问ldap服务器。

步骤5 用**root**用户登录故障节点，在这个节点上使用**ping**命令检查该节点与LdapServer节点的网络是否畅通。

- 是，执行**步骤6**。
- 否，请联系网络管理员，解决网络故障。

检查Name Service服务是否异常。

步骤6 用**root**用户登录故障节点，执行命令**cat /etc/nsswitch.conf**，查看NameService配置中的“passwd”、“group”两项配置是否正确。

正确配置请参照：“passwd: files sss”、“group: files sss”。

- 是，执行**步骤7**。
- 否，执行**步骤8**。

步骤7 用**root**用户执行**/usr/sbin/sss_cache -G**和**/usr/sbin/sss_cache -U**命令，等待2分钟，执行**id admin**和**id backup/manager**命令，查看是否能查询到结果。

- 是，执行**步骤9**。
- 否，执行**步骤13**。

步骤8 以root用户执行**vi /etc/nsswitch.conf**命令，将**步骤6**中的两项配置项改成正确配置，保存后执行**service sssd restart**命令重启sssd服务，等待2分钟，执行**id admin**和**id backup/manager**命令，查看是否能查询到结果。

- 是，执行**步骤9**。
- 否，执行**步骤13**。

步骤9 登录FusionInsight Manager界面，等待5分钟，然后查看“Sssd服务异常”告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤10**。

检查操作系统执行命令是否卡顿。

步骤10 用**root**用户登录故障节点，执行命令**id admin**，观察命令返回结果时长，观察执行命令是否缓慢（超过3s即可认为执行命令慢）。

- 是，执行**步骤11**。
- 否，执行**步骤13**。

步骤11 执行命令**cat /var/log/messages**，查看sssd是否频繁重启或者存在Can't contact LDAP server的异常信息。

sssd重启样例

```
Feb  7 11:38:16 10-132-190-105 sssd[pam]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd[nss]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd[nss]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd[be[default]]: Shutting down
Feb  7 11:38:16 10-132-190-105 sssd: Starting up
Feb  7 11:38:16 10-132-190-105 sssd[be[default]]: Starting up
```

```
Feb  7 11:38:16 10-132-190-105 sssd[nss]: Starting up
Feb  7 11:38:16 10-132-190-105 sssd[pam]: Starting up
```

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 执行命令**vi $BIGDATA_HOME/tmp/random_ldap_ip_order**，修改末尾数字，若原本为奇数则改为偶数，若原本为偶数则修改为奇数。

执行命令**vi /etc/sssd/sssd.conf**，将ldap_uri配置项的前两个IP进行颠倒，保存退出。

执行命令**ps -ef | grep sssd**查询sssd进程id，并将其kill掉，执**行/usr/sbin/sssd -D -f**，重启sssd服务，等待5分钟，再次执行**id admin**命令。

观察返回结果时长，观察执行命令是否缓慢。

- 是，执行**步骤13**。
- 否，登录其他故障节点执行**步骤10**至**步骤12**；收集日志，并排查“/etc/sssd/sssd.conf”修改前ldap_uri中第一个ldapserver节点是否故障，例如业务IP不可达、网络延时过长或者部署其他异常的软件。

收集故障信息。

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选待操作集群的“LdapClient”。

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.179 ALM-25500 KrbServer 服务不可用

告警解释

系统按30秒周期性检测组件KrbServer的服务状态。当检测到组件KrbServer服务异常时产生该告警。

当检测到组件KrbServer服务恢复时告警恢复。

告警属性

告警ID	告警级别	是否自动清除
25500	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

告警发生时，不能对集群中的组件KrbServer进行任何操作。其它组件的KrbServer认证将受影响。集群中依赖KrbServer的组件运行状态将为故障。

可能原因

- 组件KrbServer服务所在节点故障。
- OLdap服务不可用。

处理步骤

检查组件KrbServer服务所在节点是否故障。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > KrbServer > 实例”。进入KrbServer实例页面查看KrbServer服务所在节点的主机名。

步骤2 在FusionInsight Manager的“告警”页面，查看是否有“节点故障”告警产生。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 查看告警信息里的主机名是否和**步骤1**主机名一致。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 按“ALM-12006 节点故障”提供的步骤处理该告警。

步骤5 在告警列表中查看“KrbServer服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查OLdap服务是否不可用。

步骤6 在FusionInsight Manager的“告警”页面，查看是否有“OLdap资源异常”告警产生。

- 是，执行**步骤7**。
- 否，执行**步骤9**。

步骤7 按“ALM-12004 OLdap资源异常”提供的步骤处理该告警。

步骤8 在告警列表中查看“KrbServer服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“KrbServer”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.180 ALM-26051 Storm 服务不可用

告警解释

系统按照30秒的周期检测Storm服务是否可用，当集群全部的Nimbus节点异常时，Storm服务不可用，系统产生此告警。

当Storm服务恢复正常，告警自动清除。

告警属性

告警ID	告警级别	是否自动清除
26051	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

集群无法对外提供Storm服务，用户无法执行新的Storm任务。

可能原因

- Kerberos集群故障。
- ZooKeeper集群故障或假死。
- Storm集群中主备Nimbus状态异常。

处理步骤

检查Kerberos集群状态（普通模式集群跳过此步骤）

步骤1 在FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务”。

步骤2 查看Kerberos服务的运行状态是否为“良好”。

- 是，执行**步骤5**。
- 否，执行**步骤3**。

步骤3 参考“ALM-25500 KrbServer服务不可用”的相关维护信息进行操作。

步骤4 查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

检查ZooKeeper集群状态

步骤5 查看ZooKeeper服务的运行状态是否为“良好”。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 如果Zookeeper服务停止运行，则启动服务，否则参考“ALM-13000 ZooKeeper服务不可用”的相关维护信息进行操作。

步骤7 查看告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

检查主备Nimbus状态

步骤8 选择“集群 > *待操作集群的名称* > 服务 > Storm > Nimbus”，进入Nimbus实例页面。

步骤9 查看“角色”中是否存在且仅存在一个状态为主的Nimbus节点。

- 是，执行**步骤13**。
- 否，执行**步骤10**。

步骤10 勾选两个Nimbus角色实例，选择“更多 > 重启实例”，查看是否重启成功。

- 是，执行**步骤11**。
- 否，执行**步骤13**。

步骤11 重新登录FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务 > Storm > Nimbus”，查看运行状态是否为“良好”。

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 等待30秒，查看告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

收集故障信息

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选待操作集群的如下节点信息。

- KrbServer

  说明

  普通模式不需要下载KrbServer日志。

- ZooKeeper
- Storm

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.181 ALM-26052 Storm 服务可用 Supervisor 数量小于阈值

告警解释

系统每60秒周期性检测Supervisor数量，并把实际Supervisor数量和阈值相比较。当检测到Supervisor数量低于阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称*”修改阈值。

当Supervisor数量大于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
26052	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

集群已经存在的任务无法运行；集群可接收新的Storm任务，但是无法运行。

可能原因

集群中Supervisor处于异常状态。

处理步骤

检查Supervisor状态

步骤1 选择“集群 > *待操作集群的名称* > 服务 > Storm > Supervisor”，进入Storm服务管理页面。

步骤2 查看“角色”中是否存在状态为故障或者是正在恢复的Supervisor实例。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 勾选状态为故障或者正在恢复的Supervisor角色实例，选择“更多 > 重启实例”，查看是否重启成功。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 等待30秒，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

说明

Supervisor重启过程中，业务会出现中断，待Supervisor重启成功后业务恢复。

收集故障信息

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选待操作集群的“Storm”和“ZooKeeper”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.182 ALM-26053 Storm Slot 使用率超过阈值

告警解释

系统每60秒周期性检测Slot使用率，并把实际Slot使用率和阈值相比较。当检测到Slot使用率高于阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置”修改阈值。

当Slot使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
26053	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

用户无法执行新的Storm任务。

可能原因

- 集群中Supervisor处于异常状态。
- 集群中Supervisor的状态正常，但是处理能力不足。

处理步骤

检查Supervisor状态

步骤1 选择“集群 > *待操作集群的名称* > 服务 > Storm > 实例”，进入Storm实例管理页面。

步骤2 查看是否存在状态为“故障”或者是“正在恢复”的Supervisor实例。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 勾选状态为“故障”或者“正在恢复”的Supervisor角色实例，选择“更多 > 重启实例”，查看是否重启成功。

- 是，执行**步骤4**。
- 否，执行**步骤10**。

步骤4 等待一段时间，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

增加Supervisor Slot数量配置。

步骤5 登录FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务 > Storm > 配置 > 全部配置”。

步骤6 适当增加每个Supervisor角色“supervisor.slots.ports”参数中的端口号数量，并重启实例。

步骤7 等待一段时间，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤8**。

步骤8 对Supervisor进行扩容。

步骤9 等待一段时间，检查该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤10**。

> **说明**
>
> Supervisor重启过程中，业务会出现中断，待Supervisor重启成功后业务恢复。

收集故障信息。

步骤10 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤11 在“服务”勾选待操作集群的“Storm”和“ZooKeeper”。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.183 ALM-26054 Nimbus 堆内存使用率超过阈值

告警解释

系统每30秒周期性检测Storm Nimbus堆内存使用率，并把实际的Storm Nimbus堆内存使用率和阈值相比较。当连续5次检测到Storm Nimbus堆内存使用率超出阈值（默认值为80%）时产生该告警。

用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Storm > Nimbus”修改阈值。

当Storm Nimbus堆内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
26054	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Storm Nimbus堆内存使用率过高时可能造成频繁GC，甚至造成内存溢出，进而影响Storm任务提交。

可能原因

该节点Storm Nimbus实例堆内存使用量过大，或分配的堆内存不合理，导致使用量超过阈值。

处理步骤

检查堆内存使用量。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 >Storm Nimbus堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Storm > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > Nimbus > Nimbus堆内存使用率”。单击“确定”。

步骤3 查看Nimbus使用的堆内存是否已达到Nimbus设定的阈值（默认值为最大堆内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务> Storm > 配置 > 全部配置 > Nimbus > 系统”。将“NIMBUS_GC_OPTS”参数中“-Xmx”的值根据实际情况进行调整，然后单击“保存”，单击“确定”。

说明

- 建议“-Xms”和“-Xmx”设置成相同的值，避免JVM动态调整堆内存大小时影响性能。
- 当Storm集群规模越大，Worker数量越多时，可以适当调大Nimbus的GC_OPTS参数，配置建议如下：Worker数量为20个时，“-Xmx”设置为不小于1G；Worker超过100个时，“-Xmx”设置为不小于5G，以此类推。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- NodeAgent
- Storm

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.184 ALM-27001 DBService 服务不可用

告警解释

告警模块按30秒周期检测DBService服务状态。当DBService服务不可用时产生该告警。

DBService服务恢复时，告警清除。

告警属性

告警ID	告警级别	是否自动清除
27001	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

数据库服务不可用，无法对上层服务提供数据入库、查询等功能，使部分服务异常。

可能原因

- 浮动IP不存在。
- 没有主DBServer实例。
- 主备DBServer进程都异常。

处理步骤

检查集群环境中是否存在浮动IP。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > DBService > 实例”。

步骤2 查看是否有主实例存在。

- 是，执行**步骤3**。

- 否，执行**步骤9**。

步骤3 选择主DBServer实例，记录IP地址。

步骤4 以**root**用户登录上述IP所在主机，执行**ifconfig**命令查看DBService的浮动IP在该节点是否存在。

- 是，执行**步骤5**。
- 否，执行**步骤9**。

步骤5 执行**ping** *浮动IP地址*命令检查DBService的浮动IP的状态，是否能ping通。

- 是，执行**步骤6**。
- 否，执行**步骤9**。

步骤6 以**root**用户登录DBService浮动IP所在主机，执行以下命令删除浮动IP地址。

ifconfig *interface* **down**

步骤7 在FusionInsight Manager首页，选择“ 集群 > *待操作集群的名称* > 服务 > DBService > 更多 > 重启服务”重启DBService服务，检查是否启动成功。

- 是，执行**步骤8**。
- 否，执行**步骤9**。

步骤8 等待约两分钟，查看告警列表中的DBService服务不可用告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤14**。

检查主DBServer实例状态。

步骤9 选择角色状态异常的DBServer实例，记录IP地址。

步骤10 在“告警”页面，查看是否有上述IP所在主机DBServer实例“进程故障”告警产生。

- 是，执行**步骤11**。
- 否，执行**步骤19**。

步骤11 按“ALM-12007 进程故障”提供的步骤处理该告警。

步骤12 等待5分钟，查看告警列表中的DBService服务不可用告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤19**。

检查主备DBServer数据库进程状态。

步骤13 以**root**用户登录DBService浮动IP所在主机，执行**su - omm**命令切换至**omm**用户。

步骤14 执行**cd ${DBSERVER_HOME}**命令进入DBService服务的安装目录。

步骤15 执行**sh sbin/status-dbserver.sh**命令查看DBService的主备HA进程状态，状态是否查询成功。

```
HAMode
double

NodeName            HostName        HAVersion          StartTime           HAActive
HAAllResOK      HARunPhase
10_5_89_12          host01          V100R001C01        2019-06-13 21:33:09     active
normal          Actived
10_5_89_66          host03          V100R001C01        2019-06-13 21:33:09     standby
```

```
normal          Deactived

NodeName        ResName         ResStatus       ResHAStatus     ResType
10_5_89_12      floatip         Normal          Normal          Single_active
10_5_89_12      gaussDB         Active_normal   Normal          Active_standby
10_5_89_66      floatip         Stopped         Normal          Single_active
10_5_89_66      gaussDB         Standby_normal  Normal          Active_standby
```

- 是，执行**步骤16**。
- 否，执行**步骤19**。

步骤16 查看主备HA进程是否都处于abnormal状态。

- 是，执行**步骤17**。
- 否，执行**步骤19**。

步骤17 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > DBService > 更多 > 重启服务”重启DBService服务，查看界面是否提示重启成功。

- 是，执行**步骤18**。
- 否，执行**步骤19**。

步骤18 等待约两分钟，查看告警列表中的DBService服务不可用告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤19**。

收集故障信息。

步骤19 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤20 在“服务”中勾选待操作集群的“DBService”和“NodeAgent”。

步骤21 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤22 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.185 ALM-27003 DBService 主备节点间心跳中断

告警解释

DBService主节点或备节点超过7秒未收到对端的心跳消息后，系统产生告警。

当心跳恢复后，该告警恢复。

告警属性

告警ID	告警级别	是否自动清除
27003	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Local DBService HA Name	本地DBService HA名称。
Peer DBService HA Name	对端DBService HA名称。

对系统的影响

DBService主备间心跳中断时只有一个节点提供服务，一旦该节点故障，再无法切换到备节点，就会服务不可用。

可能原因

主备DBService节点间链路异常。

处理步骤

检查主备DBService服务器间的网络是否正常。

步骤1 在FusionInsight Manager页面，在告警列表中，单击此告警所在行的，查看该告警的DBService备服务器地址。

步骤2 以**root**用户登录主DBService服务器。

步骤3 执行**ping** *备DBService心跳IP地址*命令检查备DBService服务器是否可达。

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 联系网络管理员查看是否为网络故障。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 修复网络故障，查看告警列表中，该告警是否已清除。

- 是，处理完毕。

- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的如下节点信息。

- DBService
- Controller
- NodeAgent

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.186 ALM-27004 DBService 主备数据不同步

告警解释

DBService主备数据不同步，每10秒检查一次主备数据同步状态，如果连续6次查不到同步状态，或者同步状态不正常，产生告警。

当同步状态正常，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
27004	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

参数名称	参数含义
Local DBService HA Name	本地DBService HA名称。
Peer DBService HA Name	对端DBService HA名称。
SYNC_PERCENT	同步百分比。

对系统的影响

主备DBServer数据不同步，如果此时主实例异常，则会出现数据丢失或者数据异常的情况。

可能原因

- 主备节点网络不稳定。
- 备DBService异常。
- 备节点磁盘空间满。

处理步骤

检查主备节点网络是否正常。

步骤1 在FusionInsight Manager页面，选择“集群 > 服务 > DBService > 实例”，查看备DBServer实例的业务IP地址。

步骤2 以**root**用户登录主DBService节点。

步骤3 执行**ping** *备DBService心跳IP地址*命令检查备DBService节点是否可达。

- 是，执行**步骤6**。
- 否，执行**步骤4**。

步骤4 联系网络管理员查看是否为网络故障。

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 修复网络故障，查看告警列表中，该告警是否已清除。

- 是，处理完毕。
- 否，执行**步骤6**。

检查备DBService状态是否正常

步骤6 以**root**用户登录备DBService节点。

步骤7 执行**su - omm**命令切换到**omm**用户。

步骤8 进入“${DBSERVER_HOME}/sbin”目录，然后执行命令 **./status-dbserver.sh** 检查备DBService的gaussDB资源状态是否正常，查看回显中，“ResName”为“gaussDB”的一行，是否显示如下信息：

例如：

```
10_10_10_231 gaussDB Standby_normal Normal Active_standby
```

- 是，执行**步骤9**。
- 否，执行**步骤16**。

检查备节点磁盘是否已满。

步骤9 以**root**用户登录备DBService节点。

步骤10 执行命令**su - omm**切换到**omm**用户。

步骤11 进入“${DBSERVER_HOME}”目录，执行以下命令获取DBservice的数据目录。

cd ${DBSERVER_HOME}

source .dbservice_profile

echo ${DBSERVICE_DATA_DIR}

步骤12 执行**df -h**命令，查看系统磁盘分区的使用信息。

步骤13 查看DBservice数据目录空间是否已满。

- 是，执行**步骤14**。
- 否，执行**步骤16**。

步骤14 对节点磁盘进行扩容。

步骤15 磁盘扩容后，等待2分钟检查告警是否清除。

- 是，操作结束。
- 否，执行**步骤16**。

收集故障信息。

步骤16 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤17 在“服务”中勾选待操作集群的“DBService”，单击“确定”。

步骤18 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤19 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.187 ALM-27005 数据库连接数使用率超过阈值

告警解释

系统每30秒周期性检查DBServer节点的数据库连接数使用率，并把实际数据库连接数使用率和阈值相比较，当数据库连接数的使用率连续5次（可配置，默认值为5）超过

设定阈值时，系统将产生此告警，数据库连接数使用率的阈值设为90%（可配置，默认值为90%）。

平滑次数可配置，当平滑次数为1，数据库连接数使用率小于或等于阈值时，该告警恢复；当平滑次数大于1，数据库连接数使用率小于或等于阈值的90%时，该告警恢复。

告警属性

告警ID	告警级别	是否自动清除
27005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

可能导致上层服务无法连接DBService的数据库，影响正常业务。

可能原因

- 数据库连接数使用过多。
- 数据库连接数最大值设置不合理。
- 告警阈值配置或者平滑次数配置不合理。

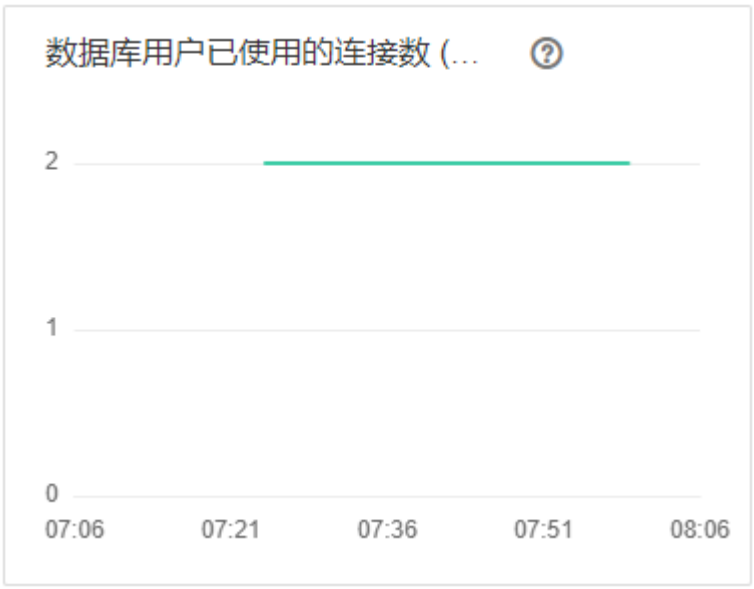
处理步骤

检查数据连接数是否使用过多

步骤1 在FusionInsight Manager主页，单击左侧服务列表的DBService服务，进入DBService监控页面。

步骤2 观察数据库用户已使用的连接数图表，如图10-39所示，用户根据业务场景评估，适当降低数据库用户连接数的使用。

图 10-39 数据库用户已使用的连接数图表



步骤3 等待2分钟查看告警是否自动恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

检查数据库连接数最大值设置是否合理

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > DBService > 配置 > 全部配置”，根据实际业务需求，将数据库连接数的最大值适当增加，如**图 10-40**所示。修改后单击“保存”，在弹出的“保存配置”页面中单击“确定”。

图 10-40 设置数据库连接数最大值

步骤5 完成数据库连接数最大值修改后，需要重启DBService服务（不要重启其上层服务）。

操作步骤：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > DBService > 更多 > 重启服务”，输入当前登录的用户密码确认身份，单击“确定”。注意，不要勾选“同时重启上层服务”，单击“确定”完成重启。

步骤6 重启服务完成后，等待2分钟查看告警是否自动恢复。

- 是，处理完毕。
- 否，执行**步骤7**。

检查告警阈值配置或者平滑次数配置是否合理

步骤7 登录FusionInsight Manager，基于实际数据库连接数使用率的情况，修改告警阈值和平滑次数配置项。选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > DBService >

数据库 > 数据库连接数使用率 (DBServer)”，单击平滑次数旁的铅笔标志，更改告警的平滑次数，如图10-41所示。

说明

平滑次数：连续检查多少次超过阈值，则发送告警。

图 10-41 设置告警平滑次数

根据数据库连接数使用率的实际情况，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > DBService > 数据库 > 数据库连接数使用率 (DBServer)”，单击“操作”栏的“修改”按钮，进入修改规则界面，修改后单击“确定”，修改即生效，如图10-42所示。

图 10-42 设置告警阈值

步骤8 等待2分钟，查看告警是否自动恢复。

- 是，处理完毕。
- 否，执行步骤9。

收集故障信息

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“DBService”。

步骤11 设置日志收集的主机，可选项，默认所有主机。

步骤12 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤13 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.188 ALM-27006 数据目录磁盘空间使用率超过阈值

告警解释

系统每30秒周期性检查DBServer主节点的数据目录磁盘空间使用率，并把实际数据目录磁盘空间使用率和阈值相比较，当数据目录磁盘空间使用率连续5次（可配置，默认值为5）超过设定阈值时，系统将产生此告警。数据目录磁盘空间使用率的阈值设为80%（可配置，默认值为80%）。

平滑次数可配置，当平滑次数为1，数据磁盘目录空间使用率小于或等于阈值时，该告警恢复；当平滑次数大于1，数据磁盘目录空间使用率小于阈值的90%时，该告警恢复。

告警属性

告警ID	告警级别	是否自动清除
27006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
设备分区名	产生告警的磁盘分区。

参数名称	参数含义
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

- 业务进程不可用。
- 当数据目录磁盘空间使用率超过90%时，数据库进入只读模式并发送告警“数据库进入只读模式”，业务数据丢失。

可能原因

- 告警阈值配置不合理。
- 数据库数据量过大或磁盘配置无法满足业务需求，导致磁盘使用率达到上限。

处理步骤

检查阈值设置是否合理

步骤1 在FusionInsight Manager，选择“运维 > 告警 > 阈值设置 > *待操作集群的名称* > DBService > 数据库 > 数据目录磁盘空间使用率”，查看该告警阈值是否合理（默认值80%为合理值）。

- 是，执行**步骤3**。
- 否，执行**步骤2**。

步骤2 根据实际服务的使用情况修改告警阈值。

步骤3 选择“集群 > *待操作集群的名称* > 服务 > DBService”，在“概览”页面查看“数据目录磁盘空间使用率”图表，检查数据目录磁盘空间使用率是否低于设置的阈值。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 等待2分钟查看告警是否自动恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

检查磁盘是否有误写入的大文件

步骤5 以**omm**用户登录DBService主管理节点。

步骤6 执行以下命令，查看数据目录磁盘空间下超过500MB的文件，检查该目录下是否有误写入的大文件存在。

source $DBSERVER_HOME/.dbservice_profile

find "$DBSERVICE_DATA_DIR"/../ -type f -size +500M

- 是，执行**步骤7**。
- 否，执行**步骤8**。

步骤7 根据实际情况处理误写入的文件，并等待2分钟，查看告警是否清除。

- 是，执行完毕。
- 否，执行**步骤8**。

收集故障信息

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“DBService”。

步骤10 设置日志收集的主机，可选项，默认所有主机。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.189 ALM-27007 数据库进入只读模式

告警解释

系统每30秒周期性检查DBServer主节点的数据目录磁盘空间使用率，当数据目录磁盘空间使用率超过90%时，系统将产生此告警。

当数据目录磁盘空间使用率低于80%时，此告警恢复。

告警属性

告警ID	告警级别	是否自动清除
27007	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。

参数名称	参数含义
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

数据库进入只读模式，业务数据丢失。

可能原因

磁盘配置无法满足业务需求，磁盘使用率达到上限。

处理步骤

检查磁盘使用率是否达到上限

步骤1 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > DBService”。

步骤2 在“概览”页面查看“数据目录磁盘空间使用率”图表，检查数据目录磁盘空间使用率是否超过90%。

- 是，执行**步骤3**。
- 否，执行**步骤13**。

步骤3 以**omm**用户登录DBServer主管理节点，执行以下命令，查看数据库是否进入只读模式。

source $DBSERVER_HOME/.dbservice_profile

gsql -U omm -W *password* **-d postgres -p 20051**

show default_transaction_read_only;

说明

其中*password*为DBService数据库的**omm**用户密码，用户可以执行**\q**退出数据库界面。

结果如下所示，查看“default_transaction_read_only”的值是否为“on”。

```
POSTGRES=# show default_transaction_read_only;
 default_transaction_read_only
-------------------------------
 on
(1 row)
```

- 是，执行**步骤4**。
- 否，执行**步骤13**。

步骤4 执行以下命令，打开“dbservice.properties”文件：

source $DBSERVER_HOME/.dbservice_profile

vi ${DBSERVICE_SOFTWARE_DIR}/tools/dbservice.properties

步骤5 修改“gaussdb_readonly_auto”的值为“OFF”，默认为“ON”。

步骤6 执行以下命令，打开dbservice.properties文件：

vi ${DBSERVICE_DATA_DIR}/postgresql.conf

步骤7 删除“default_transaction_read_only = on”。

步骤8 执行以下命令，使配置生效：

gs_ctl reload -D ${DBSERVICE_DATA_DIR}

步骤9 登录FusionInsight Manager，选择“运维 > 告警 > 告警”。单击告警“数据库进入只读模式”所在行右侧“操作”列中的“清除”，在弹出窗口中单击“确定”。手动清除该告警。

步骤10 以**omm**用户登录DBServer主管理节点，执行以下命令查看数据目录磁盘空间下超过500MB的文件，检查该目录下是否有误写入的大文件存在。

source $DBSERVER_HOME/.dbservice_profile

find "$DBSERVICE_DATA_DIR"/../ -type f -size +500M

- 是，执行**步骤11**。
- 否，执行**步骤13**。

步骤11 根据实际情况处理误写入的文件。

步骤12 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > DBService”，在“概览”页面查看“数据目录磁盘空间使用率”图表，检查数据目录磁盘空间使用率是否低于80%。

- 是，处理完毕。
- 否，执行**步骤13**。

收集故障信息

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选待操作集群的“DBService”。

步骤15 设置日志收集的主机，可选项，默认所有主机。

步骤16 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤17 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.190 ALM-29000 Impala 服务不可用

告警解释

以30s为周期检测Impala服务状态，当检测到Impala服务异常时，系统产生此告警。

当系统检测到Impala服务恢复正常，或告警处理完成时，告警解除。

告警属性

告警ID	告警级别	是否自动清除
29000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称
服务名	产生告警的服务名称
角色名	产生告警的角色名称
主机名	产生告警的主机名

对系统的影响

Impala服务异常，无法通过FusionInsight Manager对Impala进行集群操作，无法使用Impala服务功能。

可能原因

- Hive服务异常
- KrbServer服务异常
- Impala进程故障

处理步骤

检查Impala依赖的服务是否正常

步骤1 在FusionInsight Manager首页，选择“集群 > 服务”，查看Hive、KrbServer是否已停止。

- 是，启动已停止的服务，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中，查看“Impala服务不可用”告警是否清除。

- 是，处理完毕。

- 否，执行**步骤3**。

步骤3 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中，查看是否存在告警“ALM-16004 Hive服务不可用”，“ALM-25500 KrbServer服务不可用”。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 参考“ALM-16004 Hive服务不可用”，“ALM-25500 KrbServer服务不可用”告警帮助文档进行处理后，检查本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

检查Impala进程是否正常。

步骤5 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，在告警列表中查看是否存在“ALM-12007 进程故障”告警。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤6 参考“ALM-12007进程故障”告警帮助文档进行处理后，检出本告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Impala”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.191 ALM-29004 Impalad 进程内存占用率超过阈值

告警解释

以30s为周期检测Impalad进程系统内存占用率，当检测到的超过默认阈值（80%）时，系统产生此告警。

当系统检测到进程内存占用率下降到阈值以下时，告警将自动解除。

告警属性

告警ID	告警级别	是否自动清除
29004	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

内存使用过高，部分查询任务可能因为内存不足而失败。

可能原因

Impalad进程正在执行较大量查询任务。

处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > Impala > CPU和内存 > Impalad进程的内存占用率（Impalad）"，检查阈值大小。

步骤2 如阈值较小（小于80%），可根据实际需要适当增大告警阈值，检查告警是否消除。

- 是，处理完毕。
- 否，执行**步骤3**。

步骤3 如阈值已超过80%，请检查告警出现时刻是否有突发的大量并发查询任务，突发大量任务将会导致内存占用飙升，待任务执行完成后告警将自动消失，期间可能有因内存不足而执行失败或取消的任务，请重试。

说明

如内存占用超过阈值为常态化状态，需要考虑集群扩容。

----结束

告警清除

突发并发任务执行结束后告警自动清除。

参考信息

无

10.13.192 ALM-29005 Impalad JDBC 连接数超过阈值

告警解释

以30s为周期检测连接到该Impalad节点的客户端连接数，当检测到的连接数超过自定义阈值（默认60）时，系统产生此告警。

当系统检测到客户端连接数减少到阈值以下时，告警将自动解除。

告警属性

告警ID	告警级别	是否自动清除
29005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称
服务名	产生告警的服务名称
角色名	产生告警的角色名称
主机名	产生告警的主机名
Trigger Condition	系统当前指标取值满足自定义的告警设置条件

对系统的影响

后续新建立客户端连接可能会阻塞甚至失败。

可能原因

该Impalad服务维护的客户端链接过多，或者阈值设定的太小。

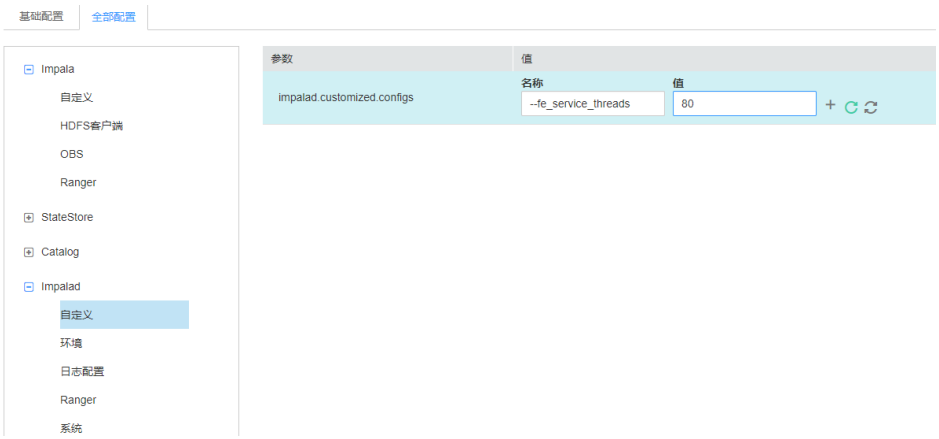
处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > Impala > 连接数 > 已经连接到Impalad进程的JDBC数量”，检查设置的阈值大小。

步骤2 检查连接到当前Impalad的JDBC应用数，并关闭闲置的应用，观察告警是否自动清除。

- 是，处理完毕。
- 否，执行**步骤3**，修改并发客户端连接数。

步骤3 在FusionInsight Manager首页，选择“集群 > Impala > 配置 > 全部配置 > Impalad > 自定义”，增加自定义参数 --fe_service_threads，该参数默认值64，请按照需要修改该值，单击“保存”按钮保存配置。



步骤4 在所有客户端的查询任务都执行完成后，选择“实例”页签，勾选所有“Impalad”实例并重启。

步骤5 重启完成后告警将消失，请重新运行使用JDBC方式连接Impalad的应用。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.193 ALM-29006 Impalad ODBC 连接数超过阈值

告警解释

以30s为周期检测连接到该Impalad节点的客户端连接数，当检测到的连接数超过自定义阈值（默认60）时，系统产生此告警。

当系统检测到客户端连接数减少到阈值以下时，告警将自动解除。

告警属性

告警ID	告警级别	是否自动清除
29006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称
服务名	产生告警的服务名称
角色名	产生告警的角色名称
主机名	产生告警的主机名
Trigger Condition	系统当前指标取值满足自定义的告警设置条件

对系统的影响

后续新建立客户端连接可能会阻塞甚至失败。

可能原因

该Impalad服务维护的客户端连接过多，或者阈值设定的太小。

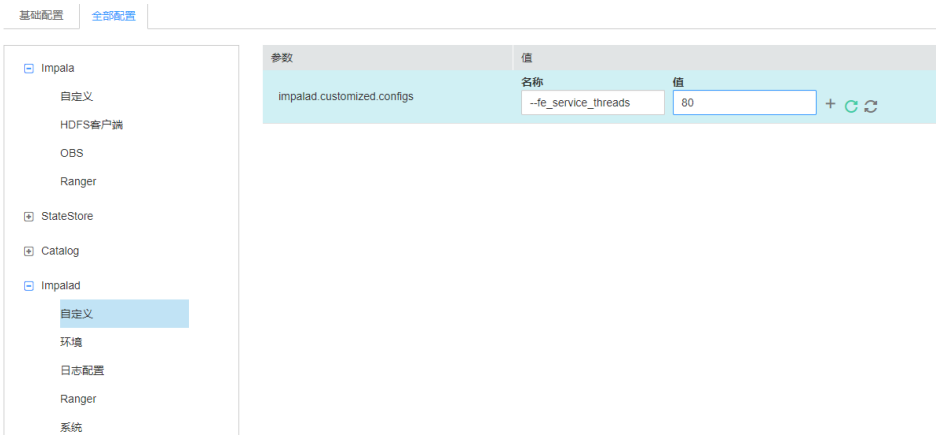
处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > Impala > 连接数 > 已经连接到Impalad进程的ODBC数量"，检查阈值大小。

步骤2 检查连接到当前Impalad进程的ODBC应用数，并关闭闲置的应用，观察告警是否自动清除。

- 是，处理完毕。
- 否，执行**步骤3**，修改并发Impalad支持的并发连接数。

步骤3 在FusionInsight Manager首页，选择“集群 > Impala > 配置 > 全部配置 > Impalad > 自定义”，增加自定义参数 --fe_service_threads，该参数默认值64，请按照需要修改该值，单击“保存”按钮保存配置。



步骤4 在所有客户端的查询任务都执行完成后，选择“实例”页签，勾选所有“Impalad”实例并重启。

步骤5 重启完成后告警将消失，请重新运行使用ODBC方式连接Impalad的应用。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.194 ALM-29100 Kudu 服务不可用

告警解释

系统每60秒周期性检测Kudu的服务状态，当检测到所有的Kudu实例都异常时，就会认为Kudu服务不可用，此时产生该告警。

至少一个Kudu实例正常后，系统认为Kudu实例服务恢复，告警清除。

告警属性

告警ID	告警级别	是否自动清除
29100	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

用户无法使用Kudu服务。

可能原因

Kudu有实例存在异常。

处理步骤

处理Kudu实例异常

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”页面，找到“ALM-29100 Kudu服务异常”告警。

步骤2 查看告警的“定位信息”一栏，记录主机名与角色名。

步骤3 选择“集群 > 服务 > Kudu > 实例”，单击**步骤2**中对应主机名的角色名称，通过查看本实例的日志，修复这个实例，然后查看是否消除各个Kudu实例异常告警。

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 在“运维 > 告警 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息

步骤5 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选待操作集群的如下节点信息。

- Kudu

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除

参考信息

无

10.13.195 ALM-29104 Tserver 进程内存占用率超过阈值

告警解释

系统每60秒周期性检测Kudu Tserver进程内存占用率，当检测到Tserver进程占用率超过阈值，此时产生该告警。

Tserver进程内存占用率恢正常后，系统认为Kudu实例服务恢复，告警清除。

告警属性

告警ID	告警级别	是否自动清除
29104	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

用户无法使用Kudu服务。

可能原因

存在KuduTserver实例内存占用率过高。

处理步骤

处理Kudu实例异常

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”页面，找到“ALM-29104 Tserver进程内存占用率超过阈值”告警，查看告警来源。

步骤2 在“运维 > 告警 > 阈值设置 > Kudu”，找到该告警的阈值，再对比集群Kudu实例的内存监控项，看是否超过阈值，处理内存使用率过高的问题，或修改阈值。

步骤3 在“运维 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**4**。

收集故障信息

步骤4 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的如下节点信息。

- Kudu

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除

参考信息

无

10.13.196 ALM-29106 Tserver 进程 CPU 占用率过高

告警解释

系统每60秒周期性检测Kudu的服务状态，当检测到Kudu Tserver进程CPU占用率过高时，此时产生该告警。

Tserver进程CPU占用率正常时，系统认为Kudu实例服务恢复，告警清除。

告警属性

告警ID	告警级别	是否自动清除
29106	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

用户无法使用Kudu服务。

可能原因

存在KuduTserver实例CPU占用率过高。

处理步骤

处理Kudu实例异常

步骤1 在FusionInsight Manager首页，选择“运维 > 告警”页面，找到“ALM-29106 Tserver进程CPU占用率超过阈值”告警，查看告警来源。

步骤2 在“运维 > 告警 > 阈值设置 > Kudu”，找到该告警的阈值，再对比集群Kudu实例的CPU使用率监控项，和阈值对比，查看超阈值情值，处理CPU使用率过高的问题，或修改阈值。

步骤3 在“运维 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**4**。

收集故障信息

步骤4 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的如下节点信息。

- Kudu

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.197 ALM-29107 Tserver 进程内存使用百分比超过阈值

告警解释

系统每60秒周期性检测Kudu的服务状态，当检测到Kudu Tserver进程内存使用百分比超过阈值，此时产生该告警。

Tserver进程内存使用百分比正常时，系统认为Kudu实例服务恢复，告警清除。

告警属性

告警ID	告警级别	是否自动清除
29107	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

用户无法使用Kudu服务。

可能原因

存在KuduTserver实例内存使用过高。

处理步骤

处理Kudu实例异常

步骤1 在FusionInsight Manager首页，选择“运维 > 告警”页面，找到“ALM-29107 Tserver进程内存使用百分比超过阈值”告警，查看告警来源。

步骤2 在“运维 > 告警 > 阈值设置 > Kudu”，找到该告警的阈值，再对比集群KuduTserver实例的内存使用百分比监控项，和阈值对比，查看阈值超过情况，找到内存使用百分比超阈值的节点。

通过增加节点、重新规划任务等方式，处理Tserver节点内存使用百分比过高的问题，或修改阈值。

步骤3 在“运维 > 告警”页签，查看该告警是否恢复。

- 是，处理完毕。
- 否，执行**4**。

收集故障信息

步骤4 在FusionInsight Manager首页，单击“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的如下节点信息。

- Kudu

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.198 ALM-38000 Kafka 服务不可用

告警解释

系统按照30秒的周期检测Kafka服务是否可用，当Kafka服务不可用，系统产生此告警。

当Kafka服务恢复正常，告警自动清除。

告警属性

告警ID	告警级别	是否自动清除
38000	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

集群无法对外提供Kafka服务，用户无法执行新的Kafka任务。

可能原因

- KrbServer服务故障。（非普通模式集群）
- ZooKeeper服务故障或无响应。
- Kafka服务中Broker实例状态异常。

处理步骤

检查KrbServer服务状态。（普通模式集群跳过此步骤）

步骤1 在FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务 > KrbServer”。

步骤2 查看KrbServer服务的运行状态是否为“良好”。

- 是，执行**步骤5**。
- 否，执行**步骤3**。

步骤3 参考“ALM-25500 KrbServer服务不可用”的处理步骤进行操作。

步骤4 再次执行**步骤2**。

检查ZooKeeper服务状态。

步骤5 查看ZooKeeper服务的运行状态是否为“良好”。

- 是，执行**步骤8**。
- 否，执行**步骤6**。

步骤6 如果ZooKeeper服务已停止，则启动ZooKeeper服务，否则参考“ALM-13000 ZooKeeper服务不可用”的处理步骤进行操作。

步骤7 再次执行**步骤5**。

检查Broker实例状态。

步骤8 选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，进入Kafka实例页面。

步骤9 查看“角色”中所有实例是否正常。

- 是，执行**步骤11**。
- 否，执行**步骤10**。

步骤10 勾选Broker所有实例，选择“更多 > 重启实例”，查看是否重启成功。

- 是，执行**步骤11**。
- 否，执行**步骤13**。

步骤11 选择“集群 > *待操作集群的名称* > 服务 > Kafka”，查看运行状态是否为“良好”。

- 是，执行**步骤12**。
- 否，执行**步骤13**。

步骤12 等待30秒，查看告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤13**。

收集故障信息。

步骤13 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤14 在“服务”中勾选待操作集群的“Kafka”。

步骤15 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤16 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.199 ALM-38001 Kafka 磁盘容量不足

告警解释

系统按60秒周期检测Kafka磁盘空间使用率，并把实际磁盘使用率和阈值相比较。磁盘使用率默认提供一个阈值范围。当检测到磁盘使用率高于阈值时产生该告警。

用户可通过“运维 > 告警 > 阈值设置”，在服务列表下面，选择“Kafka > 磁盘 > Broker磁盘使用率 (Broker)”修改阈值。

平滑次数为1，Kafka磁盘使用率小于或等于阈值时，告警恢复；平滑次数大于1，Kafka磁盘使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38001	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
设备分区名	产生告警的磁盘分区。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

磁盘容量不足会导致Kafka写入数据失败。

可能原因

- 用于存储Kafka数据的磁盘配置（如磁盘数目、磁盘大小等），无法满足当前业务数据流量，导致磁盘使用率达到上限。
- 数据保存时间配置过长，数据累积达到磁盘使用率上限。
- 业务规划不合理，导致数据分配不均，使部分磁盘达到使用率上限。

处理步骤

检查Kafka数据的磁盘配置。

步骤1 在FusionInsight Manager管理界面，选择“运维 > 告警 > 告警”。

步骤2 在告警列表中单击该告警，从“定位信息”中获得主机名。

步骤3 选择“集群 > *待操作集群的名称* > 主机”。

步骤4 在“主机”页面单击**步骤2**中获取的主机名称。

步骤5 检查“磁盘”区域中是否包含该告警中的磁盘分区名称。

- 是，执行**步骤6**。
- 否，手动清除该告警，操作结束。

步骤6 检查“磁盘”区域中包含该告警中的磁盘分区使用率是否达到百分之百。

- 是，参考**参考信息**进行处理。
- 否，执行**步骤7**

检查Kafka数据保存时间配置。

步骤7 选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置 > 全部配置”。

步骤8 查看“disk.adapter.enable”参数是否配置为“true”。

- 是，执行**步骤10**。
- 否，执行**步骤9**。

步骤9 将“disk.adapter.enable”配置为“true”，开启该功能。然后查看“adapter.topic.min.retention.hours”所配置的数据最短保存周期是否合理。

- 是，执行**步骤10**。
- 否，根据业务需求合理调整数据保存周期。

须知

启用磁盘自适应功能可能导致Topic的历史数据被清除，如果有个别Topic不能做保存周期调整，单击“全部配置”，将Topic配置在“disk.adapter.topic.blacklist”参数中。

步骤10 等待10分钟，查看故障磁盘的使用率是否有减少。

- 是，继续等待直到告警消除。
- 否，执行**步骤11**。

检查Kafka数据规划。

步骤11 选择上报告警实例主机名对应的角色“Broker”。单击图表区域右上角的下拉菜单，选择“定制”，来自定义监控项。

步骤12 在弹出的“定制”对话框中，选择“磁盘 > Broker磁盘使用率”，并单击“确定”。

关于Kafka磁盘使用情况信息会被显示。

步骤13 根据**步骤12**的显示信息，查看是否只有**步骤2**中上报告警的磁盘分区。

- 是，执行**步骤14**。
- 否，执行**步骤15**。

步骤14 重新进行磁盘规划，挂载新的磁盘，进入当前问题节点“实例配置”页面，重新配置“log.dirs”，增加其他磁盘相应路径，重启当前Kafka实例。

步骤15 查看Kafka配置的数据保存时间配置，根据业务需求和业务量权衡，考虑是否需要调小数据保存时间。

- 是，执行**步骤16**。
- 否，执行**步骤17**。

步骤16 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置 > 全部配置”，在右侧搜索框中填写配置项名称“log.retention.hours”，然后会显示该配置的当前值，此处的值为Topic默认的数据保存时间，可以适当调小该值。

说明

- 对于单独配置数据保存时间的Topic，修改Kafka服务配置页面上配置的数据保存时间不生效。
- 如果需要对某个Topic单独配置的话，可以使用Kafka客户端命令行，来单独配置该Topic。
  例如：**kafka-topics.sh --zookeeper “*ZooKeeper地址*:2181/kafka” --alter --topic “*Topic名称*” --config retention.ms=“*保存时间*”**

步骤17 查看是否由于某些Topic的Partition配置不合理导致部分磁盘使用率达到上限（例如：数据量非常大的Topic的Partition数目小于配置的磁盘个数，导致各磁盘上数据分配无法均匀，进而部分磁盘达到使用率上限）。

说明

如果不清楚哪些Topic业务数据量较大，可以根据**步骤2**中获取到的主机节点信息，登录到实例节点上，进入对应的数据目录（即**步骤14**中“log.dirs”修改之前的配置路径），查看该目录下哪些Topic的Partition目录占用的磁盘空间比较大。

- 是，执行**步骤18**。
- 否，执行**步骤19**。

步骤18 通过Kafka客户端对Topic的Partition进行扩展，命令行操作命令如下：

kafka-topics.sh --zookeeper “*ZooKeeper地址*:2181/kafka” --alter --topic “*Topic名称*” --partitions=“*新Partition数目*”

说明

- 新Partition数目建议配置为Kafka数据磁盘数量的倍数。
- 当前步骤修改可能不会很快解决当前告警，需要结合**步骤11**中的数据保存时间逐渐均衡数据。

步骤19 考虑是否需要扩容。

说明

建议当前Kafka磁盘使用率超过80%时，则需要扩容。

- 是，执行**步骤20**。
- 否，执行**步骤21**。

步骤20 扩展磁盘容量，扩展后检查告警是否消失。

- 是，操作结束。
- 否，执行**步骤22**。

步骤21 检查告警是否清除。

- 是，操作结束。
- 否，执行**步骤22**。

收集故障信息。

步骤22 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤23 在“服务”中勾选待操作集群的“Kafka”。

步骤24 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤25 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

步骤1 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，将运行状态为“正在恢复”的Broker实例停止并记录实例所在节点的管理IP地址以及对应的“broker.id”，该值可通过单击角色名称，在“实例配置”页面中选择“全部配置”，搜索“broker.id”参数获取。

步骤2 以**root**用户登录记录的管理IP地址，并执行**df -lh**命令，查看磁盘占用率为100%的挂载目录，例如“${BIGDATA_DATA_HOME}/kafka/data1”。

步骤3 进入该目录，执行**du -sh ***命令，查看该目录下各文件夹的大小。查看是否存在除“kafka-logs”目录外的其他文件，并判断是否可以删除或者迁移。

- 是，删除或者迁移相关数据，然后执行**步骤8**。
- 否，执行**步骤4**。

步骤4 进入“kafka-logs”目录，执行**du -sh ***命令，选择一个待移动的Partition文件夹，其名称命名规则为“Topic名称-Partition标识”，记录Topic及Partition。

步骤5 修改“kafka-logs”目录下的“recovery-point-offset-checkpoint”和“replication-offset-checkpoint”文件（两个文件做同样的修改）。

1. 减少文件中第二行的数字（若移出多个目录，则减少的数字为移出的目录个数）。
2. 删除待移出的Partition所在的行（行结构为“Topic名称 Partition标识 Offset”，删除前先将该行数据保存，后续此内容还要添加到目的目录下的同名文件中）。

步骤6 修改目的数据目录下（例如：“${BIGDATA_DATA_HOME}/kafka/data2/kafka-logs”）的“recovery-point-offset-checkpoint”和“replication-offset-checkpoint”文件（两个文件做同样的修改）。

- 增加文件中第二行的数字（若移入多个Partition目录，则增加的数字为移入的Partition目录个数）。
- 添加待移入的Partition行到文件末尾（行结构为“Topic名称 Partition标识 Offset”，直接复制**步骤5**中保存的行数据即可）。

步骤7 移动数据，将待移动的Partition文件夹移动到目的目录下，移动完成后执行**chown omm:wheel -R** *Partition目录*命令修改Partition目录属组。

步骤8 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，启动停止的Broker实例。

步骤9 等待5至10分钟后查看Broker实例的运行状态是否为“良好”。

- 是，修复完成后按照“ALM-38001 Kafka磁盘容量不足”告警指导彻底解决磁盘容量不足问题。
- 否，联系运维人员。

----结束

10.13.200 ALM-38002 Kafka 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测Kafka服务堆内存使用状态，当连续10次检测到Kafka实例堆内存使用率超出阈值（最大内存的95%）时产生该告警。

平滑次数为1，堆内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，堆内存使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38002	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Kafka可用内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点Kafka实例堆内存使用率过大，或配置的堆内存大小不合理，导致使用率超过阈值。

处理步骤

检查Kafka实例堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Kafka堆内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，选择上报告警实例主机名对应的角色。单击图表区域右上角的下拉菜单，选择“定制 > 进程 > Kafka堆内存使用率”，单击“确定”。

步骤3 查看Kafka使用的堆内存是否已达到Kafka设定的最大堆内存的95%。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

检查Kafka配置的堆内存大小。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置 > 全部配置 > Broker（角色） > 环境变量”。将“KAFKA_HEAP_OPTS”参数的值参考如下说明调大。

说明

- 建议“KAFKA_HEAP_OPTS”参数中“-Xmx”和“-Xms”值保持一致。
- 建议根据**步骤2**查看“Kafka堆内存使用率”，调整“KAFKA_HEAP_OPTS”的值为“Kafka使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Kafka”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.201 ALM-38004 Kafka 直接内存使用率超过阈值

告警解释

系统每30秒周期性检测Kafka服务直接内存使用状态，当连续10次检测到Kafka实例直接内存使用率超出阈值（最大内存的80%）时，产生该告警。

平滑次数为1，直接内存使用率小于或等于阈值时，告警恢复；平滑次数大于1，直接内存使用率小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38004	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Kafka可用直接内存不足，可能会造成内存溢出导致服务崩溃。

可能原因

该节点Kafka实例直接内存使用率过大，或配置的直接内存大小不合理，导致使用率超过阈值。

处理步骤

检查Kafka实例直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Kafka直接内存使用率超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，选择上报告警实例主机名对应的角色。单击图表区域右上角的下拉菜单，选择“定制 > 进程 > Kafka直接内存使用率”，单击“确定”。

步骤3 查看Kafka使用的直接内存是否已达到Kafka设定的最大直接内存的80%。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

检查Kafka配置的直接内存大小。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置 > 全部配置 > Broker（角色）> 环境变量”。将“KAFKA_HEAP_OPTS”参数中配置的“-Xmx”值参考如下说明调大。

说明

- 建议“KAFKA_HEAP_OPTS”参数中“-Xmx”和“-Xms”值保持一致。
- 建议根据步骤2查看“Kafka直接内存使用率”，调整“KAFKA_HEAP_OPTS”的值为“Kafka使用的直接内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 保存配置，并重启Kafka服务。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行步骤7。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Kafka”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.202 ALM-38005 Broker 进程垃圾回收（GC）时间超过阈值

告警解释

系统每60秒周期性检测Broker进程的垃圾回收（GC）占用时间，当连续3次检测到Broker进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。

平滑次数为1，垃圾回收（GC）时间小于或等于阈值时，告警恢复；平滑次数大于1，垃圾回收（GC）时间小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38005	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Broker进程的垃圾回收时间过长，可能影响该Broker进程正常提供服务。

可能原因

该节点Kafka实例进程的垃圾回收时间过长，或配置的直接内存大小不合理，导致进程GC频繁。

处理步骤

检查Broker进程的垃圾回收（GC）时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Broker进程垃圾回收（GC）时间超过阈值 > 定位信息”。查看告警上报的实例的主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，选择上报告警实例主机名对应的角色 。单击图表区域右上角的下拉菜单，选择“定制 > 进程 > Broker垃圾回收（GC）时间”，单击“确定”。

步骤3 查看Broker每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

检查Kafka配置的直接内存大小。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置 > 全部配置 > Broker（角色） > 环境变量”。将“KAFKA_HEAP_OPTS”参数中配置的“-Xmx”值参考如下说明调大。

说明

- 建议“KAFKA_HEAP_OPTS”参数中“-Xmx”和“-Xms”值保持一致。
- 建议根据“Kafka直接内存资源状况”调整“KAFKA_HEAP_OPTS”的值为“Kafka使用的直接内存大小”的两倍（可根据实际业务场景进行修改）。“Kafka直接内存资源状况”可在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，选择上报告警实例主机名对应的角色。单击图表区域右上角的下拉菜单，选择“定制 > 进程 > Kafka直接内存资源状况”进行查看。

步骤5 保存配置，并重启Kafka服务。

步骤6 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Kafka”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.203 ALM-38006 Kafka 未完全同步的 Partition 百分比超过阈值

告警解释

系统每60秒周期性检测Kafka服务未完全同步的Partition数占Partition总数的百分比，当连续3次检测到该比率超出阈值（默认50%）时产生该告警。

平滑次数为1，未完全同步的Partition百分比小于或等于阈值时，告警恢复；平滑次数大于1，未完全同步的Partition百分比小于或等于阈值的90%时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

Kafka服务未完全同步的Partition数过多，会影响服务的可靠性，一旦发生leader切换，可能会导致丢数据。

可能原因

部分Broker实例所在节点故障或者实例停止运行，导致Kafka中某些Partition的副本下线。

处理步骤

检查Broker实例。

步骤1 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，进入Kafka实例页面。

步骤2 查看所有Broker实例中是否有故障的节点。

- 是，记录当前节点主机名，并执行**步骤3**。
- 否，执行**步骤5**。

步骤3 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”。查看所有告警信息中是否有**步骤2**中节点主机对应的故障告警，根据对应的告警指导进行处理。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，进入Kafka实例页面。

步骤5 查看所有Broker实例中是否有已停止的实例。

- 是，执行**步骤6**。
- 否，执行**步骤7**。

步骤6 勾选所有已停止的Broker实例，单击“启动实例”。

步骤7 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“Kafka”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.204 ALM-38007 Kafka 默认用户状态异常

告警解释

系统每60秒周期性检测Kafka服务默认用户，当检测到该用户异常时发送此告警。

平滑次数为1，当用户状态恢复后，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38007	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名称。
Trigger Condition	Kafka默认用户状态异常。

对系统的影响

Kafka默认用户状态异常，会影响Broker之间的元数据同步，以及Kafka与ZooKeeper之间的交互，进而影响业务生产、消费和Topic的创建、删除等操作。

可能原因

- Sssd服务异常导致。
- 部分Broker实例停止运行。

处理步骤

检查是否有"Sssd服务异常"告警。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > Kafka默认用户状态异常 > 定位信息”。查看告警上报的实例的主机名。

步骤2 根据告警提示的主机信息，登录到该节点上。

步骤3 执行**id -Gn kafka**，查看返回结果是否报"No such user"。

- 是，记录当前节点主机名，并执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”。查看所有告警信息中是否有"Sssd服务异常"告警，根据对应的告警指导进行处理。

检查Broker实例运行状态。

步骤5 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，进入Kafka实例页面。

步骤6 查看所有Broker实例中是否有已停止的节点。

- 是，执行**步骤7**。
- 否，执行**步骤8**。

步骤7 勾选所有已停止的Broker实例，单击“启动实例”。

步骤8 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“Kafka”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.205 ALM-38008 Kafka 数据目录状态异常

告警解释

系统每60秒周期性检测Kafka数据目录状态，当检测到某数据目录状态异常时产生该告警。

平滑次数为1，当数据目录状态恢复正常后，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名称。
目录名	产生告警的目录名称。
Trigger Condition	Kafka数据目录状态异常。

对系统的影响

Kafka数据目录状态异常，会导致该数据目录上所有Partition的当前副本下线，多个节点同时出现数据目录状态异常，可能会导致部分Partition不可用。

可能原因

- 数据目录权限被篡改。
- 数据目录所在磁盘故障。

处理步骤

检查故障的数据目录权限。

步骤1 根据告警提示的主机信息，登录到该节点上。

步骤2 查看告警详细信息中所提示的数据目录及其子目录，属组是否为omm:wheel。

- 是，记录当前节点主机名，并执行**步骤4**。
- 否，执行**步骤3**。

步骤3 恢复数据目录及其子目录的属组为omm:wheel。

检查数据目录所在磁盘是否故障。

步骤4 使用omm用户，在所提示的数据目录的上一级目录下，进行创建、删除文件测试，看能够正常读写磁盘。

- 是，执行**步骤6**。
- 否，执行**步骤5**。

步骤5 更换或者修复数据目录所在磁盘，保证其可以正常读写。

步骤6 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，进入Kafka实例页面，重启**步骤2**中主机名上的Broker实例。

步骤7 等待Broker启动完成之后，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤8**。

收集故障信息。

步骤8 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤9 在“服务”中勾选待操作集群的“Kafka”。

步骤10 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤11 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.206 ALM-38009 Broker 磁盘 IO 繁忙

告警解释

系统每60秒周期性检测Kafka各个磁盘的IO情况，当检测到某个Broker上的Kafka数据目录磁盘IO超出阈值（默认80%）时，产生该告警。

平滑次数为3，当该磁盘IO低于阈值（默认80%）时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
38009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
数据目录名称	Kafka磁盘IO频繁的数据目录名称

对系统的影响

Partition所在的磁盘分区IO过于繁忙，产生告警的Kafka Topic上可能无法写入数据。

可能原因

- Topic副本数配置过多。
- 生产者消息批量写入磁盘的参数设置不合理。该Topic承担的业务流量过大，当前Partition的设置不合理。

处理步骤

检查Topic副本数配置。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击此告警所在行的，查看定位信息中上报告警的“主题名”。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Kafka > KafkaTopic监控”，搜索发生告警的Topic，查看副本数量。

步骤3 如果副本数量值大于3，则考虑减少该Topic的复制因子（减少为3）。

在FusionInsight客户端执行以下命令对Kafka Topic的副本进行重新规划：

kafka-reassign-partitions.sh --zookeeper *{zk_host}:{port}*/**kafka --reassignment-json-file** *{manual assignment json file path}* **--execute**

例如：

/opt/Bigdata/client/Kafka/kafka/bin/kafka-reassign-partitions.sh --zookeeper 10.149.0.90:2181,10.149.0.91:2181,10.149.0.92:2181/kafka --reassignment-json-file expand-cluster-reassignment.json --execute

说明

在**expand-cluster-reassignment.json**文件中描述该Topic的Partition迁移到哪些Broker。其中json文件中的内容格式为：{"partitions":[{"topic": "*topicName*","partition": 1,"replicas": [1,2,3] }],"version":1}。

步骤4 观察一段时间，看告警是否消失。如果告警没有消失，执行**步骤5**。

检查Topic的Partition规划设置。

步骤5 在“KafkaTopic监控”页面单击每一个Topic的“Topic的字节流量 > Topic输入的字节流量”，统计出“Topic输入的字节流量”值最大的Topic。查看该Topic有哪些Partition以及这些Partition所在的主机信息。

步骤6 登录到**步骤5**查询到的主机，执行**iostat -d -x**命令查看每个磁盘的最后一个指标“%util”：

```
189-39-172-162:/opt/R3/FusionInsight_Manager/software/packs # iostat -d -x
Linux 3.0.76-0.11-default (189-39-172-162)      06/26/19        _x86_64_

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
xvda              0.04    44.44    1.26   21.94    43.62   531.02    24.78     0.03    1.44   0.56   1.30
xvde              0.16   431.84   13.78   82.51   284.32  4115.90    45.70     0.06    1.41   0.64   6.21
```

- 各个磁盘的“%util”指标都比较高，则考虑对Kafka磁盘进行扩容，扩容后，参考**步骤3**，对Topic的Partition重新规划。
- 各个磁盘的“%util”指标差别较大，查看Kafka的磁盘分区配置信息。例如：${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/1_14_Broker/etc/server.properties文件中的log.dirs配置值。

  执行如下命令查看命令输出的Filesystem信息：

  df -h *log.dirs配置值*

  执行结果如下：

  ```
  189-39-172-162:/opt/R3/FusionInsight_Manager/software/packs # df -h /srv/BigData/kafka/data1/kafka-logs/
  Filesystem      Size  Used Avail Use% Mounted on
  /dev/xvda2       36G   21G   14G  62% /
  189-39-172-162:/opt/R3/FusionInsight_Manager/software/packs #
  ```

- Filesystem所在的分区与“%util”指标比较高的分区相匹配，则考虑在空闲的磁盘上规划Kafka分区，并将log.dirs设置为空闲磁盘目录，然后参考**步骤3**，对Topic的Partition重新规划，保证该Topic的Partition均匀分布到各个磁盘。

步骤7 观察一段时间，检查告警是否清除。

- 告警清除，操作结束。
- 告警没有清除，重复执行**步骤5**~**步骤6**三次。重复执行次数达到上限，执行**步骤8**。

步骤8 观察一段时间，检查告警是否清除。

- 是，操作结束。
- 否，执行**步骤9**。

收集故障信息。

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“Kafka”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.207 ALM-38010 存在单副本的 Topic

告警解释

系统在Kafka的Controller所在节点上，每60秒周期性检测各个Topic的副本数，当检测到某个Topic的副本数为1时，产生该告警。

告警属性

告警ID	告警级别	是否自动清除
38010	提示	否

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
角色名	产生告警的角色名称。
主题名	产生告警的Topic名称列表。

对系统的影响

单副本的Topic存在单点故障风险，当副本所在节点异常时，会直接导致Partition没有leader，影响该Topic上的业务。

可能原因

Topic副本数配置不合理。

处理步骤

检查Topic副本数配置。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，单击此告警所在行的，查看定位信息中上报告警的“主题名”列表。

步骤2 确认发生告警Topic是否需要增加副本。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 在FusionInsight客户端，对相关Topic的副本进行重新规划，在**add-replicas-reassignment.json**文件中描述该Topic的Partition分布信息，其中json文件中的内容格式为：{"partitions":[{"topic": "*topicName*","partition": 1,"replicas": [1,2] }],"version":1}，并执行如下命令增加副本：

kafka-reassign-partitions.sh --zookeeper *{zk_host}:{port}*/**kafka --reassignment-json-file** *{manual assignment json file path}* **--execute**

例如：

/opt/Bigdata/client/Kafka/kafka/bin/kafka-reassign-partitions.sh --zookeeper 192.168.0.90:2181,192.168.0.91:2181,192.168.0.92:2181/kafka --reassignment-json-file add-replicas-reassignment.json --execute

步骤4 执行如下命令进行确认任务执行进度：

kafka-reassign-partitions.sh --zookeeper *{zk_host}:{port}*/**kafka --reassignment-json-file** *{manual assignment json file path}* **--verify**

例如：

/opt/Bigdata/client/Kafka/kafka/bin/kafka-reassign-partitions.sh --zookeeper 192.168.0.90:2181,192.168.0.91:2181,192.168.0.92:2181/kafka --reassignment-json-file add-replicas-reassignment.json --verify

步骤5 确认处理完成或者告警无影响后，可在FusionInsight Manager页面，手动清除该告警。

步骤6 观察一段时间，检查告警是否清除或者告警无影响后，可在FusionInsight Manager页面，手动清除该告警。

- 是，操作结束。
- 否，执行**步骤7**。

收集故障信息。

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Kafka”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

确认告警已无影响，可手工清除告警。

参考信息

无。

10.13.208 ALM-43001 Spark2x 服务不可用

告警解释

系统每300秒周期性检测Spark2x服务状态，当检测到Spark2x服务不可用时产生该告警。

Spark2x服务恢复时，告警清除。

告警属性

告警ID	告警级别	是否自动清除
43001	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

用户提交的Spark任务执行失败。

可能原因

- KrbServer服务异常。
- LdapServer服务异常。
- ZooKeeper服务异常。
- HDFS服务异常。
- Yarn服务异常。
- 对应的Hive服务异常。
- Spark2x assembly包异常。

处理步骤

若告警原因为：Spark2x assembly包异常，则表示spark的包存在异常，等待10分钟左右，告警自动恢复。

检查Spark2x依赖的服务是否有服务不可用告警。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”。

步骤2 在告警列表中，查看是否存在以下告警：

- ALM-25500 KrbServer服务不可用
- ALM-25000 LdapServer服务不可用
- ALM-13000 ZooKeeper服务不可用
- ALM-14000 HDFS服务不可用
- ALM-18000 Yarn服务不可用

- ALM-16004 Hive服务不可用

说明

若集群启用了多实例功能且安装了多个Spark2x服务，请根据“定位信息”中的“服务名”值来查看具体产生告警的Spark2x服务，然后确认对应的Hive服务是否故障，Spark2x对应Hive，Spark2x1对应Hive1，以此类推。

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 根据对应服务不可用告警帮助提供的故障处理对应告警。

告警全部恢复后，等待几分钟，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤4**。

收集故障信息。

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”中勾选待操作集群的如下节点信息。（Hive为根据告警定位信息中的“服务名”确定的具体Hive服务。）

- KrbServer
- LdapServer
- ZooKeeper
- HDFS
- Yarn
- Hive

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.209 ALM-43006 JobHistory2x 进程堆内存使用超出阈值

告警解释

系统每30秒周期性检测JobHistory2x进程堆内存使用状态，当检测到JobHistory2x进程堆内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43006	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JobHistory2x进程堆内存使用率过高，会影响JobHistory2x进程运行的性能，甚至造成内存溢出导致JobHistory2x进程不可用。

可能原因

该节点JobHistory2x进程堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43006”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JobHistory2x内存使用率统计”，单击“确定”，查看JobHistory2x进程使用的堆内存是否已达到JobHistory2x进程设定的最大堆内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > 内存 > JobHistory2x进程的堆内存统计”，单击“确定”，根据告警产生时间，查看对应时间段的“JobHistory2x进程使用的堆内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JobHistory2x > 默认”，“SPARK_DAEMON_MEMORY”参数默认值为4G，可根据如下方案调整该参数值：告警时间段内JobHistory2x使用堆内存的最大值和“JobHistory2x堆内存使用率统计(JobHistory2x)”阈值的比值。若参数值调整后，仍偶现告警，可以按0.5倍速率调大。若频繁出现告警，可以按1倍速率调大。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > JobHistory2x堆内存使用率统计 (JobHistory2x)”，可查看“阈值”。

步骤5 重启所有的JobHistory2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.210 ALM-43007 JobHistory2x 进程非堆内存使用超出阈值

告警解释

系统每30秒周期性检测JobHistory2x进程非堆内存使用状态，当检测到JobHistory2x进程非堆内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43007	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JobHistory2x进程非堆内存使用率过高，会影响JobHistory2x进程运行的性能，甚至造成内存溢出导致JobHistory2x进程不可用。

可能原因

该节点JobHistory2x进程非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43007”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JobHistory2x内存使用率统计”，单击“确定”，查看JobHistory2x进程使用的非堆内存是否已达到JobHistory2x进程设定的最大非堆内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > 内存 > JobHistory2x进程的非堆内存统计”，单击“确定”，根据告警产生时间，查看对应时间段的“JobHistory2x进程使用的非堆内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JobHistory2x > 默认”，根据如下原则调整“SPARK_DAEMON_JAVA_OPTS”参数中-XX:MaxMetaspaceSize的值：告警时间段内JobHistory2x使用非堆内存的最大值和“JobHistory2x非堆内存使用率统计(JobHistory2x)”阈值的比值。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > JobHistory2x非堆内存使用率统计 (JobHistory2x)”，可查看“阈值”。

步骤5 重启所有的JobHistory2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.211 ALM-43008 JobHistory2x 进程直接内存使用超出阈值

告警解释

系统每30秒周期性检测JobHistory2x进程直接内存使用状态，当检测到JobHistory2x进程直接内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43008	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JobHistory2x进程直接内存使用率过高，会影响JobHistory2x进程运行的性能，甚至造成内存溢出导致JobHistory2x进程不可用。

可能原因

该节点JobHistory2x进程直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43008”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JobHistory2x内存使用率统计”，单击“确定”，查看JobHistory2x进程使用的直接内存是否已达到JobHistory2x进程设定的最大直接内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JobHistory2x直接内存”，单击“确定”，根据告警产生时间，查看对应时间段的“JobHistory2x进程使用的直接内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JobHistory2x > 默认”，“SPARK_DAEMON_JAVA_OPTS”参数中“-XX:MaxDirectMemorySize”的默认值为512M，可根据如下原则调整: 告警时间段内JobHistory2x使用直接内存的最大值和“JobHistory2x直接内存使用率统计 (JobHistory2x)”阈值的比值。若参数值调整后，仍偶现告警，可以按0.5倍速率调大。若频繁出现告警，可以按1倍速率调大，建议不要超过参数SPARK_DAEMON_MEMORY的值。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > JobHistory2x直接内存使用率统计 (JobHistory2x)”，可查看“阈值”。

步骤5 重启所有的JobHistory2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.212 ALM-43009 JobHistory2x 进程 GC 时间超出阈值

告警解释

系统每60秒周期性检测JobHistory2x进程的GC时间，当检测到JobHistory2x进程的GC时间超出阈值（连续3次检测超过12秒）时产生该告警。用户可通过“运维 >告警 > 阈值设置 > *待操作集群的名称* > Spark2x > GC时间 > JobHistory2x的总GC时间”修改阈值。当JobHistory2x进程 GC时间小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
43009	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

参数名称	参数含义
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC时间超出阈值，会影响JobHistory2x进程运行的性能，甚至造成JobHistory2x进程不可用。

可能原因

该节点JobHistory2x进程堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43009”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JobHistory2x的GC时间”，单击“确定”，查看JobHistory2x进程的GC时间是否大于阈值（默认12秒）。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JobHistory2x > 默认”，将“SPARK_DAEMON_MEMORY”参数的值根据如下原则调整：
“SPARK_DAEMON_MEMORY”参数默认值为4G，若偶现告警，可以按0.5倍速率调大。若告警次数比较频繁，可以按1倍速率调大。

步骤4 重启所有的JobHistory2x实例。

步骤5 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Spark2x”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.213 ALM-43010 JDBCServer2x 进程堆内存使用超出阈值

告警解释

系统每30秒周期性检测JDBCServer2x进程堆内存使用状态，当检测到JDBCServer2x进程堆内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43010	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JDBCServer2x进程堆内存使用率过高，会影响JDBCServer2x进程运行的性能，甚至造成内存溢出导致JDBCServer2x进程不可用。

可能原因

该节点JDBCServer2x进程堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43010”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x内存使用率统计”，单击“确定”，查看JDBCServer2x进程使用的堆内存是否已达到JDBCServer2x进程设定的最大堆内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x进程的堆内存统计”，单击“确定”，根据告警产生时间，查看对应时间段的“JDBCServer2x进程使用的堆内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JDBCServer2x > 性能”，“SPARK_DRIVER_MEMORY”参数的值默认4G，可根据如下原则进行调整：告警时间段内JDBCServer2x使用堆内存的最大值和“JDBCServer2x堆内存使用率统计(JDBCServer2x)”阈值的比值。若参数值调整后，仍偶现告警，可以按0.5倍速率调大。若频繁出现告警，可以按1倍速率调大。多业务量、高并发的情况可以考虑增加实例。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > JDBCServer2x堆内存使用率统计 (JDBCServer2x)”，可查看“阈值”。

步骤5 重启所有的JDBCServer2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.214 ALM-43011 JDBCServer2x 进程非堆内存使用超出阈值

告警解释

系统每30秒周期性检测JDBCServer2x进程非堆内存使用状态，当检测到JDBCServer2x进程非堆内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43011	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JDBCServer2x进程非堆内存使用率过高，会影响JDBCServer2x进程运行的性能，甚至造成内存溢出导致JDBCServer2x进程不可用。

可能原因

该节点JDBCServer2x进程非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43011”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x内存使用率统计”，单击“确定”，查看JDBCServer2x进程使用的非堆内存是否已达到JDBCServer2x进程设定的最大非堆内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x进程的非堆内存统计”，单击“确定”，根据告警产生时间，查看对应时间段的“JDBCServer2x进程使用的非堆内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JDBCServer2x > 性能”，将“spark.driver.extraJavaOptions”参数中-XX:MaxMetaspaceSize的值根据如下原则调整：告警时间段内JDBCServer2x使用的非堆内存的最大值和“JDBCServer2x非堆内存使用率统计（JDBCServer2x）”阈值的比值。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > JDBCServer2x非堆内存使用率统计（JDBCServer2x）”，可查看“阈值”。

步骤5 重启所有的JDBCServer2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.215 ALM-43012 JDBCServer2x 进程直接内存使用超出阈值

告警解释

系统每30秒周期性检测JDBCServer2x进程直接内存使用状态，当检测到JDBCServer2x进程直接内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43012	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

JDBCServer2x进程直接内存使用率过高，会影响JDBCServer2x进程运行的性能，甚至造成内存溢出导致JDBCServer2x进程不可用。

可能原因

该节点JDBCServer2x进程直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43012”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x内存使用率统计”，单击“确定”，查看JDBCServer2x进程使用的直接内存是否已达到JDBCServer2x进程设定的最大直接内存的阈值。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x直接内存”，单击“确定”，根据告警产生时间，查看对应时间段的“JDBCServer2x进程使用的直接内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JDBCServer2x > 性能”，“spark.driver.extraJavaOptions”参数中-XX:MaxDirectMemorySize的默认值为512M，可根据如下方案调整：告警时间段内JDBCServer2x使用的直接内存的最大值和“JDBCServer2x直接内存使用率统计 (JDBCServer2x)”阈值的比值。若参数值调整后，仍偶现告警，可以按0.5倍速率调大。若频繁出现告警，可以按1倍速率调大。建议不要超过“SPARK_DRIVER_MEMORY”的参数值。多业务量、高并发的情况可以考虑增加实例。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > JDBCServer2x直接内存使用率统计 (JDBCServer2x)”，可查看“阈值”。

步骤5 重启所有的JDBCServer2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.216 ALM-43013 JDBCServer2x 进程 GC 时间超出阈值

告警解释

系统每60秒周期性检测JDBCServer2x进程的GC时间，当检测到JDBCServer2x进程的GC时间超出阈值（连续3次检测超过12秒）时产生该告警。用户可通过“运维 >告警 > 阈值设置 > *待操作集群的名称* > Spark2x > GC时间 > JDBCServer2x的总GC时间”修改阈值。当JDBCServer2x进程GC时间小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
43013	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC时间超出阈值，会影响JDBCServer2x进程运行的性能，甚至造成JDBCServer2x进程不可用。

可能原因

该节点JDBCServer2x进程堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43013”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x的GC时间”，单击“确定”，查看JDBCServer2x进程的GC时间是否大于阈值（默认12秒）。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JDBCServer2x > 默认”，“SPARK_DRIVER_MEMORY”参数默认值为4G，可根据如下原则调整：若参数值调整后，仍偶现告警，可按0.5倍速率调大。若告警次数比较频繁，可以按1倍速率调大。多业务量、高并发的情况可以考虑增加实例。

步骤4 重启所有的JDBCServer2x实例。

步骤5 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Spark2x”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.217 ALM-43017 JDBCServer2x 进程 Full GC 次数超出阈值

告警解释

系统每60秒周期性检测JDBCServer2x进程的Full GC次数，当检测到JDBCServer2x进程的Full GC次数超出阈值（连续3次检测超过12次）时产生该告警。用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Spark2x > GC次数 > JDBCServer2x的Full GC次数”修改阈值。当JDBCServer2x进程Full GC次数小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
43017	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC次数超出阈值，会影响JDBCServer2x进程运行的性能，甚至造成JDBCServer2x进程不可用。

可能原因

该节点JDBCServer2x进程堆内存使用率过大，或配置的堆内存不合理，导致进程Full GC频繁。

处理步骤

检查Full GC次数

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“43017”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击上报告警的JDBCServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JDBCServer2x的Full GC次数”，单击“确定”，查看JDBCServer进程的Full GC次数是否大于阈值（默认12）。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JDBCServer2x > 性能”，“SPARK_DRIVER_MEMORY”参数的默认值为4G，可根据如下原则进行调整：若偶现告警，可以按0.5倍速率调大。若告警次数比较频繁，可以按1倍速率调大。多业务量、高并发的情况可以考虑增加实例。

步骤4 重启所有的JDBCServer2x实例。

步骤5 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Spark2x”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.218 ALM-43018 JobHistory2x 进程 Full GC 次数超出阈值

告警解释

系统每60秒周期性检测JobHistory2x进程的Full GC次数，当检测到JobHistory2x进程的Full GC次数超出阈值（连续3次检测超过12次）时产生该告警。用户可通过“运维 > 阈值设置 > *待操作集群的名称* > Spark2x > GC次数 > JobHistory2x的Full GC次数”修改阈值。当JobHistory2x进程Full GC次数小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
43018	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC次数超出阈值，会影响JobHistory2x进程运行的性能，甚至造成JobHistory2x进程不可用。

可能原因

该节点JobHistory2x进程堆内存使用率过大，或配置的堆内存不合理，导致进程Full GC频繁。

处理步骤

检查Full GC次数

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“43018”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击上报告警的JobHistory2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > JobHistory2x的Full GC次数”，单击“确定”，查看JobHistory2x进程的Full GC次数是否大于阈值（默认值12）。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“JobHistory2x > 默认”，将“SPARK_DAEMON_MEMORY”参数的默认值为4G，可根据如下原则进行调整：若偶现告警，可以按0.5倍速率调大。若告警次数比较频繁，可以按1倍速率调大。

步骤4 重启所有的JobHistory2x实例。

步骤5 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Spark2x”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.219 ALM-43019 IndexServer2x 进程堆内存使用超出阈值

告警解释

系统每30秒周期性检测IndexServer2x进程堆内存使用状态，当检测到IndexServer2x进程堆内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43019	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

IndexServer2x进程堆内存使用率过高，会影响IndexServer2x进程运行的性能，甚至造成内存溢出导致IndexServer2x进程不可用。

可能原因

该节点IndexServer2x进程堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43019”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x内存使用率统计”，单击“确定”，查看IndexServer2x进程使用的堆内存是否已达到IndexServer2x进程设定的最大堆内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x进程堆内存统计”，单击“确定”，根据告警产生时间，查看对应时间段的“IndexServer2x进程使用的堆内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“IndexServer2x > 性能”，“SPARK_DRIVER_MEMORY”参数的值默认4G，可根据如下原则进行调整：告警时间段内IndexServer2x使用堆内存的最大值和“IndexServer2x堆内存使用率统计(IndexServer2x)”阈值的比值。若参数值调整后，仍偶现告警，可以按0.5倍速率调大。若频繁出现告警，可以按1倍速率调大。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > IndexServer2x堆内存使用率统计（IndexServer2x)”，可查看“阈值”。

步骤5 重启所有的IndexServer2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.220 ALM-43020 IndexServer2x 进程非堆内存使用超出阈值

告警解释

系统每30秒周期性检测IndexServer2x进程非堆内存使用状态，当检测到IndexServer2x进程非堆内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43020	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。

参数名称	参数含义
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

IndexServer2x进程非堆内存使用率过高，会影响IndexServer2x进程运行的性能，甚至造成内存溢出导致IndexServer2x进程不可用。

可能原因

该节点IndexServer2x进程非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43020”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x内存使用率统计”，单击“确定”，查看IndexServer2x进程使用的非堆内存是否已达到IndexServer2x进程设定的最大非堆内存的阈值（默认95%）。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x进程的非堆内存统计”，单击“确定”，根据告警产生时间，查看对应时间段的“IndexServer2x进程使用的非堆内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“IndexServer2x > 性能”，将“spark.driver.extraJavaOptions”参数中-XX:MaxMetaspaceSize的值根据如下原则调整：告警时间段内IndexServer2x使用的非堆内存的最大值和“IndexServer2x非堆内存使用率统计（IndexServer2x）”阈值的比值。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > IndexServer2x非堆内存使用率统计（IndexServer2x）”，可查看“阈值”。

步骤5 重启所有的IndexServer2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.221 ALM-43021 IndexServer2x 进程直接内存使用超出阈值

告警解释

系统每30秒周期性检测IndexServer2x进程直接内存使用状态，当检测到IndexServer2x进程直接内存使用率超出阈值（最大内存的95%）时产生该告警。

告警属性

告警ID	告警级别	是否自动清除
43021	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

IndexServer2x进程直接内存使用率过高，会影响IndexServer2x进程运行的性能，甚至造成内存溢出导致IndexServer2x进程不可用。

可能原因

该节点IndexServer2x进程直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43021”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x内存使用率统计”，单击“确定”，查看IndexServer2x进程使用的直接内存是否已达到IndexServer2x进程设定的最大直接内存的阈值。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x直接内存”，单击“确定”，根据告警产生时间，查看对应时间段的“IndexServer2x进程使用的直接内存”的值，获取最大值。

步骤4 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“IndexServer2x > 性能”，“spark.driver.extraJavaOptions”参数中-XX:MaxDirectMemorySize的默认值为512M，可根据如下方案调整：告警时间段内IndexServer2x使用的直接内存的最大值和“IndexServer2x直接内存使用率统计 (IndexServer2x)”阈值的比值。若参数值调整后，仍偶现告警，可以按0.5倍速率调大。若频繁出现告警，可以按1倍速率调大。

说明

在FusionInsight Manager首页，选择“运维 > 告警 > 阈值设置 > *待操作集群名称* > Spark2x > 内存 > IndexServer2x直接内存使用率统计 (IndexServer2x)”，可查看“阈值”。

步骤5 重启所有的IndexServer2x实例。

步骤6 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤7**。

收集故障信息

步骤7 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤8 在“服务”中勾选待操作集群的“Spark2x”。

步骤9 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤10 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.222 ALM-43022 IndexServer2x 进程 GC 时间超出阈值

告警解释

系统每60秒周期性检测IndexServer2x进程的GC时间，当检测到IndexServer2x进程的GC时间超出阈值（连续3次检测超过12秒）时产生该告警。用户可通过“运维 >告警 > 阈值设置 > *待操作集群的名称* > Spark2x > GC时间 > IndexServer2x的总GC时间”修改阈值。当IndexServer2x进程GC时间小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
43022	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC时间超出阈值，会影响IndexServer2x进程运行的性能，甚至造成IndexServer2x进程不可用。

可能原因

该节点IndexServer2x进程堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“ID”为“43022”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击告警上报的IndexServer2x，进入实例“概览”页面，单击图表区域右上角的下拉菜单，选择“定制 > IndexServer2x的GC时间”，单击“确定”，查看IndexServer2x进程的GC时间是否大于阈值（默认12秒）。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“IndexServer2x > 默认”，“SPARK_DRIVER_MEMORY”参数默认值为4G，可根据如下原则调整：可将“SPARK_DRIVER_MEMORY”参数调整为默认值的1.5倍；若参数值调整后，仍偶现告警，可按0.5倍速率调大。若告警次数比较频繁，可以按1倍速率调大。

步骤4 重启所有的IndexServer2x实例。

步骤5 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Spark2x”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.223 ALM-43023 IndexServer2x 进程 Full GC 次数超出阈值

告警解释

系统每60秒周期性检测IndexServer2x进程的Full GC次数，当检测到IndexServer2x进程的Full GC次数超出阈值（连续3次检测超过12次）时产生该告警。用户可通过“运维 > 告警 > 阈值设置 > *待操作集群的名称* > Spark2x > GC次数 > IndexServer2x的Full GC次数”修改阈值。当IndexServer2x进程Full GC次数小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
43023	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

GC次数超出阈值，会影响IndexServer2x进程运行的性能，甚至造成IndexServer2x进程不可用。

可能原因

该节点IndexServer2x进程堆内存使用率过大，或配置的堆内存不合理，导致进程Full GC频繁。

处理步骤

检查Full GC次数

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”，选中“告警ID”为“43023”的告警，查看“定位信息”中的角色名以及确认主机名所在的IP地址。

步骤2 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 实例”，单击上报告警的IndexServer2x，进入实例“概览”页面，单击图表区域右上

角的下拉菜单，选择“定制 > IndexServer2x的Full GC次数”，单击“确定”，查看IndexServer2x进程的Full GC次数是否大于阈值（默认12）。

- 是，执行**步骤3**。
- 否，执行**步骤6**。

步骤3 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，选择“IndexServer2x > 性能”，“SPARK_DRIVER_MEMORY”参数的默认值为4G，可根据如下原则进行调整：若偶现告警，可以按0.5倍速率调大。若告警次数比较频繁，可以按1倍速率调大。多业务量、高并发的情况可以考虑增加实例。

步骤4 重启所有的IndexServer2x实例。

步骤5 等待10分钟，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”中勾选待操作集群的“Spark2x”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.224 ALM-44004 Presto Coordinator 资源组排队任务超过阈值

告警解释

系统通过jmx接口查询资源组的排队任务数即QueuedQueries指标，当检测到资源组排队数大于阈值时产生该告警。用户可通过"组件管理 > Presto > 服务配置（将“基础配置”切换为“全部配置”） >Presto > resource-groups " 配置资源组。用户可通过"组件管理 > Presto > 服务配置（将“基础配置”切换为“全部配置”） > Coordinator > 自定义 > resourceGroupAlarm " 配置每个资源组的阈值。

告警属性

告警ID	告警级别	可自动清除
44004	严重	是

告警参数

参数名称	参数含义
ServiceName	产生告警的服务名称。
RoleName	产生告警的角色名称。
HostName	产生告警的主机名。

对系统的影响

资源组排队超过阈值可能导致大量任务处于排队状态，presto任务时间超过预期，当资源组排队数超过该组最大排队数（maxQueued）时，会导致新的任务无法执行。

可能原因

资源组配置不合理或该资源组下提交的任务过多。

处理步骤

步骤1 用户可通过“组件管理 > Presto > 服务配置（将“基础配置”切换为“全部配置”）>Presto > resource-groups”调整资源组的配置。

步骤2 用户可通过“组件管理 > Presto > 服务配置（将“基础配置”切换为“全部配置”）> Coordinator > 自定义 > resourceGroupAlarm”修改每个资源组的阈值。

步骤3 收集故障信息。

1. 根据故障信息中的HostName登录到集群节点，在presto客户端根据附加信息中的Reource Group查询排队数。
2. 根据故障信息中的HostName登录到集群节点，查看/var/log/Bigdata/nodeagent/monitorlog/monitor.log日志，搜索Resource group info可看到资源组监控采集信息。
3. 请联系运维人员，并发送已收集的故障日志信息。

----结束

参考信息

无。

10.13.225 ALM-44005 Presto Coordinator 进程垃圾收集时间超出阈值

告警解释

系统每30s周期性采集Presto Coordinator进程的垃圾收集（GC）时间，当检测到GC时间超出阈值（连续3次检测超过5s）时产生该告警。用户可在FusionInsight Manager中通过“运维 > 阈值配置 > 服务 > Presto > 集群状态 > Coordinator进程GC时间”修改阈值。当 Coordinator进程Gc时间小于或等于告警阈值时，告警清除。

告警属性

告警ID	告警级别	可自动清除
44005	严重	是

告警参数

参数名称	参数含义
ServiceName	产生告警的服务名称。
RoleName	产生告警的角色名称。
HostName	产生告警的主机名。

对系统的影响

Coordinator进程GC时间过长，会影响Coordinator进程运行的性能，甚至造成Coordinator进程不可用。

可能原因

该节点Coordinator进程堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

步骤1 检查GC时间。

1. 登录MRS集群详情页面，选择“告警管理”。
2. 选中“告警ID”为“44005”的告警，查看“定位信息”中的角色名并确定实例的IP地址。
3. 单击“组件管理 > Presto > 实例 > Coordinator（对应上报告警实例IP地址） > 定制 > Presto进程GC时间”。单击“确定”，查看GC时间。
4. 查看Coordinator进程的GC时间是否大于5秒。
   - 是，执行**步骤1.5**。

- 否，执行**步骤2**。

5. 单击“组件管理 > Presto > 服务配置 > 全部配置 > Presto > Coordinator”。将“JAVA_OPTS”参数中的最大堆内存-Xmx值根据实际情况调大。
6. 观察界面告警是否清除。
   - 是，处理完毕。
   - 否，执行**步骤2**。

步骤2 收集故障信息。

1. 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。
2. 在“服务”中勾选操作集群的“ Presto ”，单击“确定”。
3. 单击右上角的，设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后30分钟，单击“下载”
4. 请联系运维人员，并发送已收集的故障日志信息。

----结束

参考信息

无。

10.13.226 ALM-44006 Presto Worker 进程垃圾收集时间超出阈值

告警解释

系统每30s周期性采集Presto Worker进程的垃圾收集（GC）时间，当检测到GC时间超出阈值（连续3次检测超过5s）时产生该告警。用户可在FusionInsight Manager中通过“运维 > 阈值配置 > 服务 > Presto > 集群状态 > Worker进程GC时间”修改阈值。当 Worker进程GC时间小于或等于告警阈值时，告警清除。

告警属性

告警ID	告警级别	可自动清除
44006	严重	是

告警参数

参数名称	参数含义
ServiceName	产生告警的服务名称。
RoleName	产生告警的角色名称。
HostName	产生告警的主机名。

对系统的影响

Worker进程GC时间过长，会影响Worker进程运行的性能，甚至造成Worker进程不可用。

可能原因

该节点Worker进程堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

步骤1 检查GC时间。

1. 登录MRS集群详情页面，选择“告警管理”。
2. 选中“告警ID”为“44006”的告警，查看“定位信息”中的角色名并确定实例的IP地址。
3. 单击“组件管理 > Presto > 实例 > Worker（对应上报告警实例IP地址） > 定制 > Presto进程GC时间”。单击“确定”，查看GC时间。
4. 查看Worker进程的GC时间是否大于5秒。
   - 是，执行**步骤1.5**。
   - 否，执行**步骤2**。
5. 单击“组件管理 > Presto > 服务配置 > 全部配置 > Presto > Worker”。将“JAVA_OPTS”参数中的最大堆内存-Xmx值根据实际情况调大。
6. 观察界面告警是否清除。
   - 是，处理完毕。
   - 否，执行**步骤2**。

步骤2 收集故障信息。

1. 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。
2. 在“服务”中勾选操作集群的“Presto ”，单击“确定”。
3. 单击右上角的，设置日志收集的“开始时间”和“结束时间”，分别为告警产生时间的前后30分钟，单击“下载”
4. 请联系运维人员，并发送已收集的故障日志信息。

----结束

参考信息

无。

10.13.227 ALM-45175 OBS 元数据接口调用平均时间超过阈值

告警解释

系统每30秒周期性检测OBS元数据接口调用平均时间是否超过阈值，当检测到连续超过所设置阈值次数大于平滑次数时就会产生该告警 。

当OBS元数据接口调用平均时间小于阈值时，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
45175	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

OBS元数据接口调用平均时间超过阈值，会影响上层大数据计算业务的性能，导致某些计算任务的执行时间超过阈值。

可能原因

OBS服务端出现卡顿，或OBS客户端到OBS服务端之间的网络不稳定。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > OBS元数据接口调用平均时间超过阈值”，查看“定位信息”中的角色名并确定实例的IP地址。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > meta > 实例 > meta（对应上报告警实例IP地址）”。单击图表区域右上角的下拉菜单，选择“定制”，在“OBS元数据操作”中勾选“OBS接口调用平均时间”，单击“确定”，查看OBS元数据接口调用平均时间，确定是否有接口调用时间超过阈值。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 选择“集群 > *待操作集群的名称* > 运维 > 告警 > 阈值设置 > meta > OBS元数据接口调用平均时间”，将阈值或平滑次数参数的值根据实际情况调大。

步骤4 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选操作OMS下面的“NodeAgent”、“NodeMetricAgent”、“OmmServer”、“OmmAgent”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.228 ALM-45176 OBS 元数据接口调用成功率低于阈值

告警解释

系统每30秒周期性检测OBS元数据接口调用成功率是否小于阈值，当检测到小于所设置阈值时就会产生该告警 。

当OBS元数据接口调用成功率大于阈值时，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
45176	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

OBS元数据接口调用成功率小于阈值，会影响上层大数据计算业务的正常执行，导致某些计算任务的执行失败。

可能原因

OBS服务端出现执行异常或严重超时。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > OBS元数据接口调用成功率低于阈值”，查看“定位信息”中的角色名并确定实例的IP地址。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > meta > 实例 > meta（对应上报告警实例IP地址）”。单击图表区域右上角的下拉菜单，选择“定制”，在“OBS元数据操作”中勾选“OBS接口调用成功率”，单击“确定”，查看OBS元数据接口调用成功率，确定是否有接口调用成功率低于阈值。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 选择“集群 > *待操作集群的名称* > 运维 > 告警 > 阈值设置 > meta > OBS元数据接口调用成功率”，将阈值或平滑次数参数的值根据实际情况调小。

步骤4 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选操作OMS下面的“NodeAgent”、“NodeMetricAgent”、“OmmServer”、“OmmAgent”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.229 ALM-45177 OBS 数据读操作接口调用成功率低于阈值

告警解释

系统每30秒周期性检测OBS数据读操作接口调用成功率是否小于阈值，当检测到小于所设置阈值时就会产生该告警 。

当OBS数据读操作接口调用成功率大于阈值时，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
45177	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

OBS数据读操作接口调用成功率小于阈值，会影响上层大数据计算业务的正常执行，导致某些计算任务的执行失败。

可能原因

OBS服务端出现执行异常或严重超时。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > OBS数据读操作接口调用成功率低于阈值”，查看“定位信息”中的角色名并确定实例的IP地址。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > meta > 实例 > meta（对应上报告警实例IP地址）”。单击图表区域右上角的下拉菜单，选择“定制”，在“OBS数据读操作”中勾选“OBS数据读操作接口调用成功率”，单击“确定”，查看OBS数据读操作接口调用成功率，确定是否有接口调用成功率低于阈值。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 选择“集群 > *待操作集群的名称* > 运维 > 告警 > 阈值设置 > meta > OBS数据读操作接口调用成功率”，将阈值或平滑次数参数的值根据实际情况调小。

步骤4 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选操作OMS下面的“NodeAgent”、“NodeMetricAgent”、“OmmServer”、“OmmAgent”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.230 ALM-45178 OBS 数据写操作接口调用成功率低于阈值

告警解释

系统每30秒周期性检测OBS数写操作接口调用成功率是否小于阈值，当检测到小于所设置阈值时就会产生该告警 。

当OBS数据写操作接口调用成功率大于阈值时，该告警会自动清除。

告警属性

告警ID	告警级别	是否自动清除
45178	次要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

OBS数据写操作接口调用成功率小于阈值，会影响上层大数据计算业务的正常执行，导致某些计算任务的执行失败。

可能原因

OBS服务端出现执行异常或严重超时。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > OBS数据写操作接口调用成功率低于阈值”，查看“定位信息”中的角色名并确定实例的IP地址。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > meta > 实例 > meta（对应上报告警实例IP地址）”。单击图表区域右上角的下拉菜单，选择“定制”，在“OBS数据写操作”中勾选“OBS数据写操作接口调用成功率”，单击“确定”，查看OBS数据写操作接口调用成功率，确定是否有接口调用成功率低于阈值。

- 是，执行**步骤3**。
- 否，执行**步骤5**。

步骤3 选择“集群 > *待操作集群的名称* > 运维 > 告警 > 阈值设置 > meta > OBS数据写操作接口调用成功率”，将阈值或平滑次数参数的值根据实际情况调小。

步骤4 观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选操作OMS下面的“NodeAgent”、“NodeMetricAgent”、“OmmServer”、“OmmAgent”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后30分钟，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.231 ALM-45275 Ranger 服务不可用

告警解释

告警模块按180秒周期检测Ranger服务状态，当检测到Ranger服务异常时，系统产生此告警。

当系统检测到Ranger服务恢复正常，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45275	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

当Ranger服务不可用时，Ranger无法正常工作，Ranger原生UI无法访问。

可能原因

- Ranger服务所依赖内部服务DBService故障。
- RangerAdmin角色实例异常。

处理步骤

检查DBService进程状态。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”页面，查看系统是否上报“ALM-27001 DBService服务不可用”告警。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 参考“ALM-27001 DBService服务不可用”告警帮助指导对DBService服务状态异常进行处理，待DBService告警消除后，查看“Ranger服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤3**。

检查所有RangerAdmin实例。

步骤3 以**omm**用户登录RangerAdmin实例所在节点，执行**ps -ef|grep "proc_rangeradmin"**命令查看当前节点是否存在RangerAdmin进程。

- 是，执行**步骤5**。
- 否，重启RangerAdmin故障实例或Ranger服务，执行**步骤4**。

步骤4 在告警列表中查看“Ranger服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”框中勾选待操作集群的“Ranger”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.232 ALM-45276 RangerAdmin 状态异常

告警解释

告警模块按60秒周期检测RangerAdmin状态，当检测到RangerAdmin状态异常时，系统产生此告警。

当系统检测到RangerAdmin状态恢复正常，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45276	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

当存在单个RangerAdmin状态异常时，不影响Ranger原生UI访问；当两个RangerAdmin状态异常时，Ranger原生UI无法访问，无法执行创建、修改、删除策略等操作。

可能原因

RangerAdmin端口未启动。

处理步骤

端口进程检查。

步骤1 在FusionInsight Manager页面告警列表中，单击此告警所在行的，查看该告警的主机名。

步骤2 以**omm**用户登录RangerAdmin状态异常实例所在节点，执行**ps -ef|grep "proc_rangeradmin" | grep -v grep | awk -F ' ' '{print $2}'**命令获取RangerAdmin进程***pid*，再执行**netstat -anp|grep *pid* | grep LISTEN**查看RangerAdmin进程是否监听端口，安全模式集群监听21401端口，普通模式集群监听21400端口。

- 是，执行**步骤4**。
- 否，重启RangerAdmin故障实例或Ranger服务，执行**步骤3**。

步骤3 在告警列表中查看“RangerAdmin状态异常”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤4**。

收集故障信息

步骤4 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤5 在“服务”框中勾选待操作集群的“Ranger”。

步骤6 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤7 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.233 ALM-45277 RangerAdmin 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测RangerAdmin服务堆内存使用状态，当连续10次检测到RangerAdmin实例堆内存使用率超出阈值（最大内存的95%）时产生该告警，堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45277	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

堆内存溢出可能导致服务崩溃。

可能原因

该节点RangerAdmin实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

检查堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45277 RangerAdmin堆内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > RangerAdmin堆内存使用率”，单击“确定”。

步骤3 查看RangerAdmin使用的堆内存是否已达到RangerAdmin设定的阈值（默认值为最大堆内存的95%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > RangerAdmin > 实例配置”，单击“全部配置”，选择“RangerAdmin > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前RangerAdmin设置的堆内存无法满足当前RangerAdmin进程所需的堆内存，建议根据**步骤2**查看“RangerAdmin堆内存使用率”，调整“GC_OPTS”参数中“-Xmx”的值为“RangerAdmin使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.234 ALM-45278 RangerAdmin 直接内存使用率超过阈值

告警解释

系统每60秒周期性检测RangerAdmin服务直接内存使用状态，当连续5次检测到RangerAdmin实例直接内存使用率超出阈值（最大内存的80%）时，产生该告警。当RangerAdmin直接内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45278	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

直接内存溢出可能导致服务崩溃。

可能原因

节点RangerAdmin实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45278 RangerAdmin直接内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > RangerAdmin直接内存使用率”，单击“确定”。

步骤3 查看RangerAdmin使用的直接内存是否已达到RangerAdmin设定的阈值（默认值为最大直接内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > RangerAdmin > 实例配置”，单击“全部配置”，选择“RangerAdmin > 系统”。将“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前RangerAdmin设置的直接内存无法满足当前RangerAdmin进程所需的直接内存，建议根据**步骤2**查看“RangerAdmin直接内存使用率”，调整“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值为“RangerAdmin使用的直接内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重新启动受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.235 ALM-45279 RangerAdmin 非堆内存使用率超过阈值

告警解释

系统每60秒周期性检测RangerAdmin服务非堆内存使用状态，当连续5次检测到RangerAdmin实例非堆内存使用率超出阈值（最大内存的80%）时产生该告警，非堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45279	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

非堆内存溢出可能导致服务崩溃。

可能原因

该节点RangerAdmin实例非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45279 RangerAdmin非堆内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > RangerAdmin非堆内存使用率”，单击“确定”。

步骤3 查看RangerAdmin使用的非堆内存是否已达到RangerAdmin设定的阈值（默认值为最大非堆内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > RangerAdmin > 实例配置”，单击“全部配置”，选择“RangerAdmin > 系统”。将“GC_OPTS”参数中“-XX: MaxPermSize”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前RangerAdmin实例设置非堆内存大小无法满足当前RangerAdmin进程所需的非堆内存，建议调整“GC_OPTS”参数中“-XX:MaxPermSize”的值为当前非堆内存使用量的两倍（或根据实际情况进行调整）。

步骤5 重启受影响的服务或实例观察界面告警是否清除。

- 是，处理完毕。

- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.236 ALM-45280 RangerAdmin 垃圾回收(GC)时间超过阈值

告警解释

系统每60秒周期性检测RangerAdmin进程的垃圾回收（GC）占用时间，当连续5次检测到RangerAdmin进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45280	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

导致RangerAdmin响应缓慢。

可能原因

该节点RangerAdmin实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45280 RangerAdmin进程垃圾回收（GC）时间超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > GC > RangerAdmin垃圾回收（GC）时间”，单击“确定”。

步骤3 查看RangerAdmin每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > RangerAdmin > 实例配置”，单击“全部配置”，选择“RangerAdmin > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前RangerAdmin设置的堆内存无法满足当前RangerAdmin进程所需的堆内存，建议根据**步骤2**查看“RangerAdmin堆内存使用率”，调整“GC_OPTS”参数中“-Xmx”的值为“RangerAdmin使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.237 ALM-45281 UserSync 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测UserSync服务堆内存使用状态，当连续10次检测到UserSync实例堆内存使用率超出阈值（最大内存的95%）时产生该告警，堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45281	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

堆内存溢出可能导致服务崩溃。

可能原因

该节点UserSync实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45281 UserSync堆内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > UserSync堆内存使用率”，单击“确定”。

步骤3 查看UserSync使用的堆内存是否已达到UserSync设定的阈值（默认值为最大堆内存的95%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > UserSync > 实例配置”，单击“全部配置”，选择“UserSync > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前UserSync设置的堆内存无法满足当前UserSync进程所需的堆内存，建议根据**步骤2**查看“UserSync堆内存使用率”，调整“GC_OPTS”参数中“-Xmx”的值为“UserSync使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.238 ALM-45282 UserSync 直接内存使用率超过阈值

告警解释

系统每60秒周期性检测UserSync服务直接内存使用状态，当连续5次检测到UserSync实例直接内存使用率超出阈值（最大内存的80%）时，产生该告警。当UserSync直接内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45282	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

直接内存溢出可能导致服务崩溃。

可能原因

节点UserSync实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45282 UserSync直接内存使用率超过阈值”，检查该告警的“定位信息”。查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > UserSync直接内存使用率”，单击“确定”。

步骤3 查看UserSync使用的直接内存是否已达到UserSync设定的阈值（默认值为最大直接内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > UserSync >实例配置”，单击“全部配置”，选择“UserSync > 系统”。将“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前UserSync设置的直接内存无法满足当前UserSync进程所需的直接内存，建议根据**步骤2**查看“UserSync直接内存使用率”，调整“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值为“UserSync使用的直接内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重新启动受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。

- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.239 ALM-45283 UserSync 非堆内存使用率超过阈值

告警解释

系统每60秒周期性检测UserSync服务非堆内存使用状态，当连续5次检测到UserSync实例非堆内存使用率超出阈值（最大内存的80%）时产生该告警，非堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45283	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

非堆内存溢出可能导致服务崩溃。

可能原因

该节点UserSync实例非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45283 UserSync非堆内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > UserSync非堆内存使用率”，单击“确定”。

步骤3 查看UserSync使用的非堆内存是否已达到UserSync设定的阈值（默认值为最大非堆内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > UserSync > 实例配置”，单击“全部配置”，选择“UserSync > 系统”。将“GC_OPTS”参数中“-XX: MaxPermSize”的值根据实际情况调大，并单击“保存”，并保存配置。

说明

出现此告警时，说明当前UserSync实例设置非堆内存大小无法满足当前UserSync进程所需的非堆内存，建议调整“GC_OPTS”参数中“-XX:MaxPermSize”的值为当前非堆内存使用量的两倍（或根据实际情况进行调整）。

步骤5 重启受影响的服务或实例观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.240 ALM-45284 UserSync 垃圾回收(GC)时间超过阈值

告警解释

系统每60秒周期性检测UserSync进程的垃圾回收（GC）占用时间，当连续5次检测到UserSync进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45284	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

导致UserSync响应缓慢。

可能原因

该节点UserSync实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45284 UserSync进程垃圾回收（GC）时间超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > GC > UserSync垃圾回收（GC）时间”，单击“确定”。

步骤3 查看UserSync每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > UserSync > 实例配置”，单击“全部配置”，选择“UserSync > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前UserSync设置的堆内存无法满足当前UserSync进程所需的堆内存，建议根据**步骤2**查看“UserSync堆内存使用率”，调整“GC_OPTS”参数中“-Xmx”的值为“UserSync使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.241 ALM-45285 TagSync 堆内存使用率超过阈值

告警解释

系统每60秒周期性检测TagSync服务堆内存使用状态，当连续10次检测到TagSync实例堆内存使用率超出阈值（最大内存的95%）时产生该告警，堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45285	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

堆内存溢出可能导致服务崩溃。

可能原因

该节点TagSync实例堆内存使用率过大，或配置的堆内存不合理，导致使用率超过阈值。

处理步骤

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45285 TagSync堆内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > TagSync堆内存使用率”，单击“确定”。

步骤3 查看TagSync使用的堆内存是否已达到TagSync设定的阈值（默认值为最大堆内存的95%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > TagSync > 实例配置”，选择“全部配置”，选择“TagSync > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前TagSync设置的堆内存无法满足当前TagSync进程所需的堆内存，建议根据**步骤2**查看“TagSync堆内存使用率”，调整“GC_OPTS”参数中“-Xmx”的值为“TagSync使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.242 ALM-45286 TagSync 直接内存使用率超过阈值

告警解释

系统每60秒周期性检测TagSync服务直接内存使用状态，当连续5次检测到TagSync实例直接内存使用率超出阈值（最大内存的80%）时，产生该告警。当TagSync直接内存使用率小于或等于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45286	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

直接内存溢出可能导致服务崩溃。

可能原因

节点TagSync实例直接内存使用率过大，或配置的直接内存不合理，导致使用率超过阈值。

处理步骤

检查直接内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45286 TagSync直接内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > TagSync直接内存使用率”，单击“确定”。

步骤3 查看TagSync使用的直接内存是否已达到TagSync设定的阈值（默认值为最大直接内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > TagSync > 实例配置”，选择“全部配置”，选择“TagSync > 系统”。将“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前TagSync设置的直接内存无法满足当前TagSync进程所需的直接内存，建议根据**步骤2**查看“TagSync直接内存使用率”，调整“GC_OPTS”参数中“-XX:MaxDirectMemorySize”的值为“TagSync使用的直接内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重新启动受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.243 ALM-45287 TagSync 非堆内存使用率超过阈值

告警解释

系统每60秒周期性检测TagSync服务非堆内存使用状态，当连续5次检测到TagSync实例非堆内存使用率超出阈值（最大内存的80%）时产生该告警，非堆内存使用率小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45287	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

非堆内存溢出可能导致服务崩溃。

可能原因

该节点TagSync实例非堆内存使用率过大，或配置的非堆内存不合理，导致使用率超过阈值。

处理步骤

检查非堆内存使用率。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45287 TagSync非堆内存使用率超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > CPU和内存 > TagSync非堆内存使用率”，单击“确定”。

步骤3 查看TagSync使用的非堆内存是否已达到TagSync设定的阈值（默认值为最大非堆内存的80%）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务> Ranger > 实例 > TagSync > 实例配置”，单击“全部配置”，选择“TagSync > 系统”。将“GC_OPTS”参数中“-XX: MaxPermSize”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前TagSync实例设置非堆内存大小无法满足当前TagSync进程所需的非堆内存，建议调整“GC_OPTS”参数中“-XX:MaxPermSize”的值为当前非堆内存使用量的两倍（或根据实际情况进行调整）。

步骤5 重启受影响的服务或实例观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.244 ALM-45288 TagSync 垃圾回收(GC)时间超过阈值

告警解释

系统每60秒周期性检测TagSync进程的垃圾回收（GC）占用时间，当连续5次检测到TagSync进程的垃圾回收（GC）时间超出阈值（默认12秒）时，产生该告警。垃圾回收（GC）时间小于阈值时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45288	重要	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。
Trigger Condition	系统当前指标取值满足自定义的告警设置条件。

对系统的影响

导致TagSync响应缓慢。

可能原因

该节点TagSync实例堆内存使用率过大，或配置的堆内存不合理，导致进程GC频繁。

处理步骤

检查GC时间。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警 > ALM-45288 TagSync垃圾回收（GC）时间超过阈值”，检查该告警的“定位信息”，查看告警上报的实例主机名。

步骤2 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例”，选择上报告警实例主机名对应的角色，单击图表区域右上角的下拉菜单，选择“定制 > GC > TagSync垃圾回收（GC）时间”，单击“确定”。

步骤3 查看TagSync每分钟的垃圾回收时间统计值是否大于告警阈值（默认12秒）。

- 是，执行**步骤4**。
- 否，执行**步骤6**。

步骤4 在FusionInsight Manager首页，选择“集群 > 服务 > Ranger > 实例 > TagSync > 实例配置”，单击“全部配置”，选择“TagSync > 系统”。将“GC_OPTS”参数中“-Xmx”的值根据实际情况调大，并保存配置。

说明

出现此告警时，说明当前TagSync设置的堆内存无法满足当前TagSync进程所需的堆内存，建议根据**步骤2**查看“TagSync堆内存使用率”，调整“GC_OPTS”参数中“-Xmx”的值为“TagSync使用的堆内存大小”的两倍（可根据实际业务场景进行修改）。

步骤5 重启受影响的服务或实例，观察界面告警是否清除。

- 是，处理完毕。
- 否，执行**步骤6**。

收集故障信息。

步骤6 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤7 在“服务”框中勾选待操作集群的“Ranger”。

步骤8 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后10分钟，单击“下载”。

步骤9 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

10.13.245 ALM-45425 ClickHouse 服务不可用

告警解释

告警模块按60秒周期检测ClickHouse实例状态，当检测到所有ClickHouse实例异常时，系统产生此告警。

当系统检测到任一ClickHouse实例恢复正常，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45425	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称
服务名	产生告警的服务名称
角色名	产生告警的角色名称
主机名	产生告警的主机名

对系统的影响

ClickHouse服务异常，无法通过FusionInsight Manager对ClickHouse进行集群操作，无法使用ClickHouse服务功能。

可能原因

ClickHouse故障实例节点其组件配置目录下的**metrika.xml**配置信息和ZooKeeper中对应ClickHouse实例配置不一致。

处理步骤

检查ClickHouse实例metrika.xml配置是否正常

步骤1 登录FusionInsight Manager，选择“集群 > 服务 > ClickHouse > 实例”，根据告警信息找到状态异常的ClickHouse实例。

- 是，执行**步骤2**。
- 否，执行**步骤9**。

步骤2 登录ClickHouse服务异常的实例主机节点，并通过**ping**其他正常ClickHouse实例节点IP的方式进行网络是否互通验证。

- 是，执行**步骤3**。
- 否，联系网络管理员修复网络。

步骤3 选择“集群 > 服务 > ClickHouse > 实例”，在“角色”列下面单击对应异常的实例名称，选择“实例配置”，搜索框中搜索“macros.id”，找到当前实例macros.id对应的值。

步骤4 登录ZooKeeper客户端所在主机节点，执行以下命令登录ZooKeeper客户端工具。

切换到客户端安装目录。

例如：cd /opt/client

执行以下命令配置环境变量。

source bigdata_env

执行以下命令进行用户认证。(普通模式跳过此步骤)

kinit *组件业务用户*

执行以下命令登录客户端工具。

zkCli.sh -server *ZooKeeper角色实例所在节点业务IP*: *clientPort*

步骤5 使用如下命令检查ClickHouse集群拓扑信息是否能正常获取到。

get /clickhouse/config/步骤3*中的macros.id对应的值*/metrika.xml

- 是，执行**步骤6**。
- 否，不能正常获取则执行**步骤9**。

步骤6 登录ClickHouse服务异常的实例主机节点，进入当前ClickHouse实例配置目录。

cd ${BIGDATA_HOME}**/FusionInsight_ClickHouse**_*版本号*/x_x_**ClickHouseServer/etc**

cat metrika.xml

步骤7 检查**步骤5**中获取的ZooKeeper上的集群拓扑信息是否与**步骤6**中组件配置目录下的metrika.xml是否一致。

- 是，如果确认告警还未恢复则执行**步骤9**。
- 否，执行**步骤8**。

步骤8 在FusionInsight Manager首页，选择“集群 > 服务 > ClickHouse > 更多 > 同步配置”，等待五分钟，查看服务状态是否良好，告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤9**。

收集故障信息

步骤9 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤10 在“服务”中勾选待操作集群的“ClickHouse”。

步骤11 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤12 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.246 ALM-45426 ClickHouse 服务在 ZooKeeper 的数量配额使用率超过阈值

告警解释

告警模块按60秒周期检测ClickHouse服务在ZooKeeper的数量配额使用百分比，当检测到使用百分比超过阈值（90%），系统产生此告警。

当系统检测到使用百分比低于阈值，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45426	重要（默认级别）	是

告警参数

参数名称	参数含义
来源	产生告警的集群或系统名称

参数名称	参数含义
服务名	产生告警的服务名称
角色名	产生告警的角色名称
主机名	产生告警的主机名

对系统的影响

ClickHouse在ZooKeeper的数量配额超过阈值后，无法通过FusionInsight Manager对ClickHouse进行集群操作，无法使用ClickHouse服务功能。

可能原因

ClickHouse在使用过程中，如表创建、插入或删除表数据等操作时，ClickHouse会在ZooKeeper的节点中创建znode，随着业务量的增加该znode实际数量可能会超过配置的阈值。

处理步骤

检查ClickHouse在ZooKeeper的znode节点创建数量

步骤1 登录ZooKeeper客户端所在主机节点，执行以下命令登录ZooKeeper客户端工具。

切换到客户端安装目录。

例如：**cd /opt/client**

执行以下命令配置环境变量。

source bigdata_env

执行以下命令进行用户认证。(普通模式跳过此步骤)

kinit *组件业务用户*

执行以下命令登录客户端工具。

zkCli.sh -server *ZooKeeper角色实例所在节点业务IP*: *clientPort*

步骤2 执行如下命令查看ZooKeeper上ClickHouse使用的配额情况，计算返回的结果中Output stat的count值与Output quota的count值之比是否大于0.9。

listquota /clickhouse

```
absolute path is /zookeeper/quota/clickhouse
Output quota for /clickhouse count=200000,bytes=1000000000
Output stat for /clickhouse count=2667,bytes=60063
```

如上，Output stat对应的count为：2667，Output quota的count为：200000。

- 是，执行**步骤4**。
- 否，等待五分钟查看告警是否清除，如果还没有清除请执行**步骤5**。

步骤3 在FusionInsight Manager首页，选择“集群 > 服务 > ClickHouse > 配置 > 全部配置”，搜索“clickhouse.zookeeper.quota.node.count”参数，将该参数的值调整为**步骤2**中Output stat的count值的2倍。

步骤4 重启告警信息对应的ClickHouse实例，等待五分钟，查看告警是否消除。

- 是，处理完毕。
- 否，再次执行**步骤4**，等待五分钟，查看告警是否消除，如果还没有清除请执行**步骤5**。

收集故障信息

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选待操作集群的“ClickHouse”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.247 ALM-45427 ClickHouse 服务在 ZooKeeper 的容量配额使用率超过阈值

告警解释

告警模块按60秒周期检测ClickHouse服务在ZooKeeper的容量配额使用百分比，当检测到使用百分比超过阈值（90%），系统产生此告警。

当系统检测到使用百分比低于阈值，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45427	重要（默认级别）	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称
服务名	产生告警的服务名称
角色名	产生告警的角色名称

参数名称	参数含义
主机名	产生告警的主机名

对系统的影响

ClickHouse在ZooKeeper的容量配额超过阈值后，无法通过FusionInsight Manager对ClickHouse进行集群操作，无法使用ClickHouse服务功能。

可能原因

ClickHouse在使用过程中，如表创建、插入或删除表数据等操作时，ClickHouse会在ZooKeeper的节点中创建znode，随着业务量的增加该znode实际容量可能会超过配置的阈值。

处理步骤

检查ClickHouse在ZooKeeper的znode节点容量值

步骤1 登录ZooKeeper客户端所在主机节点，执行以下命令登录ZooKeeper客户端工具。

切换到客户端安装目录。

例如：**cd /opt/client**

执行以下命令配置环境变量。

source bigdata_env

执行以下命令进行用户认证。(普通模式跳过此步骤)

kinit *组件业务用户*

执行以下命令登录客户端工具。

zkCli.sh -server *ZooKeeper角色实例所在节点业务IP*:*clientPort*

步骤2 执行如下命令查看ZooKeeper上ClickHouse使用的配额情况，计算返回的结果中Output stat的bytes值与Output quota的bytes值之比是否大于0.9。

listquota /clickhouse

```
absolute path is /zookeeper/quota/clickhouse
Output quota for /clickhouse count=200000,bytes=1000000000
Output stat for /clickhouse count=2667,bytes=60063
```

如上，Output stat对应的bytes为：60063，Output quota的bytes为：1000000000。

- 是，执行**步骤4**。
- 否，等待五分钟查看告警是否清除，若还未消除，执行**步骤5**。

步骤3 在FusionInsight Manager首页，选择“集群 > 服务 > ClickHouse > 配置 > 全部配置”，搜索“clickhouse.zookeeper.quota.size”参数，将该参数的值调整为**步骤2**中Output stat的bytes值的2倍。

步骤4 重启告警信息对应的ClickHouse实例，等待五分钟，查看告警是否消除。

- 是，处理完毕。

- 否，再次执行步骤4，等待五分钟，查看告警是否消除，若还未消除，执行步骤5。

收集故障信息

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”中勾选待操作集群的“ClickHouse”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无

10.13.248 ALM-45736 Guardian 服务不可用

告警解释

告警模块按60秒周期检测Guardian服务状态，当检测到Guardian服务异常时，系统产生此告警。

当系统检测到Guardian服务恢复正常，且告警处理完成时，告警恢复。

告警属性

告警ID	告警级别	是否自动清除
45275	紧急	是

告警参数

参数名称	参数含义
来源	产生告警的集群名称。
服务名	产生告警的服务名称。
角色名	产生告警的角色名称。
主机名	产生告警的主机名。

对系统的影响

当Guardian服务不可用时，Guardian无法正常工作。

可能原因

- Guardian服务所依赖内部服务Ranger或者HDFS故障。
- TokenServer角色实例异常。

处理步骤

检查Ranger和HDFS服务状态。

步骤1 在FusionInsight Manager首页，选择“运维 > 告警 > 告警”页面，查看系统是否上报“ALM-45275 Ranger服务不可用”或者“ALM-14000 HDFS服务不可用”告警。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 参考“ALM-45275 Ranger服务不可用”或“ALM-14000 HDFS服务不可用”告警帮助指导处理对应告警。

告警全部恢复后，等待几分钟，检查本告警是否恢复。

- 是，处理完毕。
- 否，执行**步骤3**。

检查所有TokenServer实例。

步骤3 以**omm**用户登录TokenServer实例所在节点，执行**ps -ef|grep "ranger-obs-service"**命令查看当前节点是否存在TokenServer进程。

- 是，执行**步骤5**。
- 否，重启TokenServer故障实例，执行**步骤4**。

步骤4 在告警列表中查看“Guardian服务不可用”告警是否清除。

- 是，处理完毕。
- 否，执行**步骤5**。

收集故障信息。

步骤5 在FusionInsight Manager界面，选择“运维 > 日志 > 下载”。

步骤6 在“服务”框中勾选待操作集群的“Guardian”。

步骤7 单击右上角的设置日志收集的“开始时间”和“结束时间”分别为告警产生时间的前后1小时，单击“下载”。

步骤8 请联系运维人员，并发送已收集的故障日志信息。

----结束

告警清除

此告警修复后，系统会自动清除此告警，无需手工清除。

参考信息

无。

11 MRS Manager 操作指导（适用于 2.x 及之前）

11.1 MRS Manager 简介

概述

MRS为用户提供海量数据的管理及分析功能，快速从结构化和非结构化的海量数据中挖掘您所需要的价值数据。开源组件结构复杂，安装、配置、管理过程费时费力，MRS Manager提供了企业级的大数据集群的统一管理平台：

- 提供集群状态的监控功能，您能快速掌握服务及主机的健康状态。
- 提供图形化的指标监控及定制，您能及时的获取系统的关键信息。
- 提供服务属性的配置功能，满足您实际业务的性能需求。
- 提供集群、服务、角色实例的操作功能，满足您一键启停等操作需求。

系统界面简介

MRS Manager提供统一的集群管理平台，帮助用户快捷、直观的完成集群的运行维护。MRS Manager请参考**访问MRS Manager（MRS 2.x及之前版本）**页面访问。

各操作入口的详细功能如**表11-1**所示。

表 11-1 界面操作入口功能描述

界面	功能描述
系统概览	提供柱状图、折线图、表格等多种图表方式展示所有服务的状态、各服务的主要监控指标、主机的状态统计。用户可以定制关键监控信息面板，并拖动到任意位置。系统概览支持数据自动刷新。
服务管理	提供服务监控、服务操作向导以及服务配置，帮助用户对服务进行统一管理。

界面	功能描述
主机管理	提供主机监控、主机操作向导，帮助用户对主机进行统一管理。
告警管理	提供告警查询、告警处理指导功能。帮助用户及时发现产品故障及潜在隐患，并进行定位排除，以保证系统正常运行。
审计管理	提供审计日志查询及导出功能。帮助用户查阅所有用户活动及操作。
租户管理	提供统一租户管理平台。
系统设置	用户可以进行监控和告警配置管理、备份管理。

当用户进入到“系统设置”的各子功能页面后，提供快捷方式跳转到其他System子功能页面，如表11-2所示。

快捷跳转操作示例如下所示。

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“系统设置”界面，任意单击一个功能链接，进入具体功能界面。

例如在“备份恢复管理”区域中单击“备份管理”，进入到“备份管理”界面。

步骤3 将鼠标移动到浏览器窗口的左边界，弹出“系统设置”黑色快捷菜单。鼠标移出该菜单后，该菜单收起。

步骤4 在弹出的快捷菜单上，可以单击某个功能链接直接跳转到对应的功能界面。

例如选择“维护 > 日志导出”，进入“日志导出”界面。

----结束

表 11-2 集群的 System 快捷菜单

菜单子标题	功能链接
备份恢复管理	备份管理
	恢复管理
维护	日志导出
	审计日志导出
	健康检查
监控和告警配置	Syslog配置
	阈值管理
	SNMP配置
	监控指标转储配置

菜单子标题	功能链接
	资源贡献排名配置
权限配置	用户管理
	用户组管理
	角色管理
	密码策略配置
	OMS数据库密码修改
补丁管理	补丁管理

参考信息

MapReduce服务是一项数据分析服务，用于海量数据的管理和分析，简称MRS。

MRS通过MRS Manager管理大数据组件，例如Hadoop生态体系中的组件。因此，MRS和MRS Manager管理界面上的部分概念需要区别，具体解释如表11-3：

表 11-3 差异对比参考

名词概念	MRS	MRS Manager
MapReduce服务	表示数据分析云服务，简称为MRS，包括Hive、Spark、Yarn、HDFS和ZooKeeper等组件。	为租户集群中的大数据组件提供的统一管理平台。

11.2 查看集群运行任务

操作场景

用户在MRS Manager进行操作触发运行任务时，会显示任务运行的过程与进度。关闭任务窗口后，需要通过任务管理功能，打开任务窗口。

MRS Manager默认保留10个最近运行的任务。例如重启服务、同步服务配置和执行健康检查。

操作步骤

步骤1 在MRS Manager，单击  ，打开“任务列表”。

“任务列表”可查看的信息包含：“任务名”、“状态”、“进度”、“开始时间”和“结束时间”。

步骤2 单击指定的任务名称，可查看任务执行过程中的详细信息。

----结束

11.3 监控管理

11.3.1 系统概览

MRS Manager支持将集群中所有部署角色的节点，按管理节点、控制节点和数据节点进行分类，分别计算关键主机监控指标在每类节点上的变化趋势，并在报表中按用户自定义的周期显示分布曲线图。如果一个主机属于多类节点，那么对应的指标将被统计多次。

该任务指导用户了解MRS集群的概览、及在MRS Manager查看、自定义与导出节点监控指标报表。

操作步骤

步骤1 登录MRS Manager，具体请参考**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager选择“系统概览”。

步骤3 在“时间区间”选择需要查看监控数据的时间段。可供选择的选项如下：

- 实时
- 最近3小时
- 最近6小时
- 最近24小时
- 最近一周
- 最近一个月
- 最近三个月
- 最近六个月
- 自定义：选择自定义时，在时间范围内自行选择需要查看的时间。

步骤4 单击“查看”可以查看相应时间区间的监控数据。

- MRS Manager在“服务概览”显示各个服务的“健康状态”和“角色数”。
- 单击曲线图表上侧的  图标，可显示具体的指标说明信息。

步骤5 自定义监控指标报表。

1. 单击“定制”，勾选需要在MRS Manager显示的监控指标。
   MRS Manager支持统计的指标共14个，界面最多显示12个定制的监控指标。
   - 集群主机健康状态统计
   - 集群网络读速率统计
   - 主机网络读速率分布
   - 主机网络写速率分布
   - 集群磁盘写速率统计
   - 集群磁盘占用率统计
   - 集群磁盘信息
   - 主机磁盘占用率统计

- 集群磁盘读速率统计
- 集群内存占用率统计
- 主机内存占用率分布
- 集群网络写速率统计
- 主机CPU占用率分布
- 集群CPU占用率统计

2. 单击“确定”保存并显示所选指标。

   说明

   单击“清除”可批量取消全部选中的指标项。

步骤6 用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

说明

勾选“全屏”会将“系统概览”窗口最大化。

步骤7 导出监控指标报表。

1. 选择报表的时间范围。可供选择的选项如下：
   - 实时
   - 最近3小时
   - 最近6小时
   - 最近24小时
   - 最近一周
   - 最近一个月
   - 最近三个月
   - 最近六个月
   - 自定义：选择自定义时，自行选择需要导出报表的时间。
2. 单击“导出”，Manager将生成指定时间范围内、已勾选的集群监控指标报表文件，请选择一个位置保存，并妥善保管该文件。

   说明

   如果需要查看指定时间范围的监控指标对应的分布曲线图，请单击“查看”，界面将显示用户自定义时间范围内选定指标的分布曲线图。

----结束

11.3.2 管理服务和主机监控

用户可以在日常使用中，可以在MRS Manager管理所有服务（含角色实例）和主机的状态及指标信息：

- 状态信息，包括运行、健康、配置及角色实例状态统计。
- 指标信息，各服务的主要监控指标项。
- 导出监控指标。

说明

用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

管理服务监控

步骤1 在MRS Manager，单击“服务管理”。

服务列表中标题包含“服务”、“操作状态”、“健康状态”、“配置状态”、“角色数”和“操作”。

- 服务操作状态描述如表11-4所示。

表 11-4 服务操作状态

状态	描述
已启动	服务已启动。
已停止	服务已停止。
启动失败	用户启动操作失败。
停止失败	用户停止操作失败。
未知	后台系统重启后，服务的初始状态。

- 服务健康状态如表11-5所示。

表 11-5 服务健康状态

状态	描述
良好	该服务中所有角色实例正常运行。
故障	至少一个角色实例运行状态为“故障”或被依赖的服务状态不正常。
未知	该服务中所有角色实例状态为“未知”。
正在恢复	后台系统正在尝试自动启动服务。
亚健康	该服务所依赖的服务状态不正常，异常服务的相关接口无法被外部调用。

- 服务配置状态如表11-6所示。

  表 11-6 服务配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
过期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
同步中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“服务”列按升序排列，单击**服务**、**操作状态**、**健康状态**或**配置状态**可修改排列方式。

步骤2 单击列表中指定服务名称，查看服务状态及指标信息。

步骤3 定制、导出监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。
3. 单击“导出”，导出当前查看的指标数据。

----结束

管理角色实例监控

步骤1 在MRS Manager，单击“服务管理”，在服务列表中单击服务指定名称。

步骤2 单击“实例”，查看角色状态。

角色实例列表中包含实例信息的**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**操作状态**、**健康状态**及**配置状态**。

- 角色实例的状态如表11-7所示。

  表 11-7 角色实例状态

状态	描述
已启动	角色实例已启动。
已停止	角色实例已停止。
启动失败	用户启动操作失败。
停止失败	用户停止操作失败。
退服中	角色实例正在退服。

状态	描述
已退服	角色实例已退服。
入服中	角色实例正在入服。
未知	后台系统重启后，角色实例的初始状态。

- 角色实例的健康状态如表11-8所示。

  表 11-8 角色实例健康状态

状态	描述
良好	该角色实例正常运行。
故障	该角色实例运行异常，如PID不存在，无法访问端口。
未知	角色实例所在主机与后台系统未连接。
正在恢复	后台系统正在尝试自动启动角色实例。

- 角色实例的配置状态如表11-9所示。

  表 11-9 角色实例配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
过期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
同步中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“角色”列按升序排列，单击**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**操作状态**、**健康状态**或**配置状态**可修改排列方式。

支持在“角色”筛选相同角色的全部实例。

单击“高级搜索”，在角色搜索区域中设置搜索条件，单击“搜索”，查看指定的角色信息。单击“重置”清除输入的搜索条件。支持模糊搜索条件的部分字符。

步骤3 单击列表中指定角色实例名称，查看角色实例状态及指标信息。

步骤4 定制、导出监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。

3. 单击“导出”，导出当前查看的指标数据。

----结束

管理主机监控

步骤1 在MRS Manager，单击“主机管理”，看所有主机状态。

主机列表中包括主机名称、管理IP、业务IP、机架、网络速度、操作状态、健康状态、磁盘使用率、内存使用率、CPU使用率。

- 主机操作状态如**表11-10**所示。

  表 11-10 主机操作状态

状态	描述
正常	主机及主机上的服务角色正常运行。
已隔离	主机被用户隔离，主机上的服务角色停止运行。

- 主机健康状态描述如**表11-11**所示。

  表 11-11 主机健康状态

状态	描述
良好	主机心跳检测正常。
故障	主机心跳超时未上报。
未知	执行添加操作时，主机的初始状态。

默认以“主机名称”列按升序排列，单击**主机名称**、**管理IP**、**业务IP**、**机架**、**网络速度**、**操作状态**、**健康状态**、**磁盘使用率**、**内存使用率**或**CPU使用率**可修改排列方式。

单击“高级搜索”，在搜索区域中，设置查询条件，单击“搜索”，查看指定的主机。单击“重置”清除输入的搜索条件。支持模糊搜索条件的部分字符。

步骤2 单击列表中指定的主机名称，查看单个主机状态及指标。

步骤3 定制、导出监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。
3. 单击“导出”，导出当前查看的指标数据。

----结束

11.3.3 管理资源分布

用户需要了解服务和主机关键监控指标中最高、最低或平均监控数据形成的曲线，即资源分布情况时，可以在MRS Manager上查看，支持查询1小时以内的监控数据。

用户也可以在MRS Manager上修改资源分布，使服务和主机的资源分布图表中，可以按自定义的数值显示一条或多条最高、最低监控数据形成的曲线。

部分监控指标的资源分布不记录。

操作步骤

- 查看服务监控指标的资源分布
  a. 在MRS Manager，单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“资源贡献排名”。
     “指标”中选择服务的关键指标，MRS Manager将显示过去1小时内指标的资源分布情况。
- 查看主机监控指标的资源分布
  a. 单击“主机管理”。
  b. 单击主机列表中指定的主机名称。
  c. 单击“资源贡献排名”。
     “指标”中选择主机的关键指标，MRS Manager将显示过去1小时内指标的资源分布情况。
- 配置资源分布
  a. 在MRS Manager，单击“系统设置”。
  b. 在“配置”区域“监控和告警配置”下，单击“资源贡献排名配置”。
  c. 修改资源分布的显示数量。
     - “TOP数量”填写最大值的显示数量。
     - “BOTTOM数量”填写最小值的显示数量。

     说明

     最大值与最小值的资源分布显示数量总和不能大于5。
  d. 单击“确定”保存设置。
     界面右上角提示“保存TOP数量和BOTTOM数量成功。”。

11.3.4 配置监控指标转储

用户可以在MRS Manager界面上配置监控指标数据对接参数，使集群内各监控指标数据通过FTP或SFTP协议保存到指定的FTP服务器，与第三方系统进行对接。FTP协议未加密数据可能存在安全风险，建议使用SFTP。

MRS Manager支持采集当前管理的集群内所有监控指标数据，采集的周期有30秒、60秒和300秒三种。监控指标数据在FTP服务器保存时，会根据采集周期分别保存在不同的监控文件中，监控文件命名规则为“集群名称_metric_监控指标数据采集的周期_文件保存时间.log”。

前提条件

转储服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问转储服务器的IP地址和指定端口。转储服务器的FTP服务正常。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“监控指标转储配置”。

步骤3 表11-12介绍转储参数。

表 11-12 转储参数

参数名称	参数说明
FTP IP地址	必选参数，指定监控指标数据对接后存放监控文件的FTP服务器。
FTP端口	必选参数，指定连接FTP服务器的端口。
FTP用户名	必选参数，指定登录FTP服务器的用户名。
FTP密码	必选参数，指定登录FTP服务器的密码。
保存路径	必选参数，指定监控文件在FTP服务器保存的路径。
转储时间间隔（秒）	必选参数，指定监控文件在FTP服务器保存的周期，单位为秒。
转储模式	必选参数，指定监控文件发送时使用的协议。可选协议为“FTP”和“SFTP”。
SFTP服务公钥	可选参数，指定FTP服务器的公共密钥，“模式”选择“SFTP”时此参数生效。建议配置公共密钥，否则可能存在安全风险。

步骤4 单击“确定”，设置完成。

----结束

11.4 告警管理

11.4.1 查看与手动清除告警

操作场景

用户可以在MRS Manager查看、清除告警。

一般情况下，告警处理后，系统自动清除该条告警记录。当告警不具备自动清除功能且用户已确认该告警对系统无影响时，可手动清除告警。

在MRS Manager界面可查看最近十万条告警（包括未清除的、手动清除的和自动清除的告警）。如果已清除告警超过十万条达到十一万条，系统自动将最早的一万条已清除告警转存，转存路径为主管理节点“${BIGDATA_HOME}/OMSV100R001C00x8664/workspace/data”。第一次转存告警时自动生成目录。

说明

用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

操作步骤

步骤1 在MRS Manager，单击“告警管理”，在告警列表查看告警信息。

- 告警列表每页默认显示最近的十条告警。
- 默认以“产生时间”列按降序排列，单击“告警ID”、“告警名称”、“告警级别”、“产生时间”、“定位信息”或“操作”可修改排列方式。
- 支持在“告警级别”筛选相同级别的全部告警。结果包含已清除和未清除的告警。
- 分别单击 、 、 或 可以快速筛选级别为“致命”、“严重”、“一般”或“警告”的告警。

步骤2 单击“高级搜索”显示告警搜索区域，设置查询条件后，单击“搜索”，查看指定的告警信息。单击“重置”清除输入的搜索条件。

说明

“开始时间”和“结束时间”表示时间范围的开始时间和结束时间，可以搜索此时间段内产生的告警。

查看“告警参考”章节告警帮助，按照帮助指导处理告警。如果某些场景中告警由于MRS依赖的其他云服务产生，可能需要联系对应云服务运维人员处理。

步骤3 处理完告警后，若需手动清除，单击“清除告警”，手动清除告警。

说明

如果有多个告警已完成处理，可选中一个或多个待清除的告警，单击“清除告警”，批量清除告警。每次最多批量清除300条告警。

----结束

11.4.2 配置监控与告警阈值

操作场景

配置监控与告警阈值用于关注各指标的健康情况。勾选“发送告警”后，当监控数据达到告警阈值，系统将会触发一条告警信息，将在“告警管理”中出现此告警信息。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“阈值配置”，依据规划选择监控指标并设置其基线。

步骤3 单击某一指标例如“CPU使用率”，单击“添加规则”。

步骤4 在“配置”对话框中填写监控指标规则参数。

表 11-13 监控指标规则参数

参数名	参数值	参数解释
规则名称	CPU_MAX（举例）	规则名称
参考日期	2014/11/06（举例）	查看某指标的历史参考数据
阈值类型	• 最大值 • 最小值	选择某指标的最大值或最小值，类型为“最大值”表示指标的实际值大于设置的阈值时系统将产生告警，类型为“最小值”表示指标的实际值小于设置的阈值时系统将产生告警。
告警级别	• 致命 • 严重 • 一般 • 提示	告警级别
时间范围	从00:00到23:59（举例）	设置规则生效时监控指标的具体时间段
阈值	设置数值 80（举例）	设置规则监控指标的阈值
日期	• 工作日 • 周末 • 其它	设置规则生效的日期类型
添加日期	11/06（举例）	日期选择“其他”时该参数生效。可选择多个日期。

步骤5 单击“确定”。界面右上角弹出提示“模板保存成功。”。

“发送告警”默认已勾选。Manager会检查监控指标数值是否满足阈值条件，若连续检查且不满足的次数等于“平滑次数”设置的值则发送告警，支持自定义。“检查周期(秒)”表示Manager检查监控指标的时间间隔。

步骤6 在新添加规则所在的行，单击“操作”下的“应用”，界面右上角弹出提示规则xx应用成功，完成添加。单击“操作”下的“取消应用”，界面右上角弹出提示规则xx取消成功。

----结束

11.4.3 配置 Syslog 北向参数

操作场景

该任务指导用户以Syslog方式将MRS Manager的告警事件上报到指定的监控运维系统中。

须知

Syslog协议未做加密，传输数据容易被窃取，存在安全风险。

前提条件

对接服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问对接服务器的IP地址和指定端口。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“Syslog配置”。

“Syslog服务”的开关默认为关闭，单击启用Syslog服务。

步骤3 设置表11-14所示的对接参数。

表 11-14 对接参数

参数区域	参数名称	参数说明
Syslog协议	服务IP	设置对接服务器IP地址。
	服务端口	设置对接端口。
	协议	设置协议类型，取值范围： ● “TCP” ● “UDP”

参数区域	参数名称	参数说明
	安全级别	设置上报消息的严重程度，取值范围： ● “Informational” ● “Emergency” ● “Alert” ● “Critical” ● “Error” ● “Warning” ● “Notice” ● “Debug”
	Facility	设置产生日志的模块。
	标识符	设置产品标识，默认为“MRS Manager”。
报告信息	报文格式	设置告警报告的消息格式，具体要求请参考界面帮助。
	报告告警类型	设置需要上报的告警类型。 ● “故障”表示Manager产生告警时会上报Syslog告警消息。 ● “清除”表示清除Manager告警时会上报Syslog告警消息。 ● “事件”表示Manager产生事件时会上报Syslog告警消息。
	报告告警级别	设置需要上报的告警级别。支持“提示”、“一般”、“严重”和“致命”。
未恢复告警上报设置	周期上报未恢复告警	设置是否按指定周期上报未清除的告警。“周期上报未恢复告警”的开关默认为关闭，单击启用此功能。

参数区域	参数名称	参数说明
	间隔时间（分钟）	设置周期上报未恢复告警到远程Syslog服务的时间间隔，当“周期上报未恢复告警”开关打开时启用。单位为分钟，默认值为“15”，取值范围为5分钟到一天（1440分钟）。
心跳设置	上报心跳	设置是否开启周期上报Syslog心跳消息。“周期上报未恢复告警”的开关默认为关闭，单击启用此功能。
	心跳周期（分钟）	设置周期上报心跳的时间间隔，当“上报心跳”开关打开时启用。单位为分钟，默认值为“15”，取值范围为1-60。
	心跳报文	设置心跳上报的内容，当“上报心跳”开关打开时启用，不能为空。支持数字、字母、下划线、竖线、冒号、空格、英文逗号和句号等字符，长度小于等于256。

说明

设置周期上报心跳报文后，在某些集群容错自动恢复的场景下（例如主备管理节点倒换）可能会出现报文上报中断的现象，此时等待自动恢复即可。

步骤4 单击“确定”，设置完成。

----结束

11.4.4 配置 SNMP 北向参数

操作场景

该任务指导用户采用SNMP协议把MRS Manager的告警、监控数据集成到网管平台。

前提条件

对接服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问对接服务器的IP地址和指定端口。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“SNMP配置”。

“SNMP服务”的开关默认为关闭，单击启用SNMP服务。

步骤3 设置表11-15所示的对接参数。

表 11-15 对接参数

参数名称	参数说明
版本	SNMP协议版本号，取值范围： ● v2c：低版本，安全性较低 ● v3：高版本，安全性比v2c高 推荐使用v3版本。
本地端口	本地端口，默认值“20000”，取值范围“1025”到“65535”。
读团体名	该参数仅在设置“版本”为v2c时存在，用于设置只读团体名。
写团体名	该参数仅在设置“版本”为v2c时存在，用于设置可写团体名。
安全用户名	该参数仅在设置“版本”为v3时存在，用于设置协议安全用户名。
认证协议	该参数仅在设置“版本”为v3时存在，用于设置认证协议，推荐选择SHA。
认证密码	该参数仅在设置“版本”为v3时存在，用于设置认证密钥。
确认认证密码	该参数仅在设置“版本”为v3时存在，用于确认认证密钥。
加密协议	该参数仅在设置“版本”为v3时存在，用于设置加密协议，推荐选择AES256。
加密密码	该参数仅在设置“版本”为v3时存在，用于设置加密密钥。
确认加密密码	该参数仅在设置“版本”为v3时存在，用于确认加密密钥。

说明

- “认证密码”和“加密密码”密码长度为8到16位，至少需要包含大写字母、小写字母、数字、特殊字符中的3种类型字符。两个密码不能相同。两个密码不可和安全用户名或安全用户名的逆序字符相同。
- 使用SNMP协议从安全方面考虑，需要定期修改“认证密码”和“加密密码”密码。
- 使用SNMP v3版本时，安全用户在5分钟之内连续鉴权失败5次将被锁定，5分钟后自动解锁。

步骤4 单击“Trap目标”下的“添加Trap目标”，在弹出的“添加Trap目标”对话框中填写以下参数：

- 目标标识：Trap目标标识，一般指接收Trap的网管或主机标识。长度限制1～255字节，一般由字母或数字组成。
- 目标IP：目标IP。可使用A、B、C类IP地址，要求可与管理节点的管理平面IP地址互通。
- 目标端口：接收Trap的端口，要求与对端保持一致，取值范围“0”～“65535”。
- Trap团体名：该参数仅在设置Version为v2c时存在，用于设置主动上报团体名。

单击“确定”，设置完成，退出“添加Trap目标”对话框。

步骤5 单击“确定”，设置完成。

----结束

11.5 对象管理

11.5.1 对象管理简介

MRS集群包含了各类不同的基本对象，不同对象的描述介绍如表11-16所示：

表 11-16 MRS 基本对象概览

对象	描述	举例
服务	可以完成具体业务的一类功能集合。	例如KrbServer服务和LdapServer服务。
服务实例	服务的具体实例，一般情况下可使用服务表示。	例如KrbServer服务。
服务角色	组成一个完整服务的一类功能实体，一般情况下可使用角色表示。	例如KrbServer由KerberosAdmin角色和KerberosServer角色组成。
角色实例	服务角色在主机节点上运行的具体实例。	例如运行在Host2上的KerberosAdmin，运行在Host3上的KerberosServer。
主机	一个弹性云服务器，可以运行Linux系统。	例如Host1～Host5。
机架	一组包含使用相同交换机的多个主机集合的物理实体。	例如Rack1，包含Host1～Host5。
集群	由多台主机组成的可以提供多种服务的逻辑实体。	例如名为Cluster1的集群由（Host1～Host5）5个主机组成，提供了KrbServer和LdapServer等服务。

11.5.2 查看配置

用户可以在MRS Manager上查看服务（含角色）和角色实例的配置。

操作步骤

- 查看服务的配置。
  a. 在MRS Manager，单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“服务配置”。
  d. 在“参数类别”选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
  e. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
     在服务节点下的参数属于服务配置参数，在角色节点下的参数是角色配置参数。
  f. 在“非默认”选项中选择“非默认”，界面上显示参数值为非默认值的参数。
- 查看角色实例的配置。
  a. 在MRS Manager，单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“实例”页签。
  d. 单击角色实例列表中指定的角色实例名称。
  e. 单击“实例配置”。
  f. 在“参数类别”选择“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
  g. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
  h. 在“非默认”选项中选择“非默认”，界面上显示参数值为非默认值的参数。

11.5.3 管理服务操作

用户可以在MRS Manager：

- 启动操作状态为“停止”、“停止失败”或“启动失败”服务，以使用该服务。
- 停止不再使用或异常服务。
- 重启异常或配置过期的服务，以恢复或生效服务功能。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在指定服务所在行，单击“启动”、“停止”和“重启”执行启动、停止和重启操作。

服务之间存在依赖关系。对某服务执行启动、停止和重启操作时，与该服务存在依赖关系的服务将受到影响。

具体影响如下：

- 启动某服务，该服务依赖的下层服务需先启动，服务功能才可生效。
- 停止某服务，依赖该服务的上层服务将无法提供功能。
- 重启某服务，依赖该服务且启动的上层服务需重启后才可生效。

----结束

11.5.4 配置服务参数

用户可以根据实际业务场景，在MRS Manager中快速查看和修改服务默认的配置，及导出或导入配置。

对系统的影响

- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。
- 集群中只剩下一个DBService角色实例时，不支持修改DBService服务的参数。

操作步骤

- 修改服务参数。
  a. 单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“服务配置”。
  d. 在“参数类别”选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
  e. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

     修改某个参数的值后需要取消修改，可以单击 恢复。

     说明

     如果需要批量修改服务某个角色多个实例的配置，可以使用主机组实现实例参数的批量配置。在“角色”选择角色名称，然后在“主机”打开“<选择主机>”。“主机组名”填写一个名称，“主机”列表中勾选要修改的主机并加入“已选择的主机”，单击“确定”添加主机组。添加的主机组可以在“主机”中选择，且仅在当前页面有效，刷新页面后将无法保存。
  f. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

     界面提示“操作成功。”，单击“完成”，服务成功启动。

     说明

     更新YARN服务队列的配置且不重启服务时，选择“更多 > 刷新队列”更新队列使配置生效。
- 导出服务配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 单击“服务配置”。

d. 单击“导出服务配置”，选择一个位置保存配置文件。

- 导入服务配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 单击“服务配置”。
  d. 单击“导入服务配置”。
  e. 选择一个指定的配置文件。
  f. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。
     界面提示“操作成功。”，单击“完成”，服务成功启动。

11.5.5 配置服务自定义参数

MRS各个组件支持开源的所有参数，在MRS Manager支持修改部分关键使用场景的参数，且部分组件的客户端可能不包含开源特性的所有参数。如果需要修改其他Manager未直接支持的组件参数，用户可以在Manager通过自定义配置项功能为组件添加新参数。添加的新参数最终将保存在组件的配置文件中并在重启后生效。

对系统的影响

- 配置服务属性后，需要重启此服务，重启期间无法访问服务。
- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

前提条件

用户已充分了解需要新添加的参数意义、生效的配置文件以及对组件的影响。

操作步骤

步骤1 在MRS Manager界面，单击“服务管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“服务配置”。

步骤4 在“参数类别”选择“全部配置”。

步骤5 在左侧导航栏选择“自定义”，Manager将显示当前组件的自定义参数。

“参数文件”显示保存用户新添加的自定义参数的配置文件。每个配置文件中可能支持相同名称的开源参数，设置不同参数值后生效结果由组件加载配置文件的顺序决定。自定义参数支持服务级别与角色级别，请根据业务实际需要选择。不支持单个角色实例添加自定义参数。

步骤6 根据配置文件与参数作用，在对应参数项所在行“名称”列输入组件支持的参数名，在“值”列输入此参数的参数值。

- 支持单击  和  增加或删除一条自定义参数。第一次单击  添加自定义参数后才支持删除操作。
- 修改某个参数的值后需要取消修改，可以单击  恢复。

步骤7 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

任务示例

配置Hive自定义参数

Hive依赖于HDFS，默认情况下Hive访问HDFS时是HDFS的客户端，生效的配置参数统一由HDFS控制。例如HDFS参数“ipc.client.rpc.timeout”影响所有客户端连接HDFS服务端的RPC超时时间，如果用户需要单独修改Hive连接HDFS的超时时间，可以使用自定义配置项功能进行设置。在Hive的“core-site.xml”文件增加此参数可被Hive服务识别并代替HDFS的设置。

步骤1 在MRS Manager界面，选择 “服务管理 > Hive > 服务配置”。

步骤2 在“参数类别”选择“全部配置”。

步骤3 在左侧导航栏选择Hive服务级别“自定义”，Manager将显示Hive支持的服务级别自定义参数。

步骤4 在“core-site.xml”对应参数“core.site.customized.configs”的“名称：”输入“ipc.client.rpc.timeout”，“值：”输入新的参数值，例如“150000”。单位为毫秒。

步骤5 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“是”重启服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

11.5.6 同步服务配置

操作场景

当用户发现部分服务的“配置状态”为“过期”或“失败”时，您可以尝试使用同步配置功能，以恢复配置状态。或者集群中所有服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

对系统的影响

同步服务配置后，需要重启配置过期的服务。重启时对应的服务不可用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在服务列表中，单击指定服务名称。

步骤3 在服务状态及指标信息上方，选择“更多 > 同步配置”。

步骤4 在弹出窗口勾选“重启配置过期的服务或实例。”，并单击“确定”重启配置过期的服务。

界面提示“操作成功”，单击“完成”，服务成功启动。

----结束

11.5.7 管理角色实例操作

操作场景

用户可以在MRS Manager启动操作状态为“停止”、“停止失败”或“启动失败”角色实例，以使用该角色实例，也可以停止不再使用或异常的角色实例，或者重启异常的角色实例，以恢复角色实例功能。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“实例”页签。

步骤4 勾选待操作角色实例前的复选框。

步骤5 选择“更多 > 启动实例”、“停止实例”或“重启实例”，执行相应操作。

----结束

11.5.8 配置角色实例参数

操作场景

用户可以根据实际业务场景，在MRS Manager中快速查看及修改角色实例默认的配置。支持导出或导入配置。

对系统的影响

配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

操作步骤

- 修改角色实例参数。
  a. 单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“实例”页签。
  d. 单击角色实例列表中指定的角色实例名称。
  e. 单击“实例配置”页签。
  f. 在“参数类别”选择“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
  g. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

     修改某个参数的值后需要取消修改，可以单击 恢复。

h. 单击“保存配置”，勾选“重启角色实例”并单击“确定”，重启角色实例。

界面提示“操作成功。”，单击“完成”，角色实例成功启动。

- 导出角色实例配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 选中某角色或单击“实例”。
  d. 选择指定主机上某角色实例。
  e. 单击“实例配置”。
  f. 单击“导出实例配置”，导出指定角色实例配置数据并选择一个位置保存。
- 导入角色实例配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 选中某角色或单击“实例”。
  d. 选择指定主机上某角色实例。
  e. 单击“实例配置”。
  f. 单击“导入实例配置”，导入指定角色实例配置数据。
  g. 单击“保存配置”，勾选“重启角色实例。”并单击“确定”。
  界面提示“操作成功。”，单击“完成”，角色实例成功启动。

11.5.9 同步角色实例配置

操作场景

当用户发现角色实例的“配置状态”为“过期”或“失败”时，可以在MRS Manager尝试使用同步配置功能，同步角色实例的配置数据与后台配置数据，以恢复配置状态。

对系统的影响

同步配置角色实例后需要重启配置过期的角色实例。重启时对应的角色实例不可用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”，选择服务名称。

步骤2 单击“实例”页签。

步骤3 在角色实例列表中，单击指定角色实例名称。

步骤4 在角色实例状态及指标信息上方，选择“更多 > 同步配置”。

步骤5 在弹出窗口勾选“重启配置过期的服务或实例。”，并单击“确定”重启角色实例。

界面提示“操作成功。”，单击“完成”，角色实例成功启动。

----结束

11.5.10 退服和入服务角色实例

操作场景

某个Core或Task节点出现问题时，可能导致整个集群状态显示为“异常”。MRS集群支持将数据存储在不同Core节点，用户可以在MRS Manager指定角色实例退服，使退服的角色实例不再提供服务。在排除故障后，可以将已退服的角色实例入服。

支持退服、入服的角色实例包括：

- HDFS的DataNode角色实例
- Yarn的NodeManager角色实例
- HBase的RegionServer角色实例
- Kafka的Broker角色实例

限制：

- 当DataNode数量少于或等于HDFS的副本数时，不能执行退服操作。例如HDFS副本数为3时，则系统中少于4个DataNode，将无法执行退服，Manager在执行退服操作时会等待30分钟后报错并退出执行。
- Kafka Broker数量少于或等于副本数时，不能执行退服。例如Kafka副本数为2时，则系统中少于3个节点，将无法执行退服，Manager执行退服操作时会失败并退出执行。
- 已经退服的角色实例，必须执行入服操作启动该实例，才能重新使用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 单击服务列表中相应服务。

步骤3 单击“实例”页签。

步骤4 勾选指定角色实例名称前的复选框。

步骤5 选择“更多 > 退服”或“入服”执行相应的操作。

说明

实例退服操作未完成时在其他浏览器窗口重启集群中相应服务，可能导致MRS Manager提示停止退服，实例的“操作状态”显示为“已启动”。实际上后台已将该实例退服，请重新执行退服操作同步状态。

----结束

11.5.11 管理主机操作

操作场景

当主机故障异常时，用户可能需要在MRS Manager停止主机上的所有角色，对主机进行维护检查。故障清除后，启动主机上的所有角色恢复主机业务。

操作步骤

步骤1 单击“主机管理”。

步骤2 勾选待操作主机前的复选框。

步骤3 选择“更多 > 启动所有角色”或“停止所有角色”执行相应操作。

----结束

11.5.12 隔离主机

操作场景

用户发现某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。在为集群安装补丁的场景中，也支持排除指定节点不安装补丁。

该任务指导用户在MRS Manager上根据实际业务或运维规划手工将主机隔离。隔离主机仅支持隔离非管理节点。

对系统的影响

- 主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。
- 主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。

操作步骤

步骤1 在MRS Manager单击“主机管理”。

步骤2 勾选待隔离主机前的复选框。

步骤3 选择“更多 > 隔离主机”。

步骤4 在“隔离主机”，单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功隔离，“操作状态”显示为“已隔离”

说明

已隔离的主机，可以取消隔离重新加入集群，请参见**取消隔离主机**。

----结束

11.5.13 取消隔离主机

操作场景

用户已排除主机的异常或故障后，需要将主机隔离状态取消才能正常使用。

该任务指导用户在MRS Manager上取消隔离主机。

前提条件

- 主机状态为“已隔离”。
- 主机的异常或故障已确认修复。

操作步骤

步骤1 在MRS Manager单击“主机管理”。

步骤2 勾选待取消隔离主机前的复选框。

步骤3 选择“更多 > 取消隔离主机”。

步骤4 在“取消隔离主机”，单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功取消隔离，“操作状态”显示为“正常”。

步骤5 单击已取消隔离主机的名称，显示主机“状态”，单击“启动所有角色”。

----结束

11.5.14 启动及停止集群

操作场景

集群是包含着服务组件的集合。用户可以启动或者停止集群中所有服务。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在服务列表上方，选择“更多 > 启动集群”或“停止集群”执行相应的操作。

----结束

11.5.15 同步集群配置

操作场景

当MRS Manager显示全部服务或部分服务的“配置状态”为“过期”或“失败”时，用户可以尝试使用同步配置功能，以恢复配置状态。

- 若集群中所有服务的配置状态为“失败”时，同步集群的配置数据与后台配置数据。
- 若集群中某些服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

对系统的影响

同步集群配置后，需要重启配置过期的服务。重启时对应的服务不可用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在服务列表上方，选择“更多 > 同步配置”。

步骤3 在弹出窗口勾选“重启配置过期的服务或实例。”，并单击“确定”，重启配置过期的服务。

界面提示“操作成功”，单击“完成”，集群成功启动。

----结束

11.5.16 导出集群的配置数据

操作场景

为了满足实际业务的需求，用户可以在MRS Manager中将集群所有配置数据导出，导出文件用于快速更新服务配置。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 选择“更多 > 导出集群配置”。

导出文件用于更新服务配置，请参见**配置服务参数**中**导入服务配置参数**。

----结束

11.6 日志管理

11.6.1 关于日志

日志描述

MRS集群的日志保存路径为“/var/log/Bigdata”。日志分类见下表：

表 11-17 日志分类一览表

日志类型	日志描述
安装日志	安装日志记录了Manager、集群和服务安装的程序信息，可用于定位安装出错的问题。
运行日志	运行日志记录了集群各服务运行产生的运行轨迹信息及调试信息、状态变迁、未产生影响的潜在问题和直接的错误信息。
审计日志	审计日志中记录了用户活动信息和用户操作指令信息，可用于安全事件中定位问题原因及划分事故责任。

MRS日志目录清单见下表：

表 11-18 日志目录一览表

文件目录	日志内容
/var/log/Bigdata/audit	组件审计日志。
/var/log/Bigdata/controller	日志采集脚本日志。 controller进程日志。 controller监控日志。
/var/log/Bigdata/dbservice	DBService日志。
/var/log/Bigdata/flume	Flume日志。
/var/log/Bigdata/hbase	HBase日志。
/var/log/Bigdata/hdfs	HDFS日志。
/var/log/Bigdata/hive	Hive日志。
/var/log/Bigdata/httpd	httpd日志。
/var/log/Bigdata/hue	Hue日志。
/var/log/Bigdata/kerberos	Kerberos日志。
/var/log/Bigdata/ldapclient	LDAP客户端日志。
/var/log/Bigdata/ldapserver	LDAP服务端日志。
/var/log/Bigdata/loader	Loader日志。
/var/log/Bigdata/logman	logman脚本日志管理日志。
/var/log/Bigdata/mapreduce	MapReduce日志。
/var/log/Bigdata/nodeagent	NodeAgent日志。
/var/log/Bigdata/okerberos	OMS Kerberos日志。
/var/log/Bigdata/oldapserver	OMS LDAP日志。
/var/log/Bigdata/omm	oms：“omm”服务端的复杂事件处理日志、告警服务日志、HA日志、认证与授权管理日志和监控服务运行日志。 oma：“omm”代理端的安装运行日志。 core：“omm”代理端与“HA”进程失去响应的dump日志。
/var/log/Bigdata/spark	Spark日志。
/var/log/Bigdata/sudo	omm执行sudo命令产生的日志。
/var/log/Bigdata/timestamp	时间同步管理日志。
/var/log/Bigdata/tomcat	Tomcat日志。
/var/log/Bigdata/yarn	Yarn日志。

文件目录	日志内容
/var/log/Bigdata/zookeeper	ZooKeeper日志。
/var/log/Bigdata/kafka	Kafka日志。
/var/log/Bigdata/storm	Storm日志。
/var/log/Bigdata/patch	补丁日志。

运行日志

运行日志记录的运行信息描述如表11-19所示。

表 11-19 运行信息一览表

运行日志	日志描述
服务安装前的准备日志	记录服务安装前的准备工作，如检测、配置和反馈操作的信息。
进程启动日志	记录进程启动过程中执行的命令信息。
进程启动异常日志	记录进程启动失败时产生异常的信息，如依赖服务错误、资源不足等
进程运行日志	记录进程运行轨迹信息及调试信息，如函数入口和出口打印、模块间接口消息等。
进程运行异常日志	记录导致进程运行时错误的错误信息，如输入对象为空、编解码失败等错误。
进程运行环境信息日志	记录进程运行环境的信息，如资源状态、环境变量等。
脚本日志	记录脚本执行的过程信息。
资源回收日志	记录资源回收的过程信息。
服务卸载时的清理日志	记录卸载服务时执行的步骤操作信息，如清除目录数据、执行时间等

审计日志

审计日志记录的审计信息包含Manager审计信息和组件审计信息。

表 11-20 Manager 审计信息一览表

审计日志	操作类型	操作
Manager审计日志	用户管理	创建用户 修改用户 删除用户 创建组 修改组 删除组 添加角色 修改角色 删除角色 密码策略修改 修改密码 密码重置 用户登录 用户注销 屏幕解锁 下载认证凭据 用户越权操作 用户帐号解锁 用户帐号锁定 屏幕锁定 导出用户 导出用户组 导出角色
	租户管理	保存静态配置 增加租户 删除租户 关联租户服务 删除租户服务 配置资源 创建资源 删除资源 增加资源池 修改资源池 删除资源池 恢复租户数据

审计日志	操作类型	操作
	集群管理	启动集群 停止集群 保存配置 同步集群配置 定制集群监控指标 保存监控阈值 下载客户端配置 北向接口配置 北向SNMP接口配置 创建阈值模板 删除阈值模板 应用阈值模板 保存集群监控配置数据 导出配置数据 导入集群配置数据 导出安装模板 修改阈值模板 取消阈值模板应用 屏蔽告警 发送告警 修改OMS数据库密码 修改组件数据库密码 启动集群的健康检查 更新健康检查的配置 导出集群健康检查的结果 导入证书文件 删除健康检查历史报告 导出健康检查历史报告 定制报表监控指标 导出报表监控数据 定制静态资源池监控指标 导出静态资源池监控数据

审计日志	操作类型	操作
	服务管理	启动服务 停止服务 同步服务配置 刷新服务队列 定制服务监控指标 重启服务 导出服务监控数据 导入服务配置数据 启动服务的健康检查 导出服务健康检查的结果 服务配置 上传配置文件 下载配置文件
	实例管理	同步实例配置 实例入服 实例退服 启动实例 停止实例 定制实例监控指标 重启实例 导出实例监控数据 导入实例配置数据
	主机管理	设置节点机架 启动所有角色 停止所有角色 隔离主机 取消隔离主机 定制主机监控指标 导出主机监控数据 启动主机的健康检查 导出主机健康检查的结果

审计日志	操作类型	操作
	维护管理	导出告警 清除告警 导出事件 批量清除告警 SNMP清除告警 SNMP添加trap目标 SNMP删除trap目标 SNMP检查告警 SNMP同步告警 修改审计转储配置 导出审计日志 采集日志文件 下载日志文件 上传文件 删除已上传的文件 创建备份任务 执行备份任务 停止备份任务 删除备份任务 修改备份任务 锁定备份任务 解锁备份任务 创建恢复任务 执行恢复任务 停止恢复任务 重试恢复任务 删除恢复任务

表 11-21 组件审计信息一览表

审计日志	操作类型	操作
DBService审计日志	维护管理	备份恢复操作

审计日志	操作类型	操作
HBase审计日志	DDL（数据定义）语句	创建表 删除表 修改表 增加列族 修改列族 删除列族 启用表 禁用表 用户信息修改 修改密码 用户登录
	DML（数据操作）语句	put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表）
	权限控制	给用户授权 取消用户授权
Hive审计日志	元数据操作	元数据定义，如创建数据库、表等 元数据删除，如删除数据库、表等 元数据修改，如增加列、重命名表等 元数据导入/导出
	数据维护	向表中加载数据 向表中插入数据
	权限管理	创建/删除角色 授予/回收角色 授予/回收权限
HDFS审计日志	权限管理	文件/文件夹访问权限 文件/文件夹owner信息

审计日志	操作类型	操作
	文件操作	创建文件夹 创建文件 打开文件 追加文件内容 修改文件名称 删除文件/文件夹 设置文件时间属性 设置文件副本个数 多文件合并 文件系统检查 文件链接
Mapreduce审计日志	程序运行	启动Container请求 停止Container请求 Container结束，状态为成功 Container结束，状态为失败 Container结束，状态为中止 提交任务 结束任务
LdapServer审计日志	维护管理	添加操作系统用户 添加组 添加用户到组 删除用户 删除组
KrbServer审计日志	维护管理	修改kerberos帐号密码 添加kerberos帐号 删除kerberos帐号 用户认证
Loader审计日志	安全管理	用户登录
	元数据管理	查询connector 查询framework 查询step

审计日志	操作类型	操作
	数据源连接管理	查询数据源连接 增加数据源连接 更新数据源连接 删除数据源连接 激活数据源连接 禁用数据源连接
	作业管理	查询作业 创建作业 更新作业 删除作业 激活作业 禁用作业 查询作业所有执行记录 查询作业最近执行记录 提交作业 停止作业
Hue审计日志	服务启动	启动Hue
	用户操作	用户登录 用户退出
	任务操作	创建任务 修改任务 删除任务 提交任务 保存任务 任务状态更新
Zookeeper审计日志	权限管理	设置ZNODE访问权限
	ZNODE操作	创建ZNODE 删除ZNODE 设置ZNODE数据
Storm审计日志	Nimbus	提交拓扑 中止拓扑 重分配拓扑 去激活拓扑 激活拓扑

审计日志	操作类型	操作
	UI	中止拓扑 重分配拓扑 去激活拓扑 激活拓扑

MRS的审计日志保存在数据库中，可通过“审计管理”页面查看及导出审计日志。

组件审计日志的文件信息见下表。部分组件审计日志文件保存在“/var/log/Bigdata/audit”，例如HDFS、HBase、Mapreduce、Hive、Hue、Yarn、Storm和ZooKeeper。每天凌晨3点自动将组件审计日志压缩备份到“/var/log/Bigdata/audit/bk”，最多保留最近的90个压缩备份文件，不支持修改备份时间。

其他组件审计日志文件保存在组件日志目录中。

表 11-22 组件审计日志目录

组件名称	审计日志目录
DBService	/var/log/Bigdata/audit/dbservice/dbservice_audit.log
HDFS	/var/log/Bigdata/audit/hdfs/nn/hdfs-audit-namenode.log /var/log/Bigdata/audit/hdfs/dn/hdfs-audit-datanode.log /var/log/Bigdata/audit/hdfs/jn/hdfs-audit-journalnode.log /var/log/Bigdata/audit/hdfs/zkfc/hdfs-audit-zkfc.log /var/log/Bigdata/audit/hdfs/httpfs/hdfs-audit-httpfs.log /var/log/Bigdata/audit/hdfs/router/hdfs-audit-router.log
Mapreduce	/var/log/Bigdata/audit/mapreduce/jobhistory/mapred-audit-jobhistory.log
Hive	/var/log/Bigdata/audit/hive/hiveserver/hive-audit.log /var/log/Bigdata/audit/hive/metastore/metastore-audit.log /var/log/Bigdata/audit/hive/webhcat/webhcat-audit.log
Loader	/var/log/Bigdata/loader/audit/default.audit
Hue	/var/log/Bigdata/audit/hue/hue-audits.log
ZooKeeper	/var/log/Bigdata/audit/zookeeper/quorumpeer/zk-audit-quorumpeer.log

组件名称	审计日志目录
Spark	/var/log/Bigdata/audit/spark/jdbcserver/jdbcserver-audit.log /var/log/Bigdata/audit/spark/jobhistory/jobhistory-audit.log
Yarn	/var/log/Bigdata/audit/yarn/rm/yarn-audit-resourcemanager.log /var/log/Bigdata/audit/yarn/nm/yarn-audit-nodemanager.log
Storm	/var/log/Bigdata/audit/storm/nimbus/audit.log /var/log/Bigdata/audit/storm/ui/audit.log

11.6.2 Manager 日志清单

日志描述

日志存储路径：Manager相关日志的默认存储路径为“/var/log/Bigdata/*Manager组件*”。

- ControllerService：/var/log/Bigdata/controller/（OMS安装、运行日志）
- Httpd：/var/log/Bigdata/httpd（httpd安装、运行日志）
- logman：/var/log/Bigdata/logman（日志打包工具日志）
- NodeAgent：/var/log/Bigdata/nodeagent（NodeAgent安装、运行日志）
- okerberos：/var/log/Bigdata/okerberos（okerberos安装、运行日志）
- oldapserver：/var/log/Bigdata/oldapserver（oldapserver安装、运行日志）
- MetricAgent：/var/log/Bigdata/metric_agent（MetricAgent运行日志）
- omm：/var/log/Bigdata/omm（omm安装、运行日志）
- timestamp：/var/log/Bigdata/timestamp（NodeAgent启动时间日志）
- tomcat：/var/log/Bigdata/tomcat（Web进程日志）
- patch：/var/log/Bigdata/patch（补丁安装日志）
- Sudo：/var/log/Bigdata/sudo（sudo脚本执行日志）
- OS：/var/log/*message文件*（OS系统日志）
- OS Performance：/var/log/osperf（OS性能统计日志）
- OS Statistics：/var/log/osinfo/statistics（OS参数配置信息日志）

日志归档规则：

Manager的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过10MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件。

表 11-23 Manager 日志列表

日志类型	日志文件名	描述
Controller运行日志	controller.log	记录组件安装、升级、补丁、配置、监控、告警和日常运维操作日志。
	controller_client.log	Rest接口运行日志。
	acs.log	Acs运行日志。
	acs_spnego.log	acs中spnego用户日志
	aos.log	Aos运行日志。
	plugin.log	Aos插件日志
	backupplugin.log	备份恢复进程运行日志
	controller_config.log	配置运行日志
	controller_nodesetup.log	Controller加载任务日志
	controller_root.log	Controller进程系统日志
	controller_trace.log	Controller与NodeAgent之间RPC通信日志
	controller_monitor.log	监控日志
	controller_fsm.log	状态机日志
	controller_alarm.log	Controller发送告警日志
	controller_backup.log	Controller备份恢复日志
	install.log，distributeAdapterFiles.log，install_os_optimization.log	oms安装日志
	oms_ctl.log	oms启停日志
	installntp.log	ntp安装日志
	modify_manager_param.log	修改Manager参数日志
	backup.log	OMS备份脚本运行日志
	supressionAlarm.log	告警脚本运行日志
	om.log	生成om证书日志
	backupplugin_ctl.log	备份恢复插件进程启动日志
	getLogs.log	采集日志脚本运行日志

日志类型	日志文件名	描述
	backupAuditLogs.log	审计日志备份脚本运行日志
	certStatus.log	证书定期检查日志
	distribute.log	证书分发日志
	ficertgenetrate.log	证书替换日志，包括生成二级证书、cas证书、httpd证书的日志。
	genPwFile.log	生成证书密码文件日志
	modifyproxyconf.log	修改HTTPD代理配置的日志
	importTar.log	证书导入信任库日志
Httpd	install.log	Httpd安装日志
	access_log, error_log	Httpd运行日志
logman	logman.log	日志打包工具日志。
NodeAgent	install.log，install_os_optimization.log	NodeAgent安装日志
	installntp.log	ntp安装日志
	start_ntp.log	ntp启动日志
	ntpChecker.log	ntp检查日志
	ntpMonitor.log	ntp监控日志
	heartbeat_trace.log	NodeAgent与Controller心跳日志
	alarm.log	告警日志
	monitor.log	监控日志
	nodeagent_ctl.log，start-agent.log	NodeAgent启动日志
	agent.log	NodeAgent运行日志
	cert.log	证书日志
	agentplugin.log	监控agent侧插件运行日志
	omaplugin.log	OMA插件运行日志
	diskhealth.log	磁盘健康检查日志
	supressionAlarm.log	告警脚本运行日志

日志类型	日志文件名	描述
	updateHostFile.log	更新主机列表日志
	collectLog.log	节点日志采集脚本运行日志
	host_metric_collect.log	主机指标采集运行日志
	checkfileconfig.log	文件权限配置检查运行日志
	entropycheck.log	熵值检查运行日志
	timer.log	节点定时调度日志
	pluginmonitor.log	组件监控插件日志
	agent_alarm_py.log	NodeAgent检查文件权限发送告警日志
okerberos	addRealm.log，modifyKerberosRealm.log	切域日志
	checkservice_detail.log	Okerberos健康检查日志
	genKeytab.log	生成keytab日志
	KerberosAdmin_genConfigDetail.log	启动kadmin进程时，生成kadmin.conf的运行日志
	KerberosServer_genConfigDetail.log	启动krb5kdc进程时，生成krb5kdc.conf的运行日志
	oms-kadmind.log	kadmin进程的运行日志
	oms_kerberos_install.log，postinstall_detail.log	okerberos安装日志
	oms-krb5kdc.log	krbkdc运行日志
	start_detail.log	okerberos启动日志
	realmDataConfigProcess.log	切域失败，回滚日志
	stop_detail.log	okerberos停止日志
oldapserver	ldapserver_backup.log	Oldapserver备份日志
	ldapserver_chk_service.log	Oldapserver健康检查日志
	ldapserver_install.log	Oldapserver安装日志
	ldapserver_start.log	Oldapserver启动日志

日志类型	日志文件名	描述
	ldapserver_status.log	Oldapserver进程状态检查日志。
	ldapserver_stop.log	Oldapserver停止日志
	ldapserver_wrap.log	Oldapserver服务管理日志。
	ldapserver_uninstall.log	Oldapserver卸载日志
	restart_service.log	Oldapserver重启日志
	ldapserver_unlockUser.log	记录解锁Ldap用户和管理帐户的日志
omm	omsconfig.log	OMS配置日志
	check_oms_heartbeat.log	OMS心跳运行日志
	monitor.log	OMS监控日志
	ha_monitor.log	HA_Monitor操作日志
	ha.log	HA操作日志
	fms.log	告警日志
	fms_ha.log	告警的HA监控日志
	fms_script.log	告警控制日志
	config.log	告警配置日志
	iam.log	IAM日志
	iam_script.log	IAM控制日志
	iam_ha.log	IAM的HA监控日志
	config.log	IAM配置日志
	operatelog.log	IAM操作日志
	heartbeatcheck_ha.log	OMS心跳的HA监控日志
	install_oms.log	OMS安装日志
	pms_ha.log	监控的HA监控日志
	pms_script.log	监控控制日志
	config.log	监控配置日志
	plugin.log	监控插件运行日志
	pms.log	监控日志
	ha.log	HA运行日志

日志类型	日志文件名	描述
	cep_ha.log	CEP的HA监控日志
	cep_script.log	CEP控制日志
	cep.log	CEP日志
	config.log	CEP配置日志
	omm_gaussdba.log	gaussdb的HA监控日志
	gaussdb-<SERIAL>.log	gaussdb运行日志
	gs_ctl-<DATE>.log	gaussdb控制日志的归档日志
	gs_ctl-current.log	gaussdb控制日志
	gs_guc-current.log	gaussdb操作日志
	encrypt.log	omm加密日志
	omm_agent_ctl.log	OMA控制日志
	oma_monitor.log	OMA监控日志
	install_oma.log	OMA安装日志
	config_oma.log	OMA配置日志
	omm_agent.log	OMA运行日志
	acs.log	acs资源日志。
	aos.log	aos资源日志
	controller.log	controller资源日志
	feed_watchdog.log	feed_watchdog资源日志
	floatip.log	floatip资源日志
	ha_ntp.log	ntp资源日志
	httpd.log	httpd资源日志
	okerberos.log	okerberos资源日志
	oldap.log	oldap资源日志
	tomcat.log	tomcat资源日志
	send_alarm.log	管理节点HA告警发送脚本运行日志
timestamp	restart_stamp	NodeAgent启动时间
tomcat	cas.log，localhost_access_cas_log.log	cas运行日志

日志类型	日志文件名	描述
	catalina.log，catalina.out，host-manager.log，localhost.log，manager.log	tomcat运行日志
	localhost_access_web_log.log	记录访问FusionInsight Manager系统REST接口的日志
	web.log	web进程运行日志
	northbound_ftp_sftp.log，snmp.log	北向日志
watchdog	watchdog.log，feed_watchdog.log	watchdog.log运行日志
patch	oms_installPatch.log	OMS补丁安装日志
	agent_installPatch.log	Agent补丁安装日志
	agent_uninstallPatch.log	agent补丁卸载日志
	NODE_AGENT_restoreFile.log	agent补丁恢复文件日志
	NODE_AGENT_updateFile.log	agent补丁更新文件日志
	OMA_restoreFile.log	OMA补丁恢复文件日志
	OMA_updateFile.log	OMA补丁更新文件日志
	CONTROLLER_restoreFile.log	CONTROLLER补丁恢复文件日志
	CONTROLLER_updateFile.log	CONTROLLER补丁更新文件日志
	OMS_restoreFile.log	OMS补丁恢复文件日志
	oms_uninstallPatch.log	OMS补丁卸载日志
	OMS_updateFile.log	OMS补丁更新文件日志
	createStackConf.log，decompress.log，decompress_OMS.log ，distrExtractPatchOnOMS.log，slimReduction.log，switch_adapter.log	补丁安装日志
sudo	sudo.log	sudo脚本执行日志

日志级别

Manager中提供了如表11-24所示的日志级别。日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 11-24 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO表示记录系统及各事件正常运行状态信息
DEBUG	DEBUG表示记录系统及系统的调试信息。

日志格式

Manager的日志格式如下所示：

表 11-25 日志格式

日志类型	组件	格式	示例
Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2015-06-30 00:37:09,067 INFO [pool-1-thread-1] Completed Discovering Node. com.xxx.hadoop.om.controller.tasks.nodesetup.DiscoverNodeTask.execute(DiscoverNodeTask.java:299)

11.6.3 查看及导出审计日志

操作场景

该任务指导用户在MRS Manager查看、导出审计日志工作，用于安全事件中事后追溯、定位问题原因及划分事故责任。

系统记录的日志信息包含：

- 用户活动信息，如用户登录与注销，系统用户信息变更，系统用户组信息变更等。
- 用户操作指令信息，如集群的启动、停止，软件升级等。

操作步骤

- 查看审计日志
  a. 在MRS Manager，单击“审计管理”，可直接查看默认的审计日志。
     若审计日志的审计内容长度大于256字符，请单击审计日志展开按钮展开审计详情，单击“日志文件”，下载完整文件查看信息。
     - 默认以“产生时间”列按降序排列，单击**操作类型**、**安全级别**、**产生时间**、**用户**、**主机**、**服务**、**实例或操作结果**可修改排列方式。
     - 支持在“安全级别”筛选相同级别的全部告警。结果包含已清除和未清除的告警。

     导出的审计日志文件，包含以下信息列：
     - “编号”：表示MRS Manager已生成的审计日志数量，每增加一条审计日志则编号自动加1。
     - “操作类型”：表示用户操作的操作类型，分为“告警”、“审计日志”、“备份恢复”、“集群”、“采集日志”、“主机”、“服务”、“多租户”和“用户管理”九种场景，其中“用户管理”仅在启用了Kerberos认证的集群中支持。每个场景中包含不同操作类型，例如“告警”中包含“导出告警”，“集群”中包含“启动集群”，“多租户”包含“增加租户”等。
     - “安全级别”：表示每条审计日志的安全级别，包含“高危”、“危险”、“一般”和“提示”四种。
     - “开始时间”：表示用户操作开始的时间，且时间为UTC/GMT+08:00时间。
     - “结束时间”：表示用户操作结束的时间，且时间为UTC/GMT+08:00时间。
     - “用户IP”：表示用户操作时所使用的IP地址。
     - “用户”：表示执行操作的用户名。
     - “主机”：表示用户操作发生在集群的哪个节点。如果操作不涉及节点则不保存信息。
     - “服务”：表示用户操作发生在集群的哪个服务。如果操作不涉及服务则不保存信息。
     - “实例”：表示用户操作发生在集群的哪个角色实例。如果操作不涉及角色实例则不保存信息。
     - “操作结果”：表示用户操作的结果，包含“成功”、“失败”和“未知”。
     - “内容”：表示用户操作的具体执行信息。

b. 单击“高级搜索”，在审计日志搜索区域中，设置查询条件，单击“搜索”，查看指定类型的审计日志。单击“重置”清除输入的搜索条件。

说明

“开始时间”和“结束时间”表示时间范围的开始时间和结束时间，可以搜索此时间段内产生的告警。

- 导出审计日志
  a. 在审计日志列表中，单击“导出全部”，导出所有的日志。
  b. 在审计日志列表中，勾选日志信息前的复选框，单击“导出”，导出指定日志。

11.6.4 导出服务日志

操作场景

该任务指导用户从MRS Manager导出各个服务角色生成的日志。

前提条件

- 用户已经获取帐号对应的Access Key ID（AK）和Secret Access Key（SK）。
- 用户已经在帐号的对象存储服务（OBS）中创建了并行文件系统。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 单击“维护”下方的“日志导出”。

步骤3 “服务”选择服务，“主机”填写服务所部署主机的IP，“开始时间”与“结束时间”选择对应的开始与结束时间。

步骤4 在“导出类型”选择一个日志保存的位置。只有启用了Kerberos认证的集群支持选择。

- “下载到本地”：表示将日志保存到用户当前的本地环境。然后执行**步骤8**。
- “上传到OBS”：表示将日志保存到OBS中。默认值。然后执行**步骤5**。

步骤5 在“OBS路径”填写服务日志在OBS保存的路径。

需要填写完整路径且不能以“/”开头，路径可以不存在，系统将自动创建。OBS的完整路径最大支持900个字节。

步骤6 在“桶名”输入已创建的OBS文件系统名称。

步骤7 在“AK”和“SK”输入用户的Access Key ID和Secret Access Key。

步骤8 单击“确定”完成日志下载。

----结束

11.6.5 配置审计日志导出参数

操作场景

MRS的审计日志长期保留在系统中，可能引起数据目录的磁盘空间不足问题，故通过设置导出参数及时将审计日志自动导出到OBS服务器的指定目录下，便于管理审计日志信息。

说明

审计日志导出到OBS服务器的内容包含两部分，服务审计日志和管理审计日志。

- 服务审计日志每天凌晨3点自动压缩存储到主管理节点“/var/log/Bigdata/audit/bk/”，保存的文件名格式为<yyyy-MM-dd_HH-mm-ss>.tar.gz。默认情况下，保存的文件个数为7份（即7天的日志），超过7份文件时会自动删除7天前的文件。
- 管理审计日志每次导出到OBS的数据范围是从最近一次成功导出到OBS的日期至本次执行任务的日期。管理审计日志每达到10万条时，系统自动将前9万条审计日志转储保存到本地文件中，数据库中保留1万条。转储的日志文件保存路径为主管理节点“${BIGDATA_DATA_HOME}/dbdata_om/dumpData/iam/operatelog”，保存的文件名格式为“OperateLog_store_YY_MM_DD_HH_MM_SS.csv”，保存的审计日志历史文件数最大为50个。

前提条件

- 用户已经获取帐号对应的Access Key ID（AK）和Secret Access Key（SK）。
- 用户已经在帐号的对象存储服务（OBS）中创建了并行文件系统。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“维护”下单击“审计日志导出”。

表 11-26 审计日志导出参数

参数名	参数值	参数解释
开始时间	7/24/2017 09:00:00（举例）	必选参数，指定审计日志导出的开始时间。
周期	1天（举例）	必选参数，指定审计日志转导出的时间间隔，间隔周期范围为（1~5天）。
桶名	mrs-bucket（举例）	必选参数，指定审计日志导出到OBS的文件系统名。
OBS路径	opt/omm/oms/auditLog（举例）	必选参数，指定审计日志导出到OBS的路径。
AK	XXX（举例）	必选参数，用户的Access Key ID。
SK	XXX（举例）	必选参数，用户的Secret Access Key。

说明

审计日志在OBS的存储路径细分为service_auditlog和manager_auditlog，分别用于存储服务审计日志和管理审计日志。

----结束

11.7 健康检查管理

11.7.1 执行健康检查

操作场景

该任务指导用户在日常运维中完成集群进行健康检查的工作，以保证集群各项参数、配置以及监控没有异常、能够长时间稳定运行。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

操作步骤

- 手动执行所有服务的健康检查
  a. 单击“服务管理”并选择对应服务。
  b. 选择“更多 > 启动服务健康检查”，启动服务健康检查。

  说明

  - 集群健康检查包含了Manager、服务与主机状态的检查。
  - 在MRS Manager界面，选择“系统设置 > 健康检查 > 集群健康检查”，也可以执行集群健康检查。
  - 手动执行健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。
- 手动执行单个服务的健康检查
  a. 选择“服务管理”，在服务列表中单击服务指定名称。
  b. 选择“更多 > 启动服务健康检查”启动指定服务健康检查。
- 手动执行主机健康检查
  a. 单击“主机管理”。
  b. 勾选待检查主机前的复选框。
  c. 选择“更多 > 启动主机健康检查”启动指定主机健康检查。
- 自动执行健康检查
  a. 单击“系统设置”。
  b. 单击“维护”下方的“健康检查”。

c. 单击“健康检查配置”，配置自动执行健康检查。

“定期健康检查”配置是否启用自动执行健康检查。“定期健康检查”的开关默认为关闭，单击可启用该功能，根据管理需要选择“每天”、“每周”或“每月”。

d. 单击“确定”保存配置。系统右上角弹出提示“健康检查配置保存成功。”。

11.7.2 查看并导出检查报告

操作场景

为了满足对健康检查结果的进一步具体分析，您可以在MRS Manager中查看以及导出健康检查的结果。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

前提条件

已执行健康检查。

操作步骤

步骤1 单击“服务管理”。

步骤2 选择“更多 > 查看集群健康检查报告”，查看集群健康检查的报告。

步骤3 在健康检查的报告面板上单击“导出报告”导出健康检查报告，可查看检查项的完整信息。

说明

对于存在问题的检查项，请参见**DBService健康检查指标项说明**~**ZooKeeper健康检查指标项说明**进行修复。

----结束

11.7.3 配置健康检查报告保存数

操作场景

在不同时间、不同使用场景下，MRS集群、服务和主机产生的健康检查报告结果不完全相同。如果需要保存更多的报告用于比较时，可以在MRS Manager修改健康检查报告保存的文件数。

健康检查报告保存的文件数不区分集群、服务或主机类型的健康检查报告。健康检查完成后，报告文件默认保存在主管理节点的“$BIGDATA_DATA_HOME/Manager/healthcheck”，备管理节点将自动同步。

前提条件

用户已明确业务需求，并规划好保存的时间跨度与健康检查频率，检查主备管理节点磁盘空间使用率。

操作步骤

步骤1 选择“系统设置 > 健康检查 > 健康检查配置”。

步骤2 “健康检查报告文件最大份数”参数填写健康检查报告的保存个数。默认值为“50”，取值范围为1~100。

步骤3 单击“确定”保存配置。系统右上角弹出提示“健康检查配置保存成功。”。

----结束

11.7.4 管理健康检查报告

操作场景

用户可以在MRS Manager对已保存的历史健康检查报告进行管理，即查看、下载和删除历史健康检查报告。

操作步骤

- 下载指定的健康检查报告
  a. 选择“系统设置 > 健康检查”。
  b. 在目标健康检查报告所在行，单击“下载”，下载报告文件。
- 批量下载指定的健康检查报告
  a. 选择“系统设置 > 健康检查”。
  b. 勾选多个目标健康检查报告，单击“下载文件”，下载多个报告文件。
- 删除指定的健康检查报告
  a. 选择“系统设置 > 健康检查”。
  b. 在目标健康检查报告所在行，单击“删除”，删除报告文件。
- 批量删除指定的健康检查报告
  a. 选择“系统设置 > 健康检查”。
  b. 勾选多个目标健康检查报告，单击“删除文件”，删除多个报告文件。

11.7.5 DBService 健康检查指标项说明

服务健康检查

指标项名称：服务状态

指标项含义：检查DBService服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-27001进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.6 Flume 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查Flume服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-24000进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.7 HBase 健康检查指标项说明

运行良好的 RegionServer 数

指标项名称：运行良好的RegionServer数

指标项含义：检查HBase集群中运行良好的RegionServer数。

恢复指导：如果该指标项异常，请检查RegionServer的状态是否正常并处理，然后建议检查网络是否正常。

服务健康状态

指标项名称：服务状态

指标项含义：检查HBase服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，请检查HMaster和RegionServer的状态是否正常并先处理，然后检查ZooKeeper服务的状态是否为故障并处理。使用客户端，确认是否可以正确读取HBase表中的数据，排查读数据失败的原因。最后参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.8 Host 健康检查指标项说明

Swap 使用率

指标项名称：Swap使用率

指标项含义：系统Swap使用率，计算方法：已用Swap大小/总共Swap大小。当前阈值设置为75.0%，如果使用率超过阈值，则认为不健康。

恢复指导：

1. 确认节点Swap使用率。

   登录检查结果不健康的节点，执行**free -m**查看swap总量和已使用量，如果swap使用率已超过阈值，则执行**2**。
2. 如果Swap使用率超过阈值，建议对系统进行扩容，如：增加节点。

主机文件句柄使用率

指标项名称：主机文件句柄使用率

指标项含义：系统中文件句柄的使用率，主机文件句柄使用率=已用句柄数/总共句柄数。如果使用率超过阈值，则认为不健康。

恢复指导：

1. 确认主机文件句柄使用率。

   登录检查结果不健康的节点，执行**cat /proc/sys/fs/file-nr**，输出结果的第一列和第三列分别表示系统已使用的句柄数和总句柄数，如果使用率超过阈值，则执行**2**。
2. 如果主机文件句柄使用率超过阈值，建议对系统进行排查，具体分析文件句柄的使用情况。

NTP 偏移量

指标项名称：NTP偏移量

指标项含义：NTP时间偏差。如果时间偏差超过阈值，则认为不健康。

恢复指导：

1. 确认NTP时间偏差。

   登录检查结果不健康的节点，执行**/usr/sbin/ntpq -np**查看信息，其中offset列表示时间偏差。如果时间偏差大于阈值，则执行**2**。
2. 如果该指标项异常，则需要检查时钟源配置是否正确，请联系运维人员处理。

平均负载

指标项名称：平均负载

指标项含义：系统平均负载，表示特定时间段内运行队列中的平均进程数。这里系统平均负载是通过uptime命令中得到的负载值计算得到。计算方法：（1分钟负载 + 5分钟负载 + 15分钟负载）/(3*CPU个数)。当前阈值设置为2，如果超过阈值，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点，执行**uptime**命令，命令输出的最后三列分别表示1分钟负载、5分钟负载和15分钟负载。根据系统平均负载的计算方法，如果负载超过阈值，则执行**2**。
2. 如果系统平均负载超过阈值，建议对系统进行扩容，如增加节点等。

D 状态进程

指标项名称：D状态进程

指标项含义：不可中断的睡眠进程，即D状态进程。D状态通常是进程在等待IO，比如磁盘IO，网络IO等，但是此时IO出现异常。如果系统中出现D状态进程，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12028进行处理。

硬件状态

指标项名称：硬件状态

指标项含义：检查系统硬件状态，包括CPU、内存、磁盘、电源、风扇等。该检查项通过ipmitool sdr elist获取相关硬件信息。如果相关硬件状态异常，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点。执行**ipmitool sdr elist**查看系统硬件状态，命令输出的最后一列表示对应的硬件状态。如果提示的状态在下面的故障描述表中，则任务不健康。

模块	故障描述
Processor	IERR Thermal Trip FRB1/BIST failure FRB2/Hang in POST failure FRB3/Processor startup/init failure Configuration Error SM BIOS Uncorrectable CPU-complex Error Disabled Throttled Uncorrectable machine check exception

模块	故障描述
Power Supply	Failure detected Predictive failure Power Supply AC lost AC lost or out-of-range AC out-of-range, but present Config Error: Vendor Mismatch Config Error: Revision Mismatch Config Error: Processor Missing Config Error: Power Supply Rating Mismatch Config Error: Voltage Rating Mismatch Config Error
Power Unit	240VA power down Interlock power down AC lost Soft-power control failure Failure detected Predictive failure
Memory	Uncorrectable ECC Parity Memory Scrub Failed Memory Device Disabled Correctable ECC logging limit reached Configuration Error Throttled Critical Overtemperature
Drive Slot	Drive Fault Predictive Failure Parity Check In Progress In Critical Array In Failed Array Rebuild In Progress Rebuild Aborted
Battery	Low Failed

2. 如果该指标项异常，建议联系运维人员解决处理。

主机名

指标项名称：主机名

指标项含义：检查是否设置了主机名。如果没有设置主机名，则认为不健康。如果该指标项异常，建议正确设置hostname。

恢复指导：

1. 登录检查结果不健康的节点。
2. 执行以下命令修改主机名，使节点主机名与规划的主机名保持一致：

   hostname *主机名*。例如，将主机名改为“Bigdata-OM-01”，请执行命令hostname Bigdata-OM-01。
3. 修改主机名配置文件。

   执行**vi /etc/HOSTNAME**命令编辑文件，修改文件内容为“Bigdata-OM-01”，并保存退出。

Umask

指标项名称：Umask

指标项含义：检查omm用户的umask设置是否正确。如果umask设置不等于0077，则认为不健康。

恢复指导：

1. 如果该指标异常，建议将omm用户的umask设置为0077。登录检查结果不健康的节点，执行**su - omm**切换到omm用户。
2. 执行**vi ${BIGDATA_HOME}/.om_profile**，修改**umask=0077**，保存并退出。

OMS 的 HA 状态

指标项名称：OMS的HA状态

指标项含义：检查OMS的双机资源是否正常。OMS双机资源状态的详细信息可使用**${CONTROLLER_HOME}/sbin/status-oms.sh**查看。如果有模块状态异常，认为不健康。

恢复指导：

1. 登录主管理节点，执行**su - omm**切换到**omm**用户，然后执行**${CONTROLLER_HOME}/sbin/status-oms.sh**查看OMS状态。
2. 如果floatip、okerberos、oldap等异常，可参见告警ALM-12002、ALM-12004、ALM-12005分别进行处理。
3. 如果是其他资源异常，建议查看相关异常模块的日志。

   controller资源异常：查看异常节点的/var/log/Bigdata/controller/controller.log。

   cep资源异常：查看异常节点的/var/log/Bigdata/omm/oms/cep/cep.log。

   aos资源异常：查看异常节点的/var/log/Bigdata/controller/aos/aos.log。

   feed_watchdog资源异常：查看异常节点的/var/log/Bigdata/watchdog/watchdog.log。

   httpd资源异常：查看异常节点的/var/log/Bigdata/httpd/error_log。

   fms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/fms/fms.log。

pms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/pms/pms.log。

iam资源异常：查看异常节点的/var/log/Bigdata/omm/oms/iam/iam.log。

gaussDB资源异常：查看异常节点的/var/log/Bigdata/omm/oms/db/omm_gaussdba.log。

ntp资源异常：查看异常节点的/var/log/Bigdata/omm/oms/ha/scriptlog/ha_ntp.log。

tomcat资源异常：查看异常节点的/var/log/Bigdata/tomcat/catalina.log。

4. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

安装目录及数据目录检查

指标项名称：安装目录及数据目录检查

指标项含义：该指标项首先检查安装目录（默认为“/opt/Bigdata”）所在磁盘分区根目录下的lost+found目录。如果该目录下有**omm**用户的文件，则认为异常。节点异常时，会把相关的文件放入到“lost+found”目录。该检查主要是针对这类场景，检查文件是否丢失。然后，对安装目录（如：“/opt/Bigdata”）和数据目录（如：“/srv/BigData”）进行检查。如果目录下出现非**omm**用户的文件，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点，执行**su - omm**切换到**omm**用户。检查lost+found目录是否存在omm用户的文件或文件夹。

   如果有**omm**用户文件，建议对其进行恢复后重新检查；如果没有**omm**用户文件，则执行**2**。

2. 分别对安装目录和数据目录进行排查。查看目录下是否存在非**omm**用户是文件或文件夹。如果确认这些文件是手工生成的临时文件，建议对清理后重新检查。

CPU 使用率

指标项名称：CPU使用率

指标项含义：检查CPU使用率是否超过当前设定的阈值。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12016进行处理。

内存使用率

指标项名称：内存使用率

指标项含义：检查内存使用率是否超过当前设定的阈值。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12018进行处理。

主机磁盘使用率

指标项名称：主机磁盘使用率

指标项含义：检查主机磁盘使用率是否超过当前设定的阈值。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12017进行处理。

主机磁盘写速率

指标项名称：主机磁盘写速率

指标项含义：检查主机磁盘写速率。根据业务场景不同，主机磁盘写速率大小可能存在差异，所以该指标项只反映具体的数值大小，用户需根据业务场景具体判断该指标是否健康。

恢复指导：用户根据具体的业务场景，判断当前磁盘写速率是否正常。

主机磁盘读速率

指标项名称：主机磁盘读速率

指标项含义：检查主机磁盘读速率。根据业务场景不同，主机磁盘读速率大小可能存在差异，所以该指标项只反映具体的数值大小，用户需根据业务场景具体判断该指标是否健康。

恢复指导：用户根据具体的业务场景，判断当前磁盘读速率是否正常。

主机业务平面网络状态

指标项名称：主机业务平面网络状态

指标项含义：检查集群主机业务平面网络连通性。如果出现无法连通的情况，则认为不健康。

恢复指导：如果是单平面组网，对应需检查单平面的IP。双平面组网排查恢复步骤如下：

1. 检查主备管理节点业务平面IP的网络连通性。
   如果网络异常，执行**3**。
   如果网络正常，执行**2**。
2. 检查主管理节点IP到集群内异常节点IP的网络连通性。
3. 如果网络不通，请联系运维人员排查网络问题，以保证满足业务使用。

主机状态

指标项名称：主机状态

指标项含义：检查主机状态是否正常。如果节点有故障，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-12006进行处理。

检查告警

指标项名称： 检查告警

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.9 HDFS 健康检查指标项说明

发送包的平均时间统计

指标项名称：发送包的平均时间统计

指标项含义：HDFS文件系统中DataNode每次执行SendPacket的平均时间统计，如果大于2000000纳秒，则认为不健康。

恢复指导： 如果该指标项异常，则需要检查集群的网络速度是否正常、内存或CPU使用率是否过高。同时检查集群中HDFS负载是否过高。

服务健康状态

指标项名称：服务状态

指标项含义：检查HDFS服务状态是否正常。如果节点有故障，则认为不健康。

恢复指导：如果该指标项异常，建议检查KrbServer、LdapServer、ZooKeeper三个服务的状态是否为异常并处理。然后再检查是否是HDFS SafeMode ON导致的写文件失败，并使用客户端，确认是否无法在HDFS中写入数据，排查HDFS写数据失败的原因。最后参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查HDFS服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，请参见告警进行修复。

11.7.10 Hive 健康检查指标项说明

HiveServer 允许的最大 session 数量

指标项名称：Hive允许连接的最大session数量

指标项含义：检查Hive允许连接的最大session数量。

恢复指导：如果该指标项异常，建议参见告警进行处理。

已经连接到 HiveServer 的 session 数量

指标项名称：已经连接到HiveServer的session数量

指标项含义：检查Hive连接数。

恢复指导：如果该指标项异常，建议参见告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Hive服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.11 Kafka 健康检查指标项说明

Broker 可用节点数

标项名称：Broker数目

指标项含义：检查集群中可用的Broker节点数，若集群中可用的Broker节点数小于2，则认为不健康。

恢复指导：如果该指标项异常，进入Kafka服务实例页面，单击不可用Broker实例的“主机名”，在“概要信息”中查看主机的健康状态，若为“良好”，则参见“进程故障”告警进行处理；若不为“良好”，则参见“节点故障”告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Kafka服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见“Kafka服务不可用”告警进行处理。

检查告警

指标项名称：告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.12 KrbServer 健康检查指标项说明

KerberosAdmin 服务可用性检查

指标项名称：KerberosAdmin服务可用性

指标项含义：系统对KerberosAdmin服务状态进行检查，如果检查结果不正常，则KerberosAdmin服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是KerberosAdmin服务所在节点故障，或者SlapdServer服务不可用。操作人员进行KerberosAdmin服务恢复时，请尝试如下操作：

1. 检查KerberosAdmin服务所在节点是否故障。
2. 检查SlapdServer服务是否不可用。

KerberosServer 服务可用性检查

指标项名称： KerberosServer服务可用性

指标项含义：系统对KerberosServer服务状态进行检查，如果检查结果不正常，则KerberosServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是KerberosServer服务所在节点故障，或者SlapdServer服务不可用。操作人员进行KerberosServer服务恢复时，请尝试如下操作：

1. 检查KerberosServer服务所在节点是否故障。
2. 检查SlapdServer服务是否不可用。

服务健康状态

指标项名称：服务状态

指标项含义：系统对KrbServer服务状态进行检查，如果检查结果不正常，则KrbServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是KrbServer服务所在节点故障或者LdapServer服务不可用。详细操作请参见告警ALM-25500处理。

检查告警

指标项名称：告警信息

指标项含义：系统对KrbServer服务的告警信息进行检查。如果存在告警信息，则KrbServer服务可能存在异常。

恢复指导：如果该指标项检查结果不正常，建议根据告警内容，查看对应的告警资料，并进行相应的处理。

11.7.13 LdapServer 健康检查指标项说明

SlapdServer 服务可用性检查

指标项名称：SlapdServer服务可用性

指标项含义：系统对SlapdServer服务状态进行检查。如果检查结果不正常，则SlapdServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是SlapdServer服务所在节点故障或者SlapdServer进程故障。操作人员进行SlapdServer服务恢复时，请尝试如下操作：

1. 检查SlapdServer服务所在节点是否故障。详细操作请参见告警ALM-12006处理。
2. 检查SlapdServer进程是否正常。详细操作请参见告警ALM-12007处理。

服务健康状态

指标项名称：服务状态

指标项含义：系统对LdapServer服务状态进行检查。如果检查结果不正常，则LdapServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是主LdapServer服务所在节点故障或者主LdapServer进程故障。详细操作请参见告警ALM-25000处理。

检查告警

指标项名称：告警信息

指标项含义：系统对LdapServer服务的告警信息进行检查。如果存在告警信息，则LdapServer服务可能存在异常。

恢复指导：如果该指标项检查结果不正常，建议根据告警内容，查看对应的告警资料，并进行相应的处理。

11.7.14 Loader 健康检查指标项说明

ZooKeeper 健康状态

指标项名称：ZooKeeper健康状态

指标项含义：检查ZooKeeper健康状态是否正常。如果ZooKeeper服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

HDFS 健康状态

指标项名称：HDFS健康状态

指标项含义：检查HDFS健康状态是否正常。如果HDFS服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

DBService 健康状态

指标项名称：DBService健康状态

指标项含义：检查DBService健康状态是否正常。如果DBService服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

Yarn 健康状态

指标项名称：Yarn健康状态

指标项含义：检查Yarn健康状态是否正常。如果Yarn服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

MapReduce 健康状态

指标项名称：MapReduce健康状态

指标项含义：检查MapReduce健康状态是否正常。如果MapReduce服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

Loader 进程状态

指标项名称：Loader进程状态

指标项含义：检查Loader进程状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Loader服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

告警检查

指标项名称：告警信息

指标项含义：检查Loader服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.15 MapReduce 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查MapReduce服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.16 OMS 健康检查指标项说明

OMS 状态检查

指标项名称：OMS状态检查

指标项含义：OMS状态检查包括HA状态检查和资源状态检查。 HA状态取值为active、standby和NULL，分别表示主节点、备节点和未知。资源状态取值为normal、abnormal和NULL，分别表示正常、异常和未知。HA状态为NULL时，认为不健康；资源状态为NULL或abnormal时，认为不健康。

表 11-27 OMS 状态说明表

名称	说明
HA状态	active表示主节点 standby表示备节点 NULL表示未知
资源状态	normal表示所有资源都正常 abnormal表示有异常资源 NULL表示未知

恢复指导：

1. 登录主管理节点，执行**su - omm**切换到omm用户。执行**${CONTROLLER_HOME}/sbin/status-oms.sh**查看OMS状态。
2. 如果HA状态为NULL，可能是系统在重启，这个一般是中间状态，HA后续会自动调整为正常状态。
3. 如果资源状态异常，则说明有Manager的某些组件资源异常， 可具体查看acs、aos、cep、controller、feed_watchdog、fms、gaussDB、httpd、iam、ntp、okerberos、oldap、pms、tomcat等组件状态是否正常。
4. 如果Manager组件资源异常，参见Manager组件状态检查进行处理。

Manager 组件状态检查

指标项名称：Manager组件状态检查

指标项含义：Manager组件状态检查包括组件资源运行状态和资源HA状态。资源运行状态，取值为Normal、Abnormal等；资源HA状态，取值为Normal、Exception等。Manager组件包含acs、aos、cep、controller、feed_watchdog、floatip、fms、gaussDB、heartBeatCheck、httpd、iam、ntp、okerberos、oldap、pms、tomcat等。当运行状态和HA状态不是Normal时，认为指标不健康。

表 11-28 Manager 组件状态说明表

名称	说明
资源运行状态	Normal表示正常运行 Abnormal表示运行异常 Stopped表示停止 Unknown表示状态未知 Starting表示正在启动 Stopping表示正在停止 Active_normal表示主正常运行 Standby_normal表示备正常运行 Raising_active表示正在升主 Lowing_standby表示正在降备 No_action表示没有该动作 Repairing表示正在修复 NULL表示未知
资源HA状态	Normal表示正常 Exception表示故障 Non_steady表示非稳态 Unknown表示未知 NULL表示未知

恢复指导：

1. 登录主管理节点，执行**su - omm**切换到omm用户。执行**${CONTROLLER_HOME}/sbin/status-oms.sh**查看OMS状态。
2. 如果floatip、okerberos、oldap等异常，可参见告警ALM-12002、ALM-12004、ALM-12005分别进行处理。
3. 如果是其他资源异常，建议查看相关异常模块的日志。

   controller资源异常：查看异常节点的/var/log/Bigdata/controller/controller.log。

   cep资源异常：查看异常节点的/var/log/Bigdata/omm/oms/cep/cep.log。

   aos资源异常：查看异常节点的/var/log/Bigdata/controller/aos/aos.log。

   feed_watchdog资源异常：查看异常节点的/var/log/Bigdata/watchdog/watchdog.log。

   httpd资源异常：查看异常节点的/var/log/Bigdata/httpd/error_log。

   fms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/fms/fms.log。

   pms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/pms/pms.log。

   iam资源异常：查看异常节点的/var/log/Bigdata/omm/oms/iam/iam.log。

   gaussDB资源异常：查看异常节点的/var/log/Bigdata/omm/oms/db/omm_gaussdba.log。

ntp资源异常：查看异常节点的/var/log/Bigdata/omm/oms/ha/scriptlog/ha_ntp.log。

tomcat资源异常：查看异常节点的/var/log/Bigdata/tomcat/catalina.log。

4. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

OMA 运行状态

指标项名称：OMA运行状态

指标项含义：检查OMA的运行状态，状态结果包括运行和停止两种状态，如果OMA状态为停止，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点，然后执行**su - omm**切换到**omm**用户。
2. 执行**${OMA_PATH}/restart_oma_app**，手工启动OMA，然后重新检查。如果检查结果仍然不健康，则执行**3**。
3. 如果手工启动OMA无法恢复，建议查看分析OMA日志“/var/log/Bigdata/omm/oma/omm_agent.log”。
4. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

各节点与主管理节点之间 SSH 互信

指标项名称：各节点与主管理节点之间SSH互信

指标项含义：检查SSH互信是否正常。如果使用omm用户，在主管理节点可以通过SSH登录其他节点且不需要输入密码，则认为健康；否则，不健康。或者主管理节点SSH可以直接登录其他节点，但在其他节点无法通过SSH登录主管理节点，则也认为不健康。

恢复指导：

1. 如果该指标项检查异常，表示各节点与主管理节点之间SSH互信异常。SSH互信异常时，首先检查“/home/omm”目录的权限是否为**omm**。非**omm**的目录权限可能导致SSH互信异常，建议执行**chown omm:wheel**修改权限后重新检查。如果“/home/omm”目录权限正常，则执行**2**。
2. SSH互信异常一般会导致Controller和NodeAgent之间心跳异常，进而出现节点故障的告警。这时可参见告警ALM-12006进行处理。

进程运行时间

指标项名称：NodeAgent运行时间、Controller运行时间和Tomcat运行时间

指标项含义：检查NodeAgent、Controller、Tomcat进程的运行时间。如果小于半小时（即1800s），则进程可能重启过，建议半小时后再检查。如果多次检查，进程的运行时间都小于半小时，说进程状态异常。

恢复指导：

1. 登录检查结果不健康的节点，执行**su - omm**切换到**omm**用户。
2. 根据进程名称查看进程pid，执行命令：

   ps -ef | grep NodeAgent

3. 根据pid查看进程启动时间，执行命令：

   ps -p pid -o lstart

4. 判断进程启动时间是否正常。如果进程一直反复重启，执行**5**
5. 查看对应模块日志，分析重启原因。

   NodeAgent运行时间异常，检查相关日志**/var/log/Bigdata/nodeagent/agentlog/agent.log**。

   Controller运行时间异常，检查相关日志**/var/log/Bigdata/controller/controller.log**。

   Tomcat运行时间异常，检查相关日志**/var/log/Bigdata/tomcat/web.log**。

6. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

帐户及密码过期检查

指标项名称：帐户及密码过期检查

指标项含义：该指标项检查MRS的两个操作系统用户**omm**和**ommdba**。对操作系统用户，同时检查帐户及密码的过期时间。如果帐户或密码有效期小于等于15天，则认为不健康。

恢复指导：如果帐户或密码有效期小于等于15天，建议及时联系运维人员处理。

11.7.17 Spark 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查Spark服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-28001进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.18 Storm 健康检查指标项说明

工作节点数

指标项名称：Supervisor数

指标项含义：检查集群中可用的Supervisor数，若集群中可用的Supervisor数小于1，则认为不健康。

恢复指导：如果该指标项异常，进入Streaming服务实例页面，单击不可用Supervisor实例的“主机名”，在“概要信息”中查看主机的健康状态，若为“良好”，则参见“ALM-12007 进程故障”告警进行处理；若不为“良好”，则参见“ALM-12006 节点故障”告警进行处理。

空闲 Slot 数

指标项名称：空闲Slot数

指标项含义：检查集群中空闲的Slot数，若集群中空闲slot数目小于1，则认为不健康。

恢复指导：如果该指标项异常，进入Storm服务实例页面，查看Supervisor实例的“健康状态”，若均为“良好”，则需要扩容集群Core节点；若不为良好，则参见“ALM-12007 进程故障”告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Storm服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见“ALM-26051 Storm服务不可用”告警进行处理。

检查告警

指标项名称：告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.19 Yarn 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查Yarn服务状态是否正常。如果当前无法获取NodeManager节点数时，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理并确认网络无异常。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.7.20 ZooKeeper 健康检查指标项说明

ZooKeeper 服务处理请求平均延时

指标项名称：ZooKeeper服务处理请求平均延时

指标项含义：检查ZooKeeper服务处理请求的平均延时，如果大于300毫秒，则认为不健康。

恢复指导： 如果该指标项异常，则需要检查集群的网络速度是否正常、内存或CPU使用率是否过高。

ZooKeeper 连接数使用率

指标项名称：ZooKeeper连接数使用率

指标项含义：检查ZooKeeper内存使用率是否超过80%。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，建议增加ZooKeeper服务可以使用的内存。可以通过ZooKeeper服务配置中的“GC_OPTS”配置项参数-Xmx来修改，修改完成需重启ZooKeeper服务。

服务健康状态

指标项名称：服务状态

指标项含义：检查ZooKeeper服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议检查KrbServer、LdapServer两个服务的健康状态是否为故障并进行处理。然后登录ZooKeeper客户端，确认ZooKeeper是否无法写入数据，根据错误提示排查ZooKeeper写数据失败的原因。最后参告警ALM-13000进行处理。

检查告警

指标项名称：告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

11.8 静态服务池管理

11.8.1 查看静态服务池状态

操作场景

MRS Manager支持通过静态服务资源池对没有运行在YARN上的服务资源进行管理和隔离。支持动态地管理HDFS和YARN在部署节点可使用的CPU、I/O和内存总量。系统支持基于时间的静态服务资源池自动调整策略，使集群在不同的时间段自动调整参数值，从而更有效地利用资源。

用户可以在MRS Manager查看静态服务池各个服务使用资源的监控指标结果，包含监控指标如下：

- 服务总体CPU使用率
- 服务总体磁盘I/O读速率
- 服务总体磁盘I/O写速率
- 服务总体内存使用大小

操作步骤

步骤1 在MRS Manager，单击“系统设置”，在“资源管理”区域单击“静态服务池”。

步骤2 单击“状态”。

步骤3 查看系统资源调整基数。

- “系统资源调整基数”表示集群中每个节点可以被集群服务使用的最大资源。如果节点只有一个服务，则表示此服务独占节点可用资源。如果节点有多个服务，则表示所有服务共同使用节点可用资源。
- “CPU(%)”表示节点中服务可使用的最大CPU。
- “Memory(%)”表示节点中服务可使用的最大内存。

步骤4 查看集群服务资源使用状态。

在图表区域的服务选择框中选择“所有服务”，则“图表”中会显示服务池所有服务的资源使用状态。

说明

“生效的配置组”表示集群服务当前使用的资源控制配置组。默认情况下每天所有时间均使用“default”配置组，表示集群服务可以使用节点全部CPU，以及70%的内存。

步骤5 查看单个服务资源使用状态。

在图表区域的服务选择框中选择指定服务，“图表”中会显示服务池此服务的资源使用状态。

步骤6 用户可以选择页面自动刷新间隔的设置。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

步骤7 在“时间区间”选择需要查看服务资源的时间段。可供选择的选项如下：

- 实时
- 最近3小时
- 最近6小时
- 最近24小时
- 最近一周
- 最近一个月
- 最近三个月
- 最近六个月
- 自定义：选择自定义时，在时间范围内自行选择需要查看的时间。

步骤8 单击“查看”可以查看相应时间区间的服务资源数据。

步骤9 自定义服务资源报表。

1. 单击“定制”，勾选需要显示的服务源指标。
   - 服务总体磁盘I/O读速率

- 服务总体内存使用大小
- 服务总体磁盘I/O写速率
- 服务总体CPU使用率

2. 单击“确定”保存并显示所选指标。

说明

单击“清除”可批量取消全部选中的指标项。

步骤10 导出监控指标报表。

单击“导出”，Manager将生成指定时间范围内、已勾选的服务资源指标报表文件，请选择一个位置保存，并妥善保管该文件。

说明

如果需要查看指定时间范围的监控指标对应的分布曲线图，请单击“查看”，界面将显示用户自定义时间范围内选定指标的分布曲线图。

----结束

11.8.2 配置静态服务池

操作场景

当需要控制集群服务可以使用节点的资源情况，或者在不同时间段集群服务使用节点的CPU不同，用户可以在MRS Manager调整资源基数，并自定义资源配置组。

前提条件

- 配置静态服务池后，HDFS和YARN服务需要重启，重启期间服务不可用。
- 配置静态服务池后，各服务及角色实例使用的最大资源将不能超过限制。

操作步骤

步骤1 修改系统资源调整基数。

1. 在MRS Manager界面，单击“系统设置”，在“资源管理”区域单击“静态服务池”。
2. 单击“配置”，显示服务池配置组管理页面。
3. 在“系统资源调整基数”分别修改参数“CPU(%)”和“Memory(%)”。

   修改“系统资源调整基数”将限制Flume、HBase、HDFS、Impala和YARN服务能够使用节点的最大物理CPU和内存资源百分比。如果多个服务部署在同一节点，则所有服务使用的最大物理资源百分比不能超过此参数值。
4. 单击“下一步”完成编辑。

   需要重新修改参数，可单击页面下方的“上一步”。

步骤2 修改服务池默认“default”配置组。

1. 单击“default”，在“服务池配置”表格中Flume、HBase、HDFS、Impala和YARN服务对应的“CPU LIMIT(%)”、“CPU SHARE(%)”、“I/O(%)”和“Memory(%)”填写各服务的资源使用百分比数量。

说明

- 所有服务使用的“CPU LIMIT(%)”资源配置总和可以大于100%。
- 所有服务使用的“CPU SHARE(%)”和“I/O(%)”资源配置总和需为100%。例如为HDFS和Yarn服务分配使用的CPU资源，2个服务分配到的CPU资源总和为100%。
- 所有服务使用的“Memory(%)”资源配置总和可以小于或等于100%，也可以大于100%。
- “Memory(%)”不支持动态生效，仅在“default”配置组中可以修改。

2. 单击页面空白处完成编辑，MRS Manager将根据集群硬件资源与分配情况，在“详细配置”生成服务池参数的正确配置值。
3. 如果根据业务需要，可以单击“详细配置”右侧的修改服务池的参数值。

   在“服务池配置”单击指定的服务名，“详细配置”将只显示此服务的参数。手工修改参数值并不会刷新服务使用资源的百分比显示。支持动态生效的参数，在新增加的配置组中显示名将包含配置组的编号，例如“HBase : RegionServer : dynamic-config1.RES_CPUSET_PERCENTAGE”，参数作用与“default”配置组中的参数相同。

表 11-29 静态服务池参数一览

参数名	参数描述
– RES_CPUSET_PERCENTAGE – dynamic-configX.RES_CPUSET_PERCENTAGE	配置服务使用CPU PERCENTAGE。
– RES_CPU_SHARE – dynamic-configX.RES_CPU_SHARE	配置服务使用CPU share。
– RES_BLKIO_WEIGHT – dynamic-configX.RES_BLKIO_WEIGHT	配置服务占用I/O的权重。
HBASE_HEAPSIZE	配置RegionServer的JVM最大内存。
HADOOP_HEAPSIZE	配置DataNode的JVM最大内存。
yarn.nodemanager.resource.memory-mb	配置当前节点上NodeManager可使用的内存大小。
dfs.datanode.max.locked.memory	配置DataNode用做HDFS缓存的最大内存。
FLUME_HEAPSIZE	配置每个flume实例能使用的最大JVM内存。
IMPALAD_MEM_LIMIT	配置impalad实例可使用的最大内存。

步骤3 添加自定义资源配置组。

1. 是否需要根据时间自动调整资源配置?

   是，执行**步骤3.2**。

否，执行**步骤4**。

2. 单击增加新的资源配置组。在“调度时间”，单击显示时间策略配置页面。

   根据业务需要修改以下参数，并单击“确定”保存：

   - “重复”：当勾选“重复”时表示此资源配置组按调度周期重复运行。不勾选时请设置一个资源配置组应用的日期与时间。
   - “重复策略”：支持“每天”、“每周”和“每月”。仅在“重复”模式中生效。
   - “介于”：表示资源配置组应用的开始与结束时间。请设置一个唯一的时间区间，如果与已有配置组的时间区间有重叠，则无法保存。仅在“重复”模式中生效。

   说明

   - “default”配置组会在所有未定义的时间段内生效。
   - 新增加的配置组属于动态生效的配置项集合，在配置组应用的时间区间内可直接生效。
   - 新增加的配置组可以被删除。最多增加4个动态生效的配置组。
   - 选择任一种“重复策略”，如果结束时间小于开始时间，默认标识为第二天的时间。例如“22:00”到“6:00”表示调度时间为当天22点到第二天6点。
   - 若多个配置组的“重复策略”类型不相同，则时间区间可以重叠，且生效的策略优先级从低到高的顺序为“每天”、“每周”、“每月”。例如，有“每月”与“每天”的调度配置组，时间区间分别为4:00到7:00，6:00到8:00，此时以每月的配置组为准。
   - 若多个配置组的“重复策略”类型相同，当日期不相同时，则时间区间可以重叠。例如，有两个“每周”的调度配置组，可以分别指定时间区间为周一和周三的4:00到7:00。

3. 在“服务池配置”修改各服务资源配置，并单击页面空白处完成编辑，然后执行**步骤4**。

   用户可单击“服务池配置”右侧的重新修改参数。如果根据业务需要，在“详细配置”单击，手动更新由系统生成的参数值。

步骤4 保存配置。

单击“保存”，在“保存配置”窗口勾选“重新启动受影响的服务或实例。”，单击确定保存并重启相关服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

11.9 租户管理

11.9.1 租户简介

定义

MRS集群拥有的不同资源和服务支持多个组织、部门或应用共享使用。集群提供了一个逻辑实体来统一使用不同资源和服务，这个逻辑实例就是租户。多个不同的租户统称多租户。当前仅分析集群支持租户。

原理

MRS集群提供多租户的功能，支持层级式的租户模型，支持动态添加和删除租户，实现资源的隔离，可以对租户的计算资源和存储资源进行动态配置和管理。

计算资源指租户Yarn任务队列资源，可以修改任务队列的配额，并查看任务队列的使用状态和使用统计。

存储资源目前支持HDFS存储，可以添加删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额。

MRS Manager作为MRS集群的统一租户管理平台，可以为企业提供成熟的多租户管理模型，实现集中式的租户和业务管理。租户可以在界面上根据业务需要，在集群中创建租户、管理租户。

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。默认情况下，新的计算资源和存储资源的全部权限将分配给租户的角色。
- 默认情况下，查看当前租户的资源、在当前租户中添加子租户并管理子租户资源的权限将分配给租户的角色。
- 修改租户的计算资源或存储资源，对应的角色关联权限将自动更新。

MRS Manager中最多支持512个租户。系统默认创建的租户包含“default”。和默认租户同处于最上层的租户，可以统称为一级租户。

资源池

YARN任务队列支持一种调度策略，称为标签调度（Label Based Scheduling）。通过此策略，YARN任务队列可以关联带有特定节点标签（Node Label）的NodeManager，使YARN任务在指定的节点运行，实现任务的调度与使用特定硬件资源的需求。例如，需要使用大量内存的YARN任务，可以通过标签关联具有大量内存的节点上运行，避免性能不足影响业务。

在MRS集群中，租户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个资源池。YARN任务队列通过配置队列容量策略，与指定的资源池进行关联，可以更有效地使用资源池中的资源，且互不影响。

MRS Manager中最多支持50个资源池。系统默认包含一个“Default”资源池。

11.9.2 添加租户

操作场景

当租户需要根据业务需求指定资源使用情况时，可以在MRS Manager创建租户。

前提条件

- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的完整存储路径在HDFS目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“添加租户”，打开添加租户的配置页面，参见以下表格内容为租户配置属性。

表 11-30 租户参数一览表

参数名	描述
“名称”	指定当前租户的名称，长度为3到20，可包含数字、字母和下划线。
“租户类型”	可选参数值为“叶子租户”和“非叶子租户”。当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
“动态资源”	为当前租户选择动态计算资源。系统将自动在Yarn中以租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。
“默认资源池容量 (%)”	配置当前租户在“default”资源池中使用的计算资源百分比。
“默认资源池最大容量 (%)”	配置当前租户在“default”资源池中使用的最大计算资源百分比。
“储存资源”	为当前租户选择存储资源。系统将自动在“/tenant”目录中以租户名称创建文件夹。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。
“存储空间配额 (MB)”	配置当前租户使用的HDFS存储空间配额。取值范围为“1”到“8796093022208”。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间表示所有副本文件在HDFS中占用的磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
“存储路径”	配置租户在HDFS中的存储目录。系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“tenant/ta1”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
“服务”	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
“描述”	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配Manager_tenant角色以及租户对应的角色。具体操作请参见**创建用户**。

----结束

相关任务

查看已添加的租户

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，单击已添加租户的名称。

默认在右侧显示“概述”页签。

步骤3 查看当前租户的“基本信息”、“资源配额”和“统计”。

如果HDFS处于“已停止”状态，“资源配额”中“Space”的“可用”和“已使用”会显示为“unknown”。

----结束

11.9.3 添加子租户

操作场景

当租户需要根据业务需求，将当前租户的资源进一步分配时，可以在MRS Manager添加子租户。

前提条件

- 已添加上级租户。
- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果子租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的存储目录在父租户的存储目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，将光标移动到需要添加子租户的租户节点上，单击“添加子租户”，打开添加子租户的配置页面，参见以下表格内容为租户配置属性。

表 11-31 子租户参数一览表

参数名	描述
“父租户”	显示上级父租户的名称。
“名称”	指定当前租户的名称，长度为3到20，可包含数字、字母和下划线。
“租户类型”	可选参数值为“叶子租户”和“非叶子租户”，当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
“动态资源”	为当前租户选择动态计算资源。系统将自动在Yarn父租户队列中以子租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。如果父租户未选择动态资源，子租户也无法使用动态资源。
“默认资源池容量 (%)”	配置当前租户使用的资源百分比，基数为父租户的资源总量。
“默认资源池最大容量 (%)”	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。
“储存资源”	为当前租户选择存储资源。系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。如果父租户未选择存储资源，子租户也无法使用存储资源。
“存储空间配额 (MB)”	配置当前租户使用的HDFS存储空间配额。最小值值为“1”，最大值为父租户的全部存储配额。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。若此配额大于父租户的配额，实际存储量受父租户配额影响。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“父租户中分配资源”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
“存储路径”	配置租户在HDFS中的存储目录。系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“tenant/ta1”，系统默认自动配置此参数值为“tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。支持在父目录中自定义存储路径。存储路径的父目录必需是父租户的存储目录。

参数名	描述
“服务”	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
“描述”	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成子租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配租户对应的角色。具体操作请参见**创建用户**。

----结束

11.9.4 删除租户

操作场景

当租户需要根据业务需求，将当前不再使用的租户删除时，可以在MRS Manager完成操作。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。该任务对应取消角色与用户的绑定，请参见**修改用户信息**。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，将光标移动到需要删除的租户节点上，单击“删除”。

界面显示“删除租户”对话框。根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户的数据”，否则将自动删除租户对应的存储空间。

步骤3 单击“确定”保存，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

- 租户删除后，Yarn中对应的租户任务队列不会被删除。
- 删除父租户时选择不保留数据，如果存在子租户且子租户使用了存储资源，则子租户的数据也会被删除。

----结束

11.9.5 管理租户目录

操作场景

用户根据业务需求，可以在MRS Manager对指定租户使用的HDFS存储目录，进行管理操作。支持用户对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

已添加关联了HDFS存储资源的租户。

操作步骤

- 查看租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击目标的租户。
  c. 单击“资源”页签。
  d. 查看“HDFS存储”表格。
     - 指定租户目录的“文件目录数上限”列表示文件和目录数量配额。
     - 指定租户目录的“存储空间配额 (MB)”列表示租户目录的存储空间大小。
- 添加租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击需要添加HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，单击“添加目录”。
     - “父目录”选择一个父租户的存储目录。
       该参数仅适用于子租户。如果父租户有多个目录，请选择其中任何一个。
     - “路径”填写租户目录的路径。

       说明

       - 如果当前租户不是子租户，新路径将在HDFS的根目录下创建。
       - 如果当前租户是一个子租户，新路径将在指定的目录下创建。

       完整的HDFS存储目录最多包含1023个字符。HDFS目录名称包含数字、大小写字母、空格和下划线。空格只能在HDFS目录名称的中间使用。

- “文件\目录数上限”填写文件和目录数量配额。
  “文件\目录数上限”为可选参数，取值范围从1到9223372036854775806。
- “存储空间配额 (MB)”填写租户目录的存储空间大小。
  “存储空间配额 (MB)”的取值范围从1到8796093022208。

  说明

  为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

e. 单击“确定”完成租户目录添加，系统将在HDFS根目录下创建租户的目录。

- 修改租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击需要修改HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。
     - “文件\目录数上限”填写文件和目录数量配额。
       “文件\目录数上限”为可选参数，取值范围从1到9223372036854775806。
     - “存储空间配额”填写租户目录的存储空间大小。
       “存储空间配额”的取值范围从1到8796093022208。

       说明

       为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

  e. 单击“确定”完成租户目录修改。
- 删除租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击需要删除HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，指定租户目录的“操作”列，单击“删除”。
     创建租户时设置的默认HDFS存储目录不支持删除，仅支持删除新添加的HDFS存储目录。
  e. 单击“确定”完成租户目录删除。

11.9.6 恢复租户数据

操作场景

租户的数据默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，需要手动恢复。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“基本信息”左侧的圆圈，绿色表示租户可用，灰色表示租户不可用。
2. 单击“资源”，查看“Yarn”或者“HDFS存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击“恢复租户数据”。

步骤5 在“恢复租户数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

11.9.7 添加资源池

操作场景

在MRS集群中，用户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个YARN资源池。每个NodeManager只能属于一个资源池。系统中默认包含了一个名为“Default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

该任务指导用户通过MRS Manager添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 单击“添加资源池”。

步骤4 在“添加资源池”设置资源池的属性。

- “名称”：填写资源池的名称。不支持创建名称为“Default”的资源池。

  资源池的名称，长度为1到20位，可包含数字、字母和下划线，且不能以下划线开头。

- “可用主机”：在界面左边主机列表，选择指定的主机名称，单击  ，将选中的主机加入资源池。只支持选择本集群中的主机。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

步骤6 完成资源池创建后，用户可以在资源池的列表中查看资源池的“名称”、“成员”、“类型”、“虚拟核数”与“内存”。已加入自定义资源池的主机，不再是“Default”资源池的成员。

----结束

11.9.8 修改资源池

操作场景

该任务指导用户通过MRS Manager，修改已有资源池中的成员。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“修改”。

步骤4 在“编辑资源池”修改“已添加主机”。

- 增加主机：在界面左边主机列表，选择指定的主机名称，单击 

，将选中的主机加入资源池。
- 删除主机：在界面右边主机列表，选择指定的主机名称，单击 

，将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

11.9.9 删除资源池

操作场景

该任务指导用户通过MRS Manager，删除已有资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列配置**。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

在弹出窗口中单击“确定”。

----结束

11.9.10 配置队列

操作场景

用户根据业务需求，可以在MRS Manager修改指定租户的队列配置。

前提条件

已添加关联Yarn并分配了动态资源的租户。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“动态资源计划”页签。

步骤3 单击“队列配置”页签。

步骤4 在租户队列表格，指定租户队列的“操作”列，单击“修改”。

说明

在“租户管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击 也能打开修改队列配置页面。

表 11-32 队列配置参数

参数名	描述
“最大应用数量”	表示最大应用程序数量。取值范围从“1”到“2147483647”。
“AM最大资源百分比”	表示集群中可用于运行application master的最大资源占比。取值范围从“0”到“1”。
“用户资源最小上限百分比 (%)”	表示用户使用的最小资源上限百分比。取值范围从“0”到“100”。
“用户资源上限因子”	表示用户使用的最大资源限制因子，与当前租户在集群中实际资源百分比相乘，可计算出用户使用的最大资源百分比。最小值为“0”。
“状态”	表示资源计划当前的状态，“运行”为运行状态，“停止”为停止状态。
“默认资源池”	表示队列使用的资源池。默认为“Default”，如果需要修改为其他资源，需要先配置队列容量，请参见**配置资源池的队列容量策略**。

----结束

11.9.11 配置资源池的队列容量策略

操作场景

添加资源池后，需要为YARN任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。每个队列只能配置一个资源池的队列容量策略。用户可以在任何一个资源池中查看队列并配置队列容量策略。配置队列策略后，YARN任务队列与资源池形成关联关系。

该任务指导用户通过MRS Manager配置队列策略。

前提条件

- 已添加资源池。
- 任务队列与其他资源池无关联关系。默认情况下，所有队列与“Default”资源池存在关联关系。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“动态资源计划”页签。

步骤3 在“资源池”选择指定的资源池。

“可用资源配额”：表示每个资源池默认所有资源都可分配给队列。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“修改”。

步骤5 在“修改资源分配”窗口设置任务队列在此资源池中的资源容量策略。

- “资源容量 (%)”：表示当前租户计算资源使用的资源百分比。
- “最大资源容量 (%)”：表示当前租户计算资源使用的最大资源百分比。

步骤6 单击“确定”保存配置。

----结束

11.9.12 清除队列配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，用户可以在MRS Manager清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

操作步骤

步骤1 在MRS Manager界面，单击“租户管理”。

步骤2 单击“动态资源计划”页签。

步骤3 在“资源池”选择指定的资源池。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“清除”。

在“清除队列配置”中单击“确定”，清除队列在当前资源池的配置。

说明

如果用户未配置队列的资源容量策略，则清除功能默认不可用。

----结束

11.10 备份与恢复

11.10.1 备份与恢复简介

概述

MRS Manager提供对系统内的用户数据及系统数据的备份恢复能力，备份功能按组件提供，支持备份管理系统Manager的数据（需要同时备份OMS和LdapServer）、Hive用户数据、DBService中保存的组件元数据和HDFS元数据备份。

备份恢复任务的使用场景如下：

- 用于日常备份，确保系统及组件的数据安全。
- 当系统故障导致无法工作时，使用已备份的数据完成恢复操作。
- 当主集群完全故障，需要创建一个与主集群完全相同的镜像集群，可以使用已备份的数据完成恢复操作。

表 11-33 根据业务需要备份元数据

备份类型	备份内容
OMS	默认备份集群管理系统中的数据库数据（不包含告警数据）以及配置数据。
LdapServer	备份用户信息，包括用户名、密码、密钥、密码策略、组信息。
DBService	备份DBService管理的组件（Hive）的元数据。
NameNode	备份HDFS元数据。

原理

任务

在进行备份恢复之前，需要先创建备份恢复任务，并指定任务的参数，例如任务名称、备份数据源和备份文件保存的目录类型等等。通过执行备份恢复任务，用户可完成数据的备份恢复需求。在使用Manager执行恢复HDFS、Hive和NameNode数据时，无法访问集群。

每个备份任务可同时备份不同的数据源，每个数据源将生成独立的备份文件，每次备份的所有备份文件组成一个备份文件集，可用于恢复任务。备份任务支持将备份文件保存在Linux本地磁盘、本集群HDFS与备集群HDFS中。备份任务提供全量备份或增量备份的策略，增量备份策略支持HDFS和Hive备份任务，OMS、LdapServer、DBService和NameNode备份任务默认只应用全量备份策略。

说明

任务运行规则：

- 某个任务已经处于执行状态，则当前任务无法重复执行，其他任务也无法启动。
- 周期任务自动执行时，距离该任务上次执行的时间间隔需要在120秒以上，否则任务推迟到下个周期启动。手动启动任务无时间间隔限制。
- 周期任务自动执行时，当前时间不得晚于任务开始时间120秒以上，否则任务推迟到下个周期启动。
- 周期任务锁定时无法自动执行，需要手动解锁。
- OMS、LdapServer、DBService和NameNode备份任务开始执行前，若主管理节点“LocalBackup”分区可用空间小于20GB，则无法开始执行。
- 用户在规划备份恢复任务时，请严格根据业务逻辑、数据存储结构、数据库或表关联关系，选择需要备份或者恢复的数据。系统默认创建了一个间隔为24小时的周期备份任务“default”，支持全量备份OMS、LdapServer、DBService和NameNode数据到Linux本地磁盘。

规格

表 11-34 备份恢复特性规格

项目	规格
备份或恢复任务最大数量（个）	100
同时运行的任务数量（个）	1
等待运行的任务最大数量（个）	199
Linux本地磁盘最大备份文件大小（GB）	600

表 11-35 “default”任务规格

项目	OMS	LdapServer	DBService	NameNode
备份周期	1小时			
最大备份数	2个			
单个备份文件最大大小	10MB	20MB	100MB	1.5GB
最大占用磁盘大小	20MB	40MB	200MB	3GB
备份数据保存位置	主备管理节点“*数据存放路径*/LocalBackup/”			

说明

“default”任务保存的备份数据，请用户根据企业运维要求，定期转移并保存到集群外部。

11.10.2 备份元数据

操作场景

为了确保元数据信息安全，或者用户需要对元数据功能进行重大操作（如扩容缩容、安装补丁包、升级或迁移等）前后，需要对元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。元数据包含OMS数据、LdapServer数据、DBService数据和NameNode数据。备份Manager数据包含同时备份OMS数据和LdapServer数据。

默认情况下，元数据备份由“default”任务支持。该任务指导用户通过MRS Manager创建备份任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 需要准备一个用于备份数据的备集群，且网络连通。每个集群的安全组，需分别添加对端集群的安全组入方向规则，允许安全组中所有弹性云服务器全部协议全部端口的访问请求。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“数据存放路径/LocalBackup/”是否有充足的空间。

操作步骤

步骤1 创建备份任务。

1. 在MRS Manager，选择“系统设置 > 备份管理”。
2. 单击“创建备份任务”。

步骤2 设置备份策略。

1. 在“任务名称”填写备份任务的名称。
2. 在“备份类型”选择备份任务的运行类型，“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

   创建周期备份任务，还需要填写以下参数：

   - “开始时间”：表示任务第一次启动的时间。
   - “周期”：表示任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
   - “备份策略”：表示任务每次启动时备份的数据量。支持“首次全量备份，后续增量备份”、“每次都全量备份”和“每n次进行一次全量备份”。选择“每n次进行一次全量备份”时，需要指定**n**的值。

步骤3 选择备份源。

在“备份配置”，勾选元数据选项，例如“OMS”和“LdapServer”。

步骤4 设置备份参数。

1. 在“OMS”和“LdapServer”的“路径类型”，选择一个备份目录的类型。

   备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“数据存放路径/LocalBackup/”。选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “目标实例名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

2. 单击“确定”保存。

步骤5 执行备份任务。

在备份任务列表中已创建任务的“操作”列，若“备份类型”选择“周期备份”请单击“即时备份”，若“备份类型”选择“手动备份”请单击“启动”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为备份任务名_任务创建时间，用于保存数据源的备份文件。备份文件的名称为版本号_数据源_任务执行时间.tar.gz。

----结束

11.10.3 恢复元数据

操作场景

在用户意外修改删除、数据需要找回，对元数据组件进行重大操作（如升级、重大数据调整等）后系统数据出现异常或未达到预期结果，模块全部故障完全无法使用，或者迁移数据到新集群的场景中，需要对元数据进行恢复操作。

该任务指导用户通过MRS Manager创建恢复元数据任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的元数据。
- 必须使用同一时间点的OMS和LdapServer备份数据进行恢复，否则可能造成业务和操作失败。
- MRS集群默认使用DBService保存Hive的元数据。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。

前提条件

- 检查OMS和LdapServer备份文件是否是同一时间点备份的数据。
- 检查OMS资源状态是否正常，检查LdapServer实例状态是否正常。如果不正常，不能执行恢复操作。
- 检查集群主机和服务的状态是否正常。如果不正常，不能执行恢复操作。
- 检查恢复数据时集群主机拓扑结构与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份。
- 检查恢复数据时集群中已添加的服务与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份
- 检查DBService主备实例状态是否正常。如果不正常，不能执行恢复操作。
- 停止依赖MRS集群运行的上层业务应用。
- 在MRS Manager停止所有待恢复数据的NameNode角色实例，其他的HDFS角色实例必须保持正常运行，恢复数据后重启NameNode。NameNode角色实例重启前无法访问。
- 检查NameNode备份文件是否保存在主管理节点“数据存放路径/LocalBackup/”。

操作步骤

步骤1 查看备份数据位置。

1. 在MRS Manager，选择“系统设置 > 备份管理”。
2. 在任务列表指定任务的“操作”列，单击“更多 > 查询历史”，打开备份任务执行历史记录。 在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：
   - “备份对象”表示备份的数据源。
   - “备份路径”表示备份文件保存的完整路径。
3. 选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤2 创建恢复任务。

1. 在MRS Manager，选择“系统设置 > 恢复管理”。
2. 单击“创建恢复任务”。
3. 在“任务名称”填写恢复任务的名称。

步骤3 选择恢复源。

在“恢复配置”，勾选待恢复数据的元数据组件。

步骤4 设置恢复参数。

1. 在“路径类型”，选择一个备份目录的类型。
2. 选择不同的备份目录时，对应设置如下：
   - “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。选择此参数值，还需要配置“源端路径”，表示备份文件保存位置的完整路径。例如，

“数据存放路径/LocalBackup/备份任务名_任务创建时间/数据源_任务执行时间/版本号_数据源_任务执行时间.tar.gz”。

- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz”。
  - “源实例名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。

3. 单击“确定”保存。

步骤5 执行恢复任务。

在恢复任务列表已创建任务的“操作”列，单击“启动”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“启动”，重试恢复任务。

步骤6 恢复了哪个元数据?

- 恢复了OMS和LdapServer元数据，执行**步骤7**。
- 恢复了DBService数据，任务结束。
- 恢复NameNode数据，在MRS Manager，选择“服务管理 > HDFS > 更多 > 重启服务”，任务结束。

步骤7 重启Manager使恢复数据生效。

1. 在MRS Manager，选择“LdapServer > 更多 > 重启服务”，单击“确定”，等待LdapServer服务重启成功。
2. 登录主管理节点，详情请参见**如何确认Manager的主备管理节点**。
3. 执行以下命令，重新启动OMS。

   sh ${BIGDATA_HOME}/om-0.0.1/sbin/restart-oms.sh

   提示以下信息表示命令执行成功：

   ```
   start HA successfully.
   ```

4. 在MRS Manager，选择“KrbServer > 更多 > 同步配置”，不勾选“重启配置过期的服务或实例”，单击“确定”，等待KrbServer服务配置同步及重启成功。
5. 选择“服务管理 > 更多 > 同步配置”，不勾选“重启配置过期的服务或实例”，单击“确定”，等待集群配置同步成功。
6. 选择“服务管理 > 更多 > 停止集群”。待停止集群的操作生效后，选择“服务管理 > 更多 > 启动集群”，等待集群启动成功。

----结束

11.10.4 修改备份任务

操作场景

该任务指导用户通过MRS Manager修改已创建的备份任务的配置参数，以适应业务需求的变化。不支持修改任何恢复任务配置参数，只能查看恢复任务的配置参数。

对系统的影响

修改备份任务后，新的参数在下一次执行任务时生效。

前提条件

- 已创建备份任务。
- 已根据业务实际需求，规划新的备份任务策略。

操作步骤

步骤1 在MRS Manager，选择“系统设置 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，单击“修改”，打开修改配置页面。

步骤3 在新页面中修改任务参数。

- 手动备份支持修改的参数项如下：
  - 目的端路径
  - 最大备份数
- 周期备份支持修改的参数项如下：
  - 开始时间
  - 周期
  - 目的端路径
  - 最大备份数

  说明

  - 当备份任务的“路径类型”为“LocalHDFS”时，修改备份任务时参数“目的端路径”有效。
  - 修改某个备份任务参数“目的端路径”后，第一次执行此任务默认为全量备份。

步骤4 单击“确定”保存。

----结束

11.10.5 查看备份恢复任务

操作场景

该任务指导用户通过MRS Manager查看已创建的备份恢复任务，以及任务的运行情况。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 单击“备份管理”或“恢复管理”。

步骤3 在任务列表中，查看“当次任务进度”列获取上一次任务运行的结果。绿色表示运行成功，红色表示运行失败。

步骤4 在任务列表指定任务的“操作”列，单击“更多 > 查询历史”，打开备份恢复任务运行记录。

在弹出的窗口中，在指定一次执行记录的“详情”列，单击“查看”，打开此次任务运行的日志信息。

----结束

相关任务

- 修改备份任务

  参考**修改备份任务**。
- 查看恢复任务

  在任务列表指定任务的“操作”列，单击“查询详情”，查看恢复任务。恢复任务的参数只能查看但不能修改。
- 运行备份恢复任务

  在任务列表指定任务的“操作”列，单击“启动”，启动处于准备或失败状态的备份、恢复任务。已成功执行过的恢复任务不能重新运行。
- 停止备份任务

  在任务列表指定任务的“操作”列，单击“更多 > 停止”，停止处于运行状态的备份恢复任务。
- 删除备份恢复任务

  在任务列表指定任务的“操作”列，单击“更多 > 删除”，删除备份恢复任务。删除任务后备份的数据默认会保留。
- 挂起备份任务

  在任务列表指定任务的“操作”列，单击“更多 > 挂起”，挂起备份任务。仅支持周期备份的任务，挂起后周期备份任务不再自动执行。挂起正在执行的备份任务时，该任务会停止运行。需要取消任务的挂起状态时，单击“更多 > 重新执行”。

11.11 安全管理

11.11.1 未开启 Kerberos 认证集群中的默认用户清单

用户分类

MRS集群提供以下2类用户，请用户定期修改密码，不建议使用默认密码。

用户类型	使用说明
系统用户	用于OMS系统进程运行的用户。
数据库用户	• 用于OMS数据库管理和数据访问的用户。 • 用于业务组件（Hive、Loader和DBservice）数据库的用户。

系统用户

说明

- MRS集群需要使用操作系统中ldap用户，此帐号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作用户维护。
- 首次修改“ommdba”和“omm”密码需执行重置密码操作。找回密码后建议定期修改。

类别	用户名称	初始密码	描述
MRS集群系统管理员	admin	在集群创建时由用户指定。	MRS Manager的管理员。 此外还具有以下权限： ● 具有HDFS、ZooKeeper普通用户的权限。 ● 具有提交、查询Mapreduce、YARN任务的权限，以及YARN队列管理权限和访问YARN WebUI的权限。 ● Storm中，具有提交、查询、激活、去激活、重分配、删除拓扑的权限，可以操作所有拓扑。 ● Kafka服务中，具有创建、删除、授权、Reassign、消费、写入、查询主题的权限。
MRS集群节点操作系统用户	omm	系统随机生成	MRS集群系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。
MRS集群节点操作系统用户	root	用户设置的密码。	MRS集群所属节点的登录用户。在全部节点生成，属于操作系统用户。

用户组信息

默认用户组	描述
supergroup	admin用户的主组，在关闭Kerberos认证的集群中没有额外的权限。

默认用户组	描述
check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
Manager_tenant	租户系统用户组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
System_administrator	MRS集群系统管理员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_viewer	MRS Manager系统查看员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_operator	MRS Manager系统操作员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_auditor	MRS Manager系统审计员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_administrator	MRS Manager系统管理员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
compcommon	MRS集群系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
default_1000	为租户创建的用户组。系统内部组，仅限组件间内部使用。
launcher-job	MRS系统内部组，用于使用V2接口提交作业。

操作系统用户组	描述
wheel	MRS集群系统内部运行用户“omm”的主组。
ficommon	MRS集群系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

数据库用户

MRS集群系统数据库用户包含OMS数据库用户、DBService数据库用户。

说明

数据库用户不能删除，否则可能导致集群或组件服务无法正常工作。

类别	默认用户	初始密码	描述
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作
	omm	ChangeMe@123456	OMS数据库数据访问用户
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户
	hive	HiveUser@	Hive连接DBService数据库用户
	hue	HueUser@123	Hue连接DBService数据库用户
	sqoop	SqoopUser@	Loader连接DBService数据库的用户

11.11.2 开启 Kerberos 认证集群中的默认用户清单

用户分类

MRS集群提供以下3类用户，请用户定期修改密码，不建议使用默认密码。

用户类型	使用说明
系统用户	● 通过Manager创建，是MRS集群操作运维与业务场景中主要使用的用户，包含两种类型： – “人机”用户：用于在Manager的操作运维场景，以及在组件客户端操作的场景。 – “机机”用户：用于MRS集群应用开发的场景。 ● 用于OMS系统进程运行的用户。
系统内部用户	MRS集群提供的用于进程通信、保存用户组信息和关联用户权限的内部用户。
数据库用户	● 用于OMS数据库管理和数据访问的用户。 ● 用于业务组件（Hive、Hue、Loader和DBservice）数据库的用户。

系统用户

说明

- MRS集群需要使用操作系统中ldap用户，此帐号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作用户维护。
- 首次修改“ommdba”和“omm”用户需要执行重置密码操作。找回密码后建议定期修改。

类别	用户名称	初始密码	描述
MRS集群系统管理员	admin	在集群创建时由用户指定。	Manager的管理员。 此外还具有以下权限： ● 具有HDFS、ZooKeeper普通用户的权限。 ● 具有提交、查询Mapreduce、YARN任务的权限，以及YARN队列管理权限和访问YARN WebUI的权限。 ● Storm中，具有提交、查询、激活、去激活、重分配、删除拓扑的权限，可以操作所有拓扑。 ● Kafka服务中，具有创建、删除、授权、Reassign、消费、写入、查询主题的权限。
MRS集群节点操作系统用户	omm	系统随机生成	MRS集群系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。
MRS集群节点操作系统用户	root	用户设置的密码。	MRS集群所属节点的登录用户。在全部节点生成，属于操作系统用户。

系统内部用户

说明

以下系统内部用户不能删除，否则可能导致集群或组件无法正常工作。

类别	默认用户	初始密码	描述
组件运行用户	hdfs	Hdfs@123	HDFS系统管理员，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看webUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统
	hbase	Hbase@123	HBase系统管理员，用户权限： ● 集群管理权限: 表的Enable、Disable操作，触发MajorCompact，ACL操作 ● 授权或回收权限，集群关闭等操作相关的权限 ● 表管理权限: 建表、修改表、删除表等操作权限 ● 数据管理权限：表级别、列族级别以及列级别的数据读写权限 ● 访问HBase WebUI的权限
	mapred	Mapred@123	MapReduce系统管理员，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 访问Yarn、MapReduce WebUI的权限
	spark	Spark@123	Spark系统管理员，用户权限： ● 访问Spark WebUI的权限 ● 提交Spark任务的权限

用户组信息

默认用户组	描述
hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。
hbase	普通用户组，将用户加入此用户组不会获得额外的权限。
hive	将用户加入此用户组，可以使用Hive。
spark	普通用户组，将用户加入此用户组不会获得额外的权限。
supergroup	将用户加入此用户组，可获得HBase、HDFS和Yarn的管理员权限，并可以使用Hive。
check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
Manager_tenant	租户系统用户组。系统内部组，仅限组件间内部使用。
System_administrator	MRS集群系统管理员组。系统内部组，仅限组件间内部使用。
Manager_viewer	MRS Manager系统查看员组。系统内部组，仅限组件间内部使用。
Manager_operator	MRS Manager系统操作员组。系统内部组，仅限组件间内部使用。
Manager_auditor	MRS Manager系统审计员组。系统内部组，仅限组件间内部使用。
Manager_administrator	MRS Manager系统管理员组。系统内部组，仅限组件间内部使用。
compcommon	MRS系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
default_1000	为租户创建的用户组。系统内部组，仅限组件间内部使用。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限,才能访问对应Topic。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。

默认用户组	描述
stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。
opentsdb	普通用户组，将用户加入此用户组不会获得额外的权限。
presto	普通用户组，将用户加入此用户组不会获得额外的权限。
flume	普通用户组，添加到该用户组的用户无任何额外权限。
launcher-job	MRS系统内部组，用于使用V2接口提交作业。

操作系统用户组	描述
wheel	MRS集群系统内部运行用户“omm”的主组。
ficommon	MRS集群系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

数据库用户

MRS集群系统数据库用户包含OMS数据库用户、DBService数据库用户。

说明

数据库用户不能删除，否则可能导致集群或组件服务无法正常工作。

类别	默认用户	初始密码	描述
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作
	omm	ChangeMe@123456	OMS数据库数据访问用户
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户
	hive	HiveUser@	Hive连接DBService数据库用户
	hue	HueUser@123	Hue连接DBService数据库用户
	sqoop	SqoopUser@	Loader连接DBService数据库的用户

类别	默认用户	初始密码	描述
	ranger	RangerUser@	Ranger连接DBService数据库的用户

11.11.3 修改操作系统用户密码

操作场景

该任务指导用户定期修改MRS集群节点操作系统用户“omm”、“ommdba”、"root"的登录密码，以提升系统运维安全性。

各节点“omm”、“ommdba”、"root"无需设置为统一的密码。

操作步骤

步骤1 登录Master1节点，然后登录要修改操作系统用户密码的其他节点。

步骤2 执行以下命令切换到用户。

sudo su - root

步骤3 执行如下命令，修改omm/ommdba/root用户密码。

passwd omm

passwd ommdba

passwd root

例如omm：passwd，系统显示：

```
Changing password for user omm.
New password:
```

输入用户的新密码。操作系统的密码修改策略由用户实际使用的操作系统类型决定。

```
Retype new password:
passwd: all authentication tokens updated successfully.
```

说明

MRS集群默认密码复杂度要求：

- 密码字符长度至少8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 重置的密码不能是最近5次使用过的密码。

----结束

11.11.4 修改 admin 密码

该任务指导用户定期修改集群用户“admin”的密码，以提升系统运维安全性。

修改该密码会导致已经下载的用户凭证不可用，请修改该密码后重新下载认证凭据并替换旧凭据。

在集群节点修改 admin 密码

步骤1 更新主管理节点客户端，具体请参看**更新客户端（3.x之前版本）**。

步骤2 登录主管理节点。

步骤3 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤4 执行以下命令切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤5 执行以下命令配置环境变量。

source bigdata_env

步骤6 执行以下命令，修改“admin”密码。此操作在整个集群中生效。

kpasswd admin

先输入旧密码，再输入两次新密码。

集群中，默认的密码复杂度要求：

- 密码字符长度至少8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

在 MRS Manager 页面修改 admin 密码

开启Kerberos认证的集群和开启弹性公网IP功能未开启Kerberos认证的集群支持通过MRS Manager界面修改admin密码。

步骤1 用admin帐户登录MRS Manager页面。

步骤2 单击页面右上角用户名，选择“修改密码”。

步骤3 在修改密码页面，输入“旧密码”、“新密码”、“确认新密码”。

说明

默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

步骤4 单击“确定”完成密码修改，使用新密码重新登录MRS Manager页面。

----结束

重置 admin 密码

步骤1 登录Master1节点。

步骤2 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如重置admin密码：**cpw admin**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

11.11.5 修改 Kerberos 管理员密码

操作场景

该任务指导用户定期修改MRS集群Kerberos管理员“kadmin”的密码，以提升系统运维安全性。

修改该密码会导致已经下载的用户凭证不可用，请修改该密码后重新下载认证凭据并替换旧凭据。

前提条件

已在Master1节点准备客户端。

操作步骤

步骤1 登录Master1节点。

步骤2 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，修改**kadmin/admin**密码。此操作对所有服务器生效。

kpasswd kadmin/admin

集群中，默认的密码复杂度要求：

- 密码字符长度至少8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

11.11.6 修改 LDAP 管理员和 LDAP 用户密码

操作场景

该任务指导用户定期修改MRS集群的LDAP管理员用户“rootdn:cn=root,dc=hadoop,dc=com”和LDAP用户“pg_search_dn:cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启全部服务，服务在重启时无法访问。

操作步骤

步骤1 在MRS Manager，选择“服务管理 > LdapServer > 更多”。

步骤2 单击“修改密码”。

步骤3 在“修改密码”对话框的“用户信息”选择要修改的用户。

步骤4 在“修改密码”对话框的“旧密码”输入旧密码，“新密码”和“确认密码”输入新密码。

默认密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符`~!@#$%^&*()-_=+\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序用户名相同。
- 不可与当前密码相同。

说明

LDAP管理员用户“rootdn:cn=root,dc=hadoop,dc=com”的默认密码为“LdapChangeMe@123”，LDAP用户“pg_search_dn:cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的默认密码为“pg_search_dn@123”，请定期修改密码并妥善保存。

步骤5 勾选“我已阅读此信息并了解其影响。”，单击“确定”确认修改并重启服务。

----结束

11.11.7 修改组件运行用户密码

操作场景

该任务指导用户定期修改MRS集群组件运行用户的密码，以提升系统运维安全性。

如果初始密码由系统随机生成，需要直接重置密码。

修改该密码会导致已经下载的用户凭证不可用，请修改该密码后重新下载认证凭据并替换旧凭据。

前提条件

已在Master1节点准备客户端。

操作步骤

步骤1 登录Master1节点。

步骤2 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如重置admin密码：**cpw admin**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。

- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

11.11.8 修改 OMS 数据库管理员密码

操作场景

该任务指导用户定期修改OMS数据库管理员的密码，以提升系统运维安全性。

操作步骤

步骤1 登录主管理节点。

说明

ommdba用户密码不支持在备管理节点修改，否则集群无法正常工作。只需在主管理节点执行修改操作，无需在备管理节点操作。

步骤2 执行以下命令，切换用户。

sudo su - omm

步骤3 执行以下命令，切换目录。

cd $OMS_RUN_PATH/tools

步骤4 执行以下命令，修改ommdba用户密码。

mod_db_passwd ommdba

步骤5 输入**ommdba**的原密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序用户名相同。
- 不可与前20个历史密码相同。

显示如下结果，说明修改成功：

```
Congratulations, update [ommdba] password successfully.
```

----结束

11.11.9 修改 OMS 数据库数据访问用户密码

操作场景

该任务指导用户定期修改OMS数据库访问用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启OMS服务，服务在重启时无法访问。

操作步骤

步骤1 在MRS Manager单击“系统设置”。

步骤2 在“权限配置”区域下，单击“OMS数据库密码修改”。

步骤3 在omm用户所在行，单击“操作”列下的“修改密码”，修改OMS数据库密码。

密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。
- 不可与前20个历史密码相同。

步骤4 单击“确定”，等待界面提示“操作成功”后单击“完成”。

步骤5 在omm用户所在行，单击“操作”列下的“重启OMS服务”，重启OMS数据库。

说明

如果修改了密码但未重启OMS数据库，则omm用户的状态变为“Waiting to restart”且无法再修改密码，直到重启OMS数据库

步骤6 在弹出的对话框中，勾选“我已阅读此信息并了解其影响。”，单击“确定”，重新启动OMS服务。

----结束

11.11.10 修改组件数据库用户密码

操作场景

该任务指导用户定期修改组件数据库用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启服务，服务在重启时无法访问。

操作步骤

步骤1 在MRS Manager单击“服务管理”，单击待修改数据库用户服务的名称。

步骤2 确定修改哪个组件数据库用户密码。

- 修改DBService数据库用户密码，直接执行**步骤3**。
- 修改Hive或者Hue或者Loader数据库用户密码，需要先停止服务再执行**步骤3**。

单击“停止服务”。

步骤3 选择“更多 > 修改密码”。

步骤4 根据界面信息，输入新旧密码。

密码复杂度要求：

- DBService数据库用户密码字符长度为16～32位。Hive或Hue或Loader数据库用户密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序用户名相同。
- 不可与前20个历史密码相同。

步骤5 单击“确定”，系统自动重新启动对应的服务。界面提示“操作成功”，单击“完成”。

----结束

11.11.11 更新集群密钥

操作场景

在创建集群时，系统将自动生成加密密钥key值以对集群的部分安全信息（例如所有数据库用户密码、密钥文件访问密码等）进行加密存储。在集群安装成功后，建议用户定期通过以下操作手动更改密钥值。

对系统的影响

- 更新集群密钥后，集群中新增加一个随机生成的新密钥，用于加密解密新保存的数据。旧的密钥不会删除，用于解密旧的加密数据。在修改安全信息后，例如修改数据库用户密码，新密码将使用新的密钥加密。
- 更新集群密钥需要停止集群，集群停止时无法访问。

前提条件

停止依赖集群运行的上层业务应用。

操作步骤

步骤1 在MRS Manager，选择“服务管理 > 更多 > 停止集群”。

在弹出窗口勾选“我已阅读此信息并了解影响。”，单击“确定”，界面提示“操作成功。”，单击“完成”，集群成功停止。

步骤2 登录主管理节点。

步骤3 执行以下命令切换用户：

sudo su - omm

步骤4 执行以下命令，防止超时退出。

TMOUT=0

步骤5 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-0.0.1/tools

步骤6 执行以下命令，更新集群密钥。

sh updateRootKey.sh

根据界面提示，输入**y**：

```
The root key update is a critical operation.
Do you want to continue?(y/n):
```

界面提示以下信息表示更新密钥成功：

```
...
Step 4-1: The key save path is obtained successfully.
...
Step 4-4: The root key is sent successfully.
```

步骤7 在MRS Manager界面，选择“服务管理 > 更多 > 启动集群”。

在弹出的提示框中单击“是”，开始启动集群。界面提示“操作成功。”，单击“完成”，集群成功启动。

----结束

11.12 权限管理

11.12.1 创建角色

操作场景

该任务指导管理员用户在MRS Manager创建角色，并对Manager和组件进行授权管理。

MRS Manager支持的角色数为1000。

前提条件

管理员用户已明确业务需求。

操作步骤

步骤1 在MRS Manager，选择“系统设置 > 角色管理”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

“角色名称”为必选参数，字符长度为3到30，可以包含数字、字母和下划线。“描述”为可选参数。

步骤3 设置角色“权限”。

1. 单击“服务名称”，并选择一个“视图名称”。
2. 勾选一个或多个权限。

说明

- “权限”为可选参数。
- 在选择“视图”设置组件的权限时，可通过右上角的“搜索”框输入资源名称，然后单击 显示搜索结果。
- 搜索范围仅包含当前权限目录，无法搜索子目录。搜索关键字支持模糊搜索，不区分大小写。支持搜索下一页的结果。

表 11-36 Manager 权限描述

支持权限管理的资源	权限设置说明
“Alarm”	Manager告警功能授权，勾选“View”表示可以查看告警，勾选“Management”表示可以管理告警。
“Audit”	Manager审计日志功能授权，勾选“View”表示可以查看审计，勾选“Management”表示可以管理审计。
“Dashboard”	Manager概览功能授权，勾选“View”表示可以查看集群概览。
“Hosts”	Manager集群节点管理功能授权，勾选“View”表示可以查看节点，勾选“Management”表示可以管理节点。
“Services”	MRS集群服务管理功能授权，勾选“View”表示可以查看服务，勾选“Management”表示可以管理服务。
“System_cluster_management”	MRS集群管理授权，勾选“Management”表示可以使用MRS补丁管理功能。
“System_configuration”	MRS集群配置功能授权，勾选“Management”表示可以使用Manager配置MRS集群。
“System_task”	MRS集群任务功能授权，勾选“Management”表示可以使用Manager管理MRS集群的周期任务。
“Tenant”	Manager多租户管理功能授权，勾选“Management”表示可以查看Manager的租户管理页面。

表 11-37 HBase 权限描述

支持权限管理的资源	权限设置说明
“SUPER_USER_GROUP”	选中时表示授予HBase管理员权限。
“Global”	HBase的一种资源类型，表示HBase整体组件。

支持权限管理的资源	权限设置说明
“Namespace”	HBase的一种资源类型，表示命名空间，用来保存HBase表。具体权限： ● “Admin”：表示管理此命名空间的权限。 ● “Create”：表示在此命名空间创建HBase表的权限。 ● “Read”：表示访问此命名空间的权限。 ● “Write”：表示写入此命名空间数据的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“Table”	HBase的一种资源类型，表示数据表，用来保存数据。具体权限： ● “Admin”：表示管理此数据表的权限。 ● “Create”：表示在此数据表创建列族和列的权限。 ● “Read”：表示读取数据表的权限。 ● “Write”：表示写入数据到表的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“ColumnFamily”	HBase的一种资源类型，表示列族，用来保存数据。具体权限： ● “Create”：表示在此列族创建列的权限。 ● “Read”：表示读取列族的权限。 ● “Write”：表示写入数据到列族的权限。
“Qualifier”	HBase的一种资源类型，表示列，用来保存数据。具体权限： ● “Read”：表示读取列的权限。 ● “Write”：表示写入数据到列的权限。

HBase中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如命名空间“default”添加了“Read”和“Write”权限，则命名空间中的表、列族和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 11-38 HDFS 权限描述

支持权限管理的资源	权限设置说明
“Folder”	HDFS的一种资源类型，表示HDFS目录，可以保存文件或子目录。具体权限： ● “Read”：表示访问此HDFS目录的权限。 ● “Write”：表示在此HDFS目录写入数据的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。
“Files”	HDFS的一种资源类型，表示HDFS中的文件。具体权限： ● “Read”：表示访问此文件的权限。 ● “Write”：表示写入此文件的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。

HDFS中每一级目录的权限默认不会传递到下级目录类型。例如目录“tmp”添加了“Read”和“Execute”，需要同时勾选“递归”才能为子目录添加权限。

表 11-39 Hive 权限描述

支持权限管理的资源	权限设置说明
“Hive Admin Privilege”	选中时表示授予Hive管理员权限。
“Database”	Hive的一种资源类型，表示Hive数据库，用来保存Hive表。具体权限： ● “Select”：表示查询Hive数据库的权限。 ● “Delete”：表示在Hive数据库执行删除操作的权限。 ● “Insert”：表示在Hive数据库执行插入操作的权限。 ● “Create”：表示在Hive数据库执行创建操作的权限。

支持权限管理的资源	权限设置说明
“Table”	Hive的一种资源类型，表示Hive表，用来保存数据。具体权限： ● “Select”：表示查询Hive表的权限。 ● “Delete”：表示在Hive表执行删除操作的权限。 ● “Update”：表示为角色添加Hive表的“Update”权限。 ● “Insert”：表示在Hive表执行插入操作的权限。 ● “Grant of Select”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Select”权限。 ● “Grant of Delete”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Delete”权限。 ● “Grant of Update”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Update”权限。 ● “Grant of Insert”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Insert”权限。

Hive中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如数据库“default”添加了“Select”和“Insert”权限，则数据库中的表和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 11-40 YARN 权限描述

支持权限管理的资源	权限设置说明
“Cluster Admin Operations”	选中时表示授予YARN管理员权限。
“root”	YARN的根队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Parent Queue”	YARN的一种资源类型，表示父队列，可以包含子队列。根队列也属于父队列的一种。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Leaf Queue”	YARN的一种资源类型，表示叶子队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。

YARN中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如队列“root”添加了“Submit”权限，则子队列自动添加该权限。子队列继承的权限不在“权限”表格显示被选中。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 11-41 Hue 权限描述

支持权限管理的资源	权限设置说明
“Storage Policy Admin”	选中时表示授予Hue中存储策略管理员权限。

步骤4 单击“确定”完成，返回“角色管理”。

----结束

相关任务

修改角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要修改角色所在的行，单击“修改”，修改角色信息。

说明

修改角色分配的权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要删除角色所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

11.12.2 创建用户组

操作场景

该任务指导管理员用户通过MRS Manager创建新用户组并指定其操作权限，使用户组可以统一管理加入用户组的单个或多个用户。用户加入用户组后，可获得用户组具有的操作权限。

MRS Manager支持用户组数为100。

前提条件

管理员用户已明确业务需求，并已创建业务场景需要的角色。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在组列表上方，单击“添加用户组”。

步骤4 填写“组名”和“描述”。

“组名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。“描述”为可选参数。

步骤5 在“角色”，单击“选择添加角色”选择指定的角色并添加。

如果不添加角色，则当前创建的用户组没有使用MRS集群的权限。

步骤6 单击“确定”完成用户组创建。

----结束

相关任务

修改用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要修改用户组所在的行，单击“修改”，修改用户组信息。

说明

为用户组修改分配的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要删除用户组所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

11.12.3 创建用户

操作场景

该任务指导管理员根据实际业务场景需要，通过MRS Manager创建新用户并指定其操作权限以满足业务使用。

MRS Manager支持的用户数为1000。

如需对新创建用户的密码使用新的密码策略，请先**修改密码策略**，再参考本章节创建用户。

前提条件

管理员已明确业务需求，并已创建业务场景需要的角色和用户组。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在用户列表上方，单击“添加用户”。

步骤4 根据界面提示配置参数，填写“用户名”。

说明

- 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。
- 使用已创建的用户时，请输入和用户名完全一样的大小写字符。
- “用户名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。
- “root”、“omm”和“ommdba”为系统保留用户，请选择其他用户名。

步骤5 设置“用户类型”，可选值包括“人机”和“机机”。

- “人机”用户：用于在MRS Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。
- “机机”用户：用于MRS应用开发的场景。选择该值用户密码随机生成，无需填写。

步骤6 在“用户组”，单击“选择添加的用户组”，选择对应用户组将用户添加进去。

说明

- 如果用户组添加了角色，则用户可获得对应角色中的权限。
- 为新用户分配Hive的权限，请将用户加入hive组。
- 如果用户需要管理租户资源，用户组必须分配了Manager_tenant角色以及租户对应的角色。

步骤7 在“主组”选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。

步骤8 根据业务实际需要在“分配角色权限”，单击“选择并绑定角色”为用户添加角色。

说明

- 创建用户时，如果用户从用户组获得的权限还不满足业务需要，则可以再分配其他已创建的角色。为新用户分配角色授权，最长可能需要3分钟时间生效。
- 创建用户时添加角色可细化用户的权限。
- 没有为新用户分配角色时，此用户可以访问HDFS、HBase、Yarn、Spark和Hue的WebUI。

步骤9 根据业务实际需要“描述”。

“描述”为可选参数。

步骤10 单击“确定”完成用户创建。

第一次在MRS集群中使用新创建的用户，例如登录Manager或者使用集群客户端，需要修改密码，具体请参见《修改操作用户密码》。

----结束

11.12.4 修改用户信息

操作场景

该任务指导管理员用户在MRS Manager修改已创建的用户信息，包括修改用户组、主组、角色和描述。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要修改用户所在的行，单击“修改”，修改用户信息。

说明

为用户修改用户组或分配的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

11.12.5 锁定用户

该任务指导管理员用户将MRS集群中的用户锁定。用户被锁定后，不能在MRS Manager重新登录或在集群中重新进行安全认证。

可通过以下两种方式锁定用户，锁定后的用户需要管理员手动解锁或者等待锁定时间结束才能恢复使用：

- 自动锁定：通过设置密码策略中的“允许输入错误次数”，将超过登录失败次数的用户自动锁定。具体操作请参见**修改密码策略**。
- 手动锁定：由管理员通过手动的方式将用户锁定。

以下将具体介绍手动锁定。不支持锁定“机机”用户。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要锁定用户所在行，单击“锁定用户”，锁定用户。

步骤4 在弹出的提示窗口，单击“确定”完成锁定操作。

----结束

11.12.6 解锁用户

在用户输入错误密码次数大于允许输入错误次数，造成用户被锁定或者用户被管理员手动锁定后需要解锁用户的场景下，管理员用户可以通过MRS Manager为锁定的用户解锁。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要解锁用户所在行，选择“解锁用户”，解锁用户。

步骤4 在弹出的提示窗口，单击“确定”完成解锁操作。

----结束

11.12.7 删除用户

MRS集群用户不再需要使用时，管理员可以在MRS Manager中删除此用户。

说明

如果删除用户A后，再次准备重新创建同名用户A，如果该用户A已经提交过作业（客户端提交或者MRS console页面提交），那么需要在删除该用户A的同时，删除该用户A残留的文件夹，否则使用重新创建的同名用户A提交作业会失败。

删除用户残留文件夹操作方法为：依次登录MRS集群的Core节点，在每个Core节点上执行如下两条命令，其中如下命令中“$user”为具体的以用户名命名的文件夹。

cd /srv/BigData/hadoop/data1/nm/localdir/usercache/

rm -rf $user

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要删除用户所在的行，选择“更多 > 删除”。

步骤4 单击“确定”完成删除操作。

----结束

11.12.8 修改操作用户密码

操作场景

出于MRS集群安全的考虑，“人机”类型系统用户密码必须定期修改。该任务指导用户通过MRS Manager完成修改自身密码工作。

如需对用户修改的密码使用新的密码策略，请先**修改密码策略**，再参考本章节修改密码。

对系统的影响

修改MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

前提条件

- 从管理员获取当前的密码策略。
- 从管理员获取MRS Manager访问地址。

操作步骤

步骤1 在MRS Manager，移动鼠标到界面右上角的  。

在弹出菜单，选择“修改密码”。

步骤2 分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

11.12.9 初始化系统用户密码

操作场景

该任务指导管理员在用户遗忘密码或公共帐号密码需要定期修改时，通过MRS Manager初始化密码。初始化密码后用户首次使用需要修改密码。

对系统的影响

初始化MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

初始化“人机”用户密码

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要初始化密码用户所在行，单击“更多 > 初始化密码”，按界面提示信息修改用户密码。

在弹出窗口中输入当前登录的管理员密码确认管理员身份，单击“确定”，然后在“初始化密码”单击“确定”。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。

- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

初始化“机机”用户密码

步骤1 根据业务情况，准备好客户端，并登录安装客户端的节点。

步骤2 执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如：**cpw oms/manager**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

11.12.10 下载用户认证文件

操作场景

用户开发大数据应用程序并在支持Kerberos认证的MRS集群中运行程序时，需要准备访问MRS集群的用户认证文件。认证文件中的keytab文件可用于认证用户身份。

该任务指导管理员用户通过MRS Manager下载用户认证文件并导出keytab文件。

说明

- 如果选择下载“人机”用户的认证文件，在下载前需要使用Manager修改过一次此用户的密码使管理员设置的初始密码失效，否则导出的keytab文件无法使用。请参见**修改操作用户密码**。
- 修改用户密码后，之前导出的keytab将失效，需要重新导出。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在需导出keytab文件用户所在的行，选择“更多 > 下载认证凭据”下载认证文件，待文件自动生成后指定保存位置，并妥善保管该文件。

步骤4 使用解压程序打开认证文件。

- “user.keytab”表示用户keytab文件，用于认证用户身份。
- “krb5.conf”表示认证服务器配置文件，应用程序在进行用户认证身份时根据该文件的配置信息连接认证服务器。

----结束

11.12.11 修改密码策略

操作场景

该任务指导管理员用户设置密码安全规则、用户登录安全规则及用户锁定规则。由于“机机”用户密码随机生成，在MRS Manager设置密码策略只影响“人机”用户。

如需对新增用户的密码或用户修改的密码使用新的密码策略，请先参考本章节修改密码策略，再**创建用户**或**修改密码**。

须知

密码策略涉及用户管理的安全性，请根据业务安全要求谨慎修改，否则会有安全性风险。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 单击“密码策略配置”。

步骤3 根据界面提示，修改密码策略，具体参数见下表。

表 11-42 密码策略参数说明

参数名称	描述
最小密码长度	密码包含的最小字符个数，取值范围是8～32。默认值为“8”。

参数名称	描述
字符类型的数目	密码字符包含大写字母、小写字母、数字、空格和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=）的最小种类。可选择数值为“3”和“4”。默认值“3”表示至少必须使用大写字母、小写字母、数字、特殊符号和空格中的任意3种。
密码有效期（天）	密码有效使用天数，取值范围0～90，0表示永久有效。默认值为“90”。
密码失效提前提醒天数	提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录MRS Manager时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒。默认值为“5”。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔（分钟），取值范围为0～1440。“0”表示永远有效，“1440”表示1天。默认值为“5”。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。默认值为“5”。
用户锁屏时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。默认值为“5”。

----结束

11.13 MRS 多用户权限管理

11.13.1 MRS 集群中的用户与权限

概述

- **MRS集群用户**

  Manager中的安全帐号，包含用户名、密码等属性，MRS集群的使用者通过这类用户访问集群中的资源。每个启用Kerberos认证的MRS集群可以有多个用户。

- **MRS集群角色**

  用户在实际使用MRS集群时需根据业务场景获取访问资源的权限，访问资源的权限是定义到MRS集群对象上的。集群的角色就是包含一个或者多个权限的集合。例如，HDFS中某个目录的访问权限，需要在指定的目录上配置，并保存在角色中。

Manager支持MRS集群用户权限管理功能，使权限管理与用户管理更加直观、易用。

- 权限管理：使用RBAC（Role-Based Access Control）方式，即基于角色授予权限，形成权限的集合。用户通过分配一个或多个已授权的角色取得对应的权限。

- 用户管理：使用Manager统一管理MRS集群用户，并通过Kerberos协议认证用户，LDAP协议存储用户信息。

权限管理

MRS集群提供的权限包括Manager和各组件（例如HDFS、HBase、Hive和Yarn等）的操作维护权限，在实际应用时需根据业务场景为各用户分别配置不同权限。为了提升权限管理的易用性，Manager引入角色的功能，通过选取指定的权限并统一授予角色，以权限集合的形式实现了权限集中查看和管理，提升了权限管理的易用性和用户体验。

角色可以理解为集中一个或多个权限的逻辑体，角色被授予指定的权限，用户通过绑定角色获得对应的权限。

一个角色可以有多个权限，一个用户可以绑定多个角色。

- 角色1：授予操作权限A和B，用户a和用户b通过绑定角色1取得对应的权限。
- 角色2：授予操作权限C，用户c和用户d通过绑定角色2取得对应的权限。
- 角色3：授予操作权限D和F，用户a通过绑定配角色3取得对应的权限。

例如，MRS集群用户绑定了管理员角色，那么这个用户成为MRS集群的管理员用户。

Manager界面显示系统默认创建的角色如**表11-43**所示。

表 11-43 Manager 默认角色与描述

默认角色	角色描述
default	为租户创建的角色。
Manager_administrator	Manager管理员，具有Manager的管理权限。
Manager_auditor	Manager审计员，具有查看和管理审计信息的权限。
Manager_operator	Manager操作员，具有除租户管理、配置管理和集群管理功能以外的权限。
Manager_viewer	Manager查看员，具有查看系统概览，服务，主机，告警，审计日志等信息的权限。
System_administrator	系统管理员，具有Manager的管理权限及所有服务管理员的所有权限。
Manager_tenant	Manager租户管理页面查看角色，具有Manager“租户管理”页面查看权限。

通过Manager创建角色时支持对Manager和组件进行授权管理，如**表11-44**所示。

表 11-44 Manager 与组件授权管理

授权类型	授权描述
Manager	Manager访问与登录权限。

授权类型	授权描述
HBase	HBase管理员权限设置和表、列族授权。
HDFS	HDFS中的目录和文件授权。
Hive	• Hive Admin Privilege Hive管理员权限。 • Hive Read Write Privileges Hive数据表管理权限，可设置与管理已创建的表的数据操作权限。
Hue	存储策略管理员权限。
Yarn	• Cluster Admin Operations Yarn管理员权限。 • Scheduler Queue 队列资源管理。

用户管理

支持Kerberos认证的MRS集群使用Kerberos协议和LDAP（Lightweight Directory Access Protocol）协议来配合工作，实现用户管理：

- Kerberos用于在用户登录Manager与使用组件客户端时认证用户身份，未启用Kerberos认证的集群则不认证用户身份。
- LDAP用于存储用户记录、用户组信息与权限信息等用户信息。

MRS集群支持在Manager执行创建用户或者修改用户等任务时，系统自动完成更新Kerberos和LDAP的用户数据，用户登录Manager或使用组件客户端时，系统自动完成认证用户身份和获取用户信息。这样一方面保证了用户管理的安全性，另一方面简化了用户管理的操作任务。Manager还提供了用户组功能，可对单个或多个用户进行分类管理：

- 用户组为一批用户的集合，可对用户进行分类管理。系统中的用户可以单独存在也可以加入到某个用户组中。
- 对已分配角色的用户组来说，当用户添加到该组时，用户组分配的角色权限将授权给用户。

MRS 3.x之前版本集群MRS Manager界面显示系统默认创建的用户组如**表11-45**所示。

MRS 3.x及之后版本集群FusionInsight Manager界面显示系统默认创建的用户组请参考**用户组**。

表 11-45 Manager 默认用户组与描述

用户组名称	描述
hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。

用户组名称	描述
hbase	普通用户组，将用户加入此用户组不会获得额外的权限。
hive	将用户加入此用户组，可以使用Hive。
spark	普通用户组，将用户加入此用户组不会获得额外的权限。
supergroup	将用户加入此用户组，可获得HBase、HDFS和Yarn的管理员权限，并可以使用Hive。
flume	普通用户组。添加到该用户组的用户无任何额外权限。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限,才能访问对应Topic。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。
stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。

启用Kerberos认证的MRS集群默认创建“admin”用户帐号，用于集群管理员维护集群。

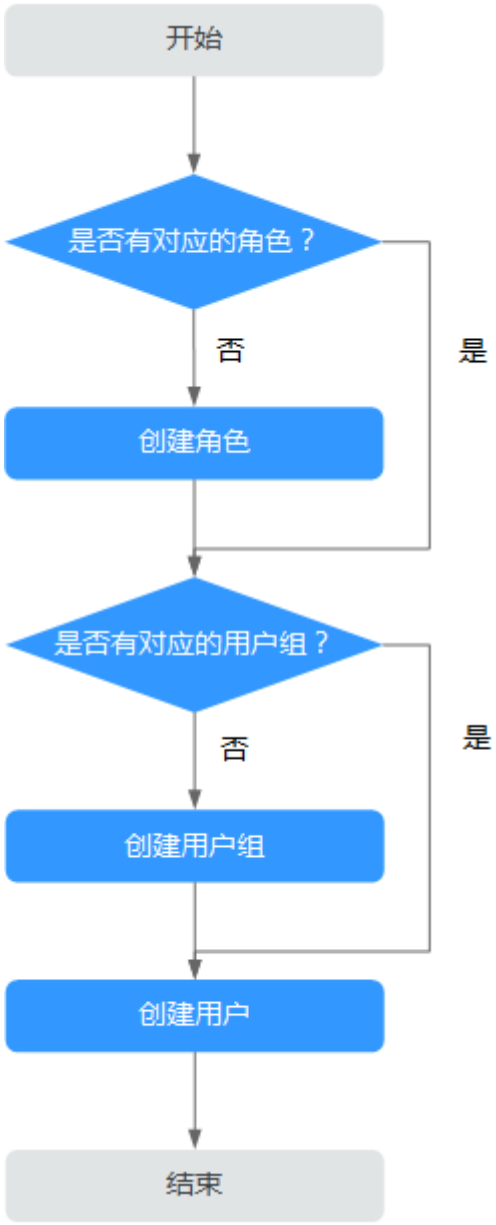
流程概述

在实际业务中，MRS集群用户需要先明确大数据的业务场景，规划集群用户对应的权限。然后在Manager界面创建角色，并设置角色包含的权限以匹配业务的需求。如果需要统一管理单个或多个相同业务场景中的用户，Manager提供了用户组的功能，管理员可以创建用户组。

说明

如果角色设置HDFS、HBase、Hive或Yarn各组件的权限，仅可以使用组件自身功能。如果还需要使用Manager，请在角色中添加对应的Manager权限。

图 11-1 创建用户流程示意



11.13.2 开启 Kerberos 认证集群中的默认用户清单

用户分类

MRS集群提供以下3类用户，请用户定期修改密码，不建议使用默认密码。

用户类型	使用说明
系统用户	● 通过Manager创建，是MRS集群操作运维与业务场景中主要使用的用户，包含两种类型： – “人机”用户：用于在Manager的操作运维场景，以及在组件客户端操作的场景。 – “机机”用户：用于MRS集群应用开发的场景。 ● 用于OMS系统进程运行的用户。
系统内部用户	MRS集群提供的用于进程通信、保存用户组信息和关联用户权限的内部用户。
数据库用户	● 用于OMS数据库管理和数据访问的用户。 ● 用于业务组件（Hive、Hue、Loader和DBservice）数据库的用户。

系统用户

说明

- MRS集群需要使用操作系统中ldap用户，此帐号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作用户维护。
- 首次修改“ommdba”和“omm”用户需要执行重置密码操作。找回密码后建议定期修改。

类别	用户名称	初始密码	描述
MRS集群系统管理员	admin	在集群创建时由用户指定。	Manager的管理员。 此外还具有以下权限： ● 具有HDFS、ZooKeeper普通用户的权限。 ● 具有提交、查询Mapreduce、YARN任务的权限，以及YARN队列管理权限和访问YARN WebUI的权限。 ● Storm中，具有提交、查询、激活、去激活、重分配、删除拓扑的权限，可以操作所有拓扑。 ● Kafka服务中，具有创建、删除、授权、Reassign、消费、写入、查询主题的权限。
MRS集群节点操作系统用户	omm	系统随机生成	MRS集群系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。

类别	用户名称	初始密码	描述
MRS集群节点操作系统用户	root	用户设置的密码。	MRS集群所属节点的登录用户。在全部节点生成，属于操作系统用户。

系统内部用户

说明

以下系统内部用户不能删除，否则可能导致集群或组件无法正常工作。

类别	默认用户	初始密码	描述
组件运行用户	hdfs	Hdfs@123	HDFS系统管理员，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看webUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统
	hbase	Hbase@123	HBase系统管理员，用户权限： ● 集群管理权限: 表的Enable、Disable操作，触发MajorCompact，ACL操作 ● 授权或回收权限，集群关闭等操作相关的权限 ● 表管理权限: 建表、修改表、删除表等操作权限 ● 数据管理权限：表级别、列族级别以及列级别的数据读写权限 ● 访问HBase WebUI的权限

类别	默认用户	初始密码	描述
	mapred	Mapred@123	MapReduce系统管理员，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 访问Yarn、MapReduce WebUI的权限
	spark	Spark@123	Spark系统管理员，用户权限： ● 访问Spark WebUI的权限 ● 提交Spark任务的权限

用户组信息

默认用户组	描述
hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。
hbase	普通用户组，将用户加入此用户组不会获得额外的权限。
hive	将用户加入此用户组，可以使用Hive。
spark	普通用户组，将用户加入此用户组不会获得额外的权限。
supergroup	将用户加入此用户组，可获得HBase、HDFS和Yarn的管理员权限，并可以使用Hive。
check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
Manager_tenant	租户系统用户组。系统内部组，仅限组件间内部使用。
System_administrator	MRS集群系统管理员组。系统内部组，仅限组件间内部使用。
Manager_viewer	MRS Manager系统查看员组。系统内部组，仅限组件间内部使用。
Manager_operator	MRS Manager系统操作员组。系统内部组，仅限组件间内部使用。
Manager_auditor	MRS Manager系统审计员组。系统内部组，仅限组件间内部使用。

默认用户组	描述
Manager_administrator	MRS Manager系统管理员组。系统内部组，仅限组件间内部使用。
compcommon	MRS系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
default_1000	为租户创建的用户组。系统内部组，仅限组件间内部使用。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限,才能访问对应Topic。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。
stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。
opentsdb	普通用户组，将用户加入此用户组不会获得额外的权限。
presto	普通用户组，将用户加入此用户组不会获得额外的权限。
flume	普通用户组，添加到该用户组的用户无任何额外权限。
launcher-job	MRS系统内部组，用于使用V2接口提交作业。

操作系统用户组	描述
wheel	MRS集群系统内部运行用户“omm”的主组。
ficommon	MRS集群系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

数据库用户

MRS集群系统数据库用户包含OMS数据库用户、DBService数据库用户。

说明

数据库用户不能删除，否则可能导致集群或组件服务无法正常工作。

类别	默认用户	初始密码	描述
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作
	omm	ChangeMe@123456	OMS数据库数据访问用户
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户
	hive	HiveUser@	Hive连接DBService数据库用户
	hue	HueUser@123	Hue连接DBService数据库用户
	sqoop	SqoopUser@	Loader连接DBService数据库的用户
	ranger	RangerUser@	Ranger连接DBService数据库的用户

11.13.3 创建角色

操作场景

该任务指导管理员用户在Manager创建角色，并对Manager和组件进行授权管理。

Manager支持的角色数为1000。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。
MRS 3.x及之后版本集群请参考**角色管理**章节。

前提条件

- 管理员用户已明确业务需求。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，选择“系统设置 > 角色管理”。

步骤3 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

“角色名称”为必选参数，字符长度为3到30，可以包含数字、字母和下划线。“描述”为可选参数。

步骤4 设置角色“权限”。

1. 单击“服务名称”，并选择一个“视图名称”。
2. 勾选一个或多个权限。

说明

- “权限”为可选参数。
- 在选择“视图”设置组件的权限时，可通过右上角的“搜索”框输入资源名称，然后单击  显示搜索结果。
- 搜索范围仅包含当前权限目录，无法搜索子目录。搜索关键字支持模糊搜索，不区分大小写。支持搜索下一页的结果。

表 11-46 Manager 权限描述

支持权限管理的资源	权限设置说明
“Alarm”	Manager告警功能授权，勾选“View”表示可以查看告警，勾选“Management”表示可以管理告警。
“Audit”	Manager审计日志功能授权，勾选“View”表示可以查看审计，勾选“Management”表示可以管理审计。
“Dashboard”	Manager概览功能授权，勾选“View”表示可以查看集群概览。
“Hosts”	Manager集群节点管理功能授权，勾选“View”表示可以查看节点，勾选“Management”表示可以管理节点。
“Services”	MRS集群服务管理功能授权，勾选“View”表示可以查看服务，勾选“Management”表示可以管理服务。
“System_cluster_management”	MRS集群管理授权，勾选“Management”表示可以使用MRS补丁管理功能。
“System_configuration”	MRS集群配置功能授权，勾选“Management”表示可以使用Manager配置MRS集群。
“System_task”	MRS集群任务功能授权，勾选“Management”表示可以使用Manager管理MRS集群的周期任务。
“Tenant”	Manager多租户管理功能授权，勾选“Management”表示可以查看Manager的租户管理页面。

表 11-47 HBase 权限描述

支持权限管理的资源	权限设置说明
“SUPER_USER_GROUP”	选中时表示授予HBase管理员权限。
“Global”	HBase的一种资源类型，表示HBase整体组件。

支持权限管理的资源	权限设置说明
“Namespace”	HBase的一种资源类型，表示命名空间，用来保存HBase表。具体权限： ● “Admin”：表示管理此命名空间的权限。 ● “Create”：表示在此命名空间创建HBase表的权限。 ● “Read”：表示访问此命名空间的权限。 ● “Write”：表示写入此命名空间数据的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“Table”	HBase的一种资源类型，表示数据表，用来保存数据。具体权限： ● “Admin”：表示管理此数据表的权限。 ● “Create”：表示在此数据表创建列族和列的权限。 ● “Read”：表示读取数据表的权限。 ● “Write”：表示写入数据到表的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“ColumnFamily”	HBase的一种资源类型，表示列族，用来保存数据。具体权限： ● “Create”：表示在此列族创建列的权限。 ● “Read”：表示读取列族的权限。 ● “Write”：表示写入数据到列族的权限。
“Qualifier”	HBase的一种资源类型，表示列，用来保存数据。具体权限： ● “Read”：表示读取列的权限。 ● “Write”：表示写入数据到列的权限。

HBase中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如命名空间“default”添加了“Read”和“Write”权限，则命名空间中的表、列族和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 11-48 HDFS 权限描述

支持权限管理的资源	权限设置说明
“Folder”	HDFS的一种资源类型，表示HDFS目录，可以保存文件或子目录。具体权限： ● “Read”：表示访问此HDFS目录的权限。 ● “Write”：表示在此HDFS目录写入数据的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。
“Files”	HDFS的一种资源类型，表示HDFS中的文件。具体权限： ● “Read”：表示访问此文件的权限。 ● “Write”：表示写入此文件的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。

HDFS中每一级目录的权限默认不会传递到下级目录类型。例如目录“tmp”添加了“Read”和“Execute”，需要同时勾选“递归”才能为子目录添加权限。

表 11-49 Hive 权限描述

支持权限管理的资源	权限设置说明
“Hive Admin Privilege”	选中时表示授予Hive管理员权限。
“Database”	Hive的一种资源类型，表示Hive数据库，用来保存Hive表。具体权限： ● “Select”：表示查询Hive数据库的权限。 ● “Delete”：表示在Hive数据库执行删除操作的权限。 ● “Insert”：表示在Hive数据库执行插入操作的权限。 ● “Create”：表示在Hive数据库执行创建操作的权限。

支持权限管理的资源	权限设置说明
“Table”	Hive的一种资源类型，表示Hive表，用来保存数据。具体权限： ● “Select”：表示查询Hive表的权限。 ● “Delete”：表示在Hive表执行删除操作的权限。 ● “Update”：表示为角色添加Hive表的“Update”权限。 ● “Insert”：表示在Hive表执行插入操作的权限。 ● “Grant of Select”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Select”权限。 ● “Grant of Delete”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Delete”权限。 ● “Grant of Update”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Update”权限。 ● “Grant of Insert”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Insert”权限。

Hive中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如数据库“default”添加了“Select”和“Insert”权限，则数据库中的表和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 11-50 YARN 权限描述

支持权限管理的资源	权限设置说明
“Cluster Admin Operations”	选中时表示授予YARN管理员权限。
“root”	YARN的根队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Parent Queue”	YARN的一种资源类型，表示父队列，可以包含子队列。根队列也属于父队列的一种。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Leaf Queue”	YARN的一种资源类型，表示叶子队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。

YARN中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如队列“root”添加了“Submit”权限，则子队列自动添加该权限。子队列继承的权限不在“权限”表格显示被选中。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 11-51 Hue 权限描述

支持权限管理的资源	权限设置说明
“Storage Policy Admin”	选中时表示授予Hue中存储策略管理员权限。

步骤5 单击“确定”完成，返回“角色管理”。

----结束

相关任务

修改角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要修改角色所在的行，单击“修改”，修改角色信息。

说明

修改角色分配的权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要删除角色所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

11.13.4 创建用户组

操作场景

该任务指导管理员用户通过Manager创建新用户组并指定其操作权限，使用户组可以统一管理加入用户组的单个或多个用户。用户加入用户组后，可获得用户组具有的操作权限。

Manager支持用户组数为100。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**用户组管理**章节。

前提条件

- 管理员用户已明确业务需求，并已创建业务场景需要的角色。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户组管理”。

步骤4 在组列表上方，单击“添加用户组”。

步骤5 填写“组名”和“描述”。

“组名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。“描述”为可选参数。

步骤6 在“角色”，单击“选择添加角色”选择指定的角色并添加。

如果不添加角色，则当前创建的用户组没有使用MRS集群的权限。

步骤7 单击“确定”完成用户组创建。

----结束

相关任务

修改用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要修改用户组所在的行，单击“修改”，修改用户组信息。

说明

为用户组修改分配的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要删除用户组所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

11.13.5 创建用户

操作场景

该任务指导管理员根据实际业务场景需要，通过Manager创建新用户并指定其操作权限以满足业务使用。

Manager支持的用户数为1000。

如需对新创建用户的密码使用新的密码策略，请先**修改密码策略**，再参考本章节创建用户。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**创建用户**章节。

前提条件

- 管理员已明确业务需求，并已创建业务场景需要的角色和用户组。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在用户列表上方，单击“添加用户”。

步骤5 根据界面提示配置参数，填写“用户名”。

说明

- 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。
- 使用已创建的用户时，请输入和用户名完全一样的大小写字符。
- “用户名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。
- “root”、“omm”和“ommdba”为系统保留用户，请选择其他用户名。

步骤6 设置“用户类型”，可选值包括“人机”和“机机”。

- “人机”用户：用于在MRS Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。
- “机机”用户：用于MRS应用开发的场景。选择该值用户密码随机生成，无需填写。

步骤7 在“用户组”，单击“选择添加的用户组”，选择对应用户组将用户添加进去。

说明

- 如果用户组添加了角色，则用户可获得对应角色中的权限。
- 为新用户分配Hive的权限，请将用户加入hive组。
- 如果用户需要管理租户资源，用户组必须分配了Manager_tenant角色以及租户对应的角色。
- 通过Manager创建的用户无法添加到通过IAM用户同步功能同步的用户组中。

步骤8 在“主组”选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。

步骤9 根据业务实际需要在“分配角色权限”，单击“选择并绑定角色”为用户添加角色。

说明

- 创建用户时，如果用户从用户组获得的权限还不满足业务需要，则可以再分配其他已创建的角色。为新用户分配角色授权，最长可能需要3分钟时间生效。
- 创建用户时添加角色可细化用户的权限。
- 没有为新用户分配角色时，此用户可以访问HDFS、HBase、Yarn、Spark和Hue的WebUI。

步骤10 根据业务实际需要“描述”。

“描述”为可选参数。

步骤11 单击“确定”完成用户创建。

第一次在MRS集群中使用新创建的用户，例如登录Manager或者使用集群客户端，需要修改密码，具体请参见**修改操作用户密码**。

----结束

11.13.6 修改用户信息

操作场景

该任务指导管理员用户在Manager修改已创建的用户信息，包括修改用户组、主组、角色和描述。

开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**修改用户信息**章节。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要修改用户所在的行，单击“修改”，修改用户信息。

说明

为用户修改用户组或分配的角色权限，最长可能需要3分钟时间生效。

步骤5 单击“确定”完成修改操作。

----结束

11.13.7 锁定用户

该任务指导管理员用户将MRS集群中的用户锁定。用户被锁定后，不能在Manager重新登录或在集群中重新进行安全认证。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

可通过以下两种方式锁定用户，锁定后的用户需要管理员手动解锁或者等待锁定时间结束才能恢复使用：

- 自动锁定：通过设置密码策略中的“允许输入错误次数”，将超过登录失败次数的用户自动锁定。具体操作请参见**修改密码策略**。
- 手动锁定：由管理员通过手动的方式将用户锁定。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**锁定用户**章节。

以下将具体介绍手动锁定。不支持锁定“机机”用户。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要锁定用户所在行，单击“锁定用户”，锁定用户。

步骤5 在弹出的提示窗口，单击“确定”完成锁定操作。

----结束

11.13.8 解锁用户

在用户输入错误密码次数大于允许输入错误次数，造成用户被锁定或者用户被管理员手动锁定后需要解锁用户的场景下，管理员用户可以通过Manager为锁定的用户解锁。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**解锁用户**章节。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要解锁用户所在行，选择“解锁用户”，解锁用户。

步骤5 在弹出的提示窗口，单击“确定”完成解锁操作。

----结束

11.13.9 删除用户

MRS集群用户不再需要使用时，管理员可以在MRS Manager中删除此用户。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持删除用户操作。

说明

如果删除用户A后，再次准备重新创建同名用户A，如果该用户A已经提交过作业（客户端提交或者MRS console页面提交），那么需要在删除该用户A的同时，删除该用户A残留的文件夹，否则使用重新创建的同名用户A提交作业会失败。

删除用户残留文件夹操作方法为：依次登录MRS集群的Core节点，在每个Core节点上执行如下两条命令，其中如下命令中“$user”为具体的以用户名命名的文件夹。

cd /srv/BigData/hadoop/data1/nm/localdir/usercache/

rm -rf $user

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**删除用户**章节。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要删除用户所在的行，选择“更多 > 删除”。

图 11-2 删除用户

步骤5 单击“确定”完成删除操作。

----结束

11.13.10 修改操作用户密码

操作场景

出于MRS集群安全的考虑，“人机”类型系统用户密码必须定期修改。该任务指导用户通过MRS Manager完成修改自身密码工作。

如需对用户修改的密码使用新的密码策略，请先**修改密码策略**，再参考本章节修改密码。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**修改用户密码**章节。

对系统的影响

修改MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

前提条件

- 从管理员获取当前的密码策略。
- 从管理员获取MRS Manager访问地址。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，移动鼠标到界面右上角的。

在弹出菜单，选择“修改密码”。

步骤3 分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

11.13.11 初始化系统用户密码

操作场景

该任务指导管理员在用户遗忘密码或公共帐号密码需要定期修改时，通过Manager初始化密码。初始化密码后用户首次使用需要修改密码。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**初始化用户密码**章节。

对系统的影响

初始化MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

初始化“人机”用户密码

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要初始化密码用户所在行，单击“更多 > 初始化密码”，按界面提示信息修改用户密码。

在弹出窗口中输入当前登录的管理员密码确认管理员身份，单击“确定”，然后在“初始化密码”单击“确定”。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

初始化“机机”用户密码

步骤1 根据业务情况，准备好客户端，并登录安装客户端的节点。

步骤2 执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

kadmin -p kadmin/admin

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如：**cpw oms/manager**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。

- 不能与用户名或倒序的用户名相同。

----结束

11.13.12 下载用户认证文件

操作场景

用户开发大数据应用程序并在支持Kerberos认证的MRS集群中运行此程序时，需要准备访问MRS集群的“机机”用户认证文件。认证文件中的keytab文件可用于认证用户身份。

该任务指导管理员用户通过Manager下载“机机”用户认证文件并导出keytab文件。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

如果选择下载“人机”用户的认证文件，在下载前需要使用Manager修改过一次此用户的密码使管理员设置的初始密码失效，否则导出的keytab文件无法使用。请参见**修改操作用户密码**。

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**导出认证凭据文件**章节。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在需导出keytab文件用户所在的行，选择“更多 > 下载认证凭据”下载认证文件，待文件自动生成后指定保存位置，并妥善保管该文件。

步骤5 使用解压程序打开认证文件。

- “user.keytab”表示用户keytab文件，用于认证用户身份。
- “krb5.conf”表示认证服务器配置文件，应用程序在进行用户认证身份时根据该文件的配置信息连接认证服务器。

----结束

11.13.13 修改密码策略

操作场景

须知

密码策略涉及用户管理的安全性，请根据业务安全要求谨慎修改，否则会有安全性风险。

该任务指导管理员用户设置密码安全规则、用户登录安全规则及用户锁定规则。由于“机机”用户密码随机生成，在MRS Manager设置密码策略只影响“人机”用户。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

如需对新增用户的密码或用户修改的密码使用新的密码策略，请先参考本章节修改密码策略，再**创建用户**或**修改密码**。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**配置密码策略**章节。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 单击“密码策略配置”。

步骤4 根据界面提示，修改密码策略，具体参数见下表。

表 11-52 密码策略参数说明

参数名称	描述
最小密码长度	密码包含的最小字符个数，取值范围是8～32。默认值为“8”。
字符类型的数目	密码字符包含大写字母、小写字母、数字、空格和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=）的最小种类。可选择数值为“3”和“4”。默认值“3”表示至少必须使用大写字母、小写字母、数字、特殊符号和空格中的任意3种。
密码有效期（天）	密码有效使用天数，取值范围0～90，0表示永久有效。默认值为“90”。
密码失效提前提醒天数	提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录MRS Manager时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒。默认值为“5”。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔（分钟），取值范围为0～1440。“0”表示永远有效，“1440”表示1天。默认值为“5”。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。默认值为“5”。
用户锁屏时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。默认值为“5”。

----结束

11.13.14 配置跨集群互信

操作场景

集群A需要访问另一个集群B的资源前，需要管理员用户为这两个集群设置互信。

如果未配置跨集群互信，每个集群资源仅能被本集群用户访问。MRS自动为每个集群定义一个唯一且不重复的“域名”，用于表示用户的基本使用范围。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**配置跨Manager集群互信**章节。

对系统的影响

- 配置跨集群互信后，外部集群的用户可以在本集群中跨域使用，请根据业务与安全要求，定期检视集群中用户的权限。
- 安全集群配置跨集群互信，需要重启KrbServer服务，集群在重启期间无法使用。
- 配置跨集群互信后，互信的两个集群均会增加内部用户“krbtgt/*本集群域名@外部集群域名*”、“krbtgt/*外部集群域名@本集群域名*”，不支持删除。。

操作步骤

步骤1 在MRS管理控制台，分别查看两个集群的所有安全组。

- 当两个集群的安全组相同，请执行**步骤3**。
- 当两个集群的安全组不相同，请执行**步骤2**。

步骤2 在VPC管理控制台，分别为每个安全组添加规则。

规则的“协议”为“ANY”，“方向”为“入方向”。

“源地址”为“安全组”且是对端集群的安全组。

- 为A集群的安全组添加入方向规则，源地址选择B集群（对端集群）的安全组。
- 为B集群的安全组添加入方向规则，源地址选择A集群（对端集群）的安全组。

说明

未开启Kerberos认证的普通集群执行**步骤1**~**步骤2**即可完成跨集群互信配置，开启Kerberos认证的安全集群请继续执行后续步骤进行配置。

步骤3 参见**访问MRS Manager（MRS 2.x及之前版本）**分别登录两个集群MRS Manager，单击“服务管理”，查看全部组件的“健康状态”结果，是否全为“良好”？

- 是，执行**步骤4**。
- 否，任务结束，联系支持人员检查状态。

步骤4 查看配置信息。

1. 分别在两个集群MRS Manager，选择“服务管理 > KrbServer > 实例”，查看两个KerberosServer部署主机的“管理IP”.。
2. 单击“服务配置”，将“基础配置”切换为“全部配置”并在左侧导航树上选择“KerberosServer > 端口”，查看“kdc_ports”的值，默认值为“21732”。

3. 单击“域”，查看“default_realm”的值。

步骤5 在其中一个集群的MRS Manager，修改配置参数“peer_realms”。

表 11-53 相关参数

参数名	描述
“realm_name”	填写互信集群的域名，即**步骤4**中获得的“default_realm”的值。
“ip_port”	填写互信集群的KDC地址，参数值格式为：*外部集群KerberosServer部署的节点IP地址:kdc_port*。 两个KerberosServer的IP地址使用逗号分隔，例如KerberosServer部署在10.0.0.1和10.0.0.2上，则对应参数值为“10.0.0.1:21732,10.0.0.2:21732”。

说明

- 如果需要配置与多个集群的互信关系，请单击 添加新项目，并填写参数值。删除多余的配置项请单击 。
- 最多支持与16个集群配置互信，且本集群的不同互信集群之间默认不存在互信关系，需要另外添加。

步骤6 单击“保存配置”，在弹出窗口中勾选“重新启动受影响的服务或实例。”，单击“确定”重启服务。若未勾选“重新启动受影响的服务或实例。”，请手动重启受影响的服务或实例。

界面提示“操作成功”，单击“完成”，服务成功启动。

步骤7 退出MRS Manager，重新登录正常表示配置已成功。

步骤8 在另外一个集群的MRS Manager，重复**步骤5**到**步骤7**。

----结束

后续操作

配置跨集群互信后，因在MRS Manager修改了服务配置参数并重启了服务，请重新准备好客户端配置文件并更新客户端。

场景1：

A集群和B集群（对端集群、互信集群）是同类型集群，例如均是分析集群或者流式集群，请参见**更新客户端（3.x之前版本）**分别更新客户端配置文件。

- 更新A集群的客户端配置文件。
- 更新B集群的客户端配置文件。

场景2：

A集群和B集群（对端集群、互信集群）是不同类型集群，请执行如下步骤分别更新对端集群的配置文件到本端集群和本端集群自身的配置文件。

- 将A集群的客户端配置文件更新到B集群上。
- 将B集群的客户端配置文件更新到A集群上。
- 更新A集群的客户端配置文件。
- 更新B集群的客户端配置文件。

步骤1 登录MRS Manager(A集群)。

步骤2 单击“服务管理”，然后单击“下载客户端”。

步骤3 “客户端类型”选择“仅配置文件”。

步骤4 “下载路径”选择“远端主机”。

步骤5 将“主机IP”设置为B集群的主Master节点IP地址，设置“主机端口”为“22”，并将“存放路径”设置为“/tmp”。

- 如果使用SSH登录B集群的默认端口“22”被修改，请将“主机端口”设置为新端口。
- “存放路径”最多可以包含256个字符。

步骤6 “登录用户”设置为“root”。

如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。

步骤7 在“登录方式”选择“密码”或“SSH私钥”。

- 密码：输入创建集群时设置的root用户密码。
- SSH私钥：选择并上传创建集群时使用的密钥文件。

步骤8 单击“确定”开始生成客户端文件。

若界面显示以下提示信息表示客户端包已经成功保存。单击“关闭”。

```
下载客户端文件到远端主机成功。
```

若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从**步骤2**执行重新开始下载客户端。

```
连接到服务器失败，请检查网络连接或参数设置。
```

步骤9 使用VNC方式，登录弹性云服务器（B集群）。参见 。

步骤10 执行以下命令切换到客户端目录，例如“/opt/Bigdata/client”：

cd /opt/Bigdata/client

步骤11 执行以下命令，将A集群的客户端配置更新到B集群上：

sh refreshConfig.sh *客户端安装目录 客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/Bigdata/client /tmp/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

说明

步骤**步骤1**~**步骤11**的操作也可以参考**更新客户端（3.x之前版本）**页面的方法二操作。

步骤12 参见**步骤1**~**步骤11**，将B集群的客户端配置文件更新到A集群上。

步骤13 参见**更新客户端（3.x之前版本）**，分别更新本端集群自身的客户端配置文件：

- 更新A集群的客户端配置文件。
- 更新B集群的客户端配置文件。

----结束

11.13.15 配置并使用互信集群的用户

操作场景

配置完跨集群互信后，需要在互信的集群上设置用户的权限，这样本集群中的用户才能访问互信集群中同名用户可访问的资源。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

MRS 3.x及之后版本集群请参考**配置跨Manager集群互信**章节。

前提条件

已完成跨集群互信配置，然后刷新两个集群的客户端。

操作步骤

步骤1 在集群A的MRS Manager，选择“系统设置 > 用户管理”，检查互信集群B的用户，是否在A集群中已存在相同名字用户。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 单击用户名左侧的 展开用户的详细信息，检查该用户所在的用户组和角色分配的权限是否满足本次业务需求。

例如，集群A的“admin”用户拥有查看本集群HDFS中目录“/tmp”并创建文件的权限，然后执行**步骤4**。

步骤3 创建业务所需要使用的用户，同时关联业务所需要的用户组或者角色。然后执行**步骤4**。

步骤4 选择“服务管理 > HDFS > 实例”，查看“NameNode(主)”的“管理IP”。

步骤5 登录集群B的客户端节点。

例如在Master2节点更新客户端，则在该节点登录客户端，具体参见**使用MRS客户端**。

步骤6 执行以下命令，查看集群A中的目录“/tmp”。

hdfs dfs -ls hdfs://192.168.6.159:9820/tmp

其中，**192.168.6.159**是集群A中主NameNode的IP地址，9820是客户端与NameNode通信的默认端口。

步骤7 执行以下命令，在集群A中的目录“/tmp”创建一个文件。

hdfs dfs -touchz hdfs://192.168.6.159:9820/tmp/mrstest.txt

访问集群A，在目录“/tmp”中可查询到mrstest.txt文件，则表示配置跨集群互信成功。

----结束

11.13.16 配置 MRS 多用户访问 OBS 细粒度权限

开启细粒度权限时，用户通过该指导配置访问OBS权限，实现MRS用户对OBS文件系统下的目录权限控制。

如需对MRS的用户访问OBS的资源进行详细控制，可通过该功能实现。例如，您只允许用户组A访问某一OBS文件系统中的日志文件，您可以执行以下操作来实现：

1. 为MRS集群配置OBS访问权限的委托，实现使用ECS自动获取的临时AK/SK访问OBS。避免了AK/SK直接暴露在配置文件中的风险。
2. 在IAM中创建一个只允许访问某一OBS文件系统中的日志文件的策略，并创建一个绑定该策略权限的委托。
3. 在MRS集群中，新建的委托与MRS集群中的用户组A进行绑定，即可实现用户组A只拥有访问某一OBS文件系统中的日志文件的权限。

在以下场景运行作业时，提交作业的用户名为内置用户名，无法实现MRS多用户访问OBS：

- spark-beeline在安全集群中提交作业的内置用户名为spark，在普通集群中提交作业的内置用户名为omm。
- hbase shell在安全集群提交作业的内置用户名为hbase，在普通集群中提交作业的内置用户名为omm。
- Presto在安全集群提交作业的内置用户名为omm、hive，在普通集群提交作业的内置用户名为omm（当通过“组件管理 > Presto > 服务配置”，选择“全部配置”并搜索修改参数hive.hdfs.impersonation.enabled的值为true可以实现MRS多用户访问OBS细粒度权限功能）。

前提条件

- 开启细粒度权限控制的用户，权限管理请参考**创建MRS操作用户**。
- 需要对和OBS细粒度策略有一定了解。

步骤一：给集群配置有 OBS 访问权限的委托

步骤1 请参考**配置存算分离集群（委托方式）**配置OBS访问权限的委托。

配置的委托对该集群上所有用户（包括内置用户）及用户组生效，如需对集群上的用户及用户组访问OBS的权限进行控制请继续执行后续步骤。

----结束

步骤二：在 IAM 服务创建策略及委托

创建拥有不同访问权限的策略，并将策略与委托进行绑定，具体操作请参考**在IAM服务创建策略及委托**。

步骤三：在 MRS 集群详情页面配置 OBS 权限控制映射关系

步骤1 在MRS控制台，选择“集群列表 > 现有集群”并单击集群名称。

步骤2 在“概览”页签的基本信息区域，单击“OBS权限控制”右侧的“单击管理”。

步骤3 单击“添加映射”，并参考**表11-54**配置相关参数。

表 11-54 OBS 权限控制参数

参数	说明
IAM委托	选择**步骤2**中创建的委托。
类型	• User：在用户级别进行映射 • Group：用户组级别进行映射 **说明** • 用户级别的映射优先级大于用户组级别的映射。若选择Group，建议在“MRS用户（组）”一栏，填写用户的主组名称。 • 请避免同个用户名（组）出现在多个映射记录上的情况。
MRS 用户（组）	MRS中的用户（组）的名称，以英文逗号进行分隔。 **说明** • 对于没有配置在OBS权限控制的用户，且没有配置AK、SK时，将以MRS_ECS_DEFAULT_AGENCY中的的权限访问OBS。对于组件内置用户不建议绑定在委托中。 • 如需对组件内置用户在以下场景提交作业时配置委托，要求如下： – 如需对spark-beeline的操作进行权限控制，安全集群时配置用户名“spark”，普通集群时配置用户名“omm”。 – 如需对hbase shell的操作进行权限控制，安全集群时配置用户名“hbase”，普通集群时配置用户名“omm”。 – 如需对Presto的操作进行权限控制，安全集群时配置用户名“omm”、“hive”和登录客户端的用户名，普通集群时配置用户名“omm”和登录客户端的用户名。 – 如需使用Hive在beeline模式下创建表时，配置内置用户“hive”。

步骤4 单击“确定”。

步骤5 勾选“我同意授权MRS用户（组）与IAM委托之间的信任关系。”，并单击“确定”，完成MRS用户与OBS权限的映射关系。

当集群详情页面“概览”页签的“OBS权限控制”后出现 或OBS权限控制的映射表已刷新，表示映射生效（过程大约需要1分钟）。

在关系列表的“操作”列可以对已添加的关系进行编辑和删除。

说明

- 对于没有配置在OBS权限控制的用户，且没有配置AK、SK时，将以集群配置的委托在“对象存储服务”项目下所拥有的权限访问OBS。
- 无论用户是否配置OBS权限控制，只要配置AK、SK时，将以AK、SK的权限访问OBS。
- 映射关系的修改、创建、删除需要用户有Security Administrator权限。
- 修改映射关系后，若想使之在spark-beeline中生效，需要重启Spark服务，若想使之在hive beeline中生效，需要退出beeline重新进入，若想使之在Presto服务中生效，需要重启Presto服务。

----结束

在开启 OBS 权限控制功能时各组件访问 OBS 的说明

步骤1 以root用户登录集群任意一个节点，密码为用户创建集群时设置的root密码。

步骤2 配置环境变量（MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。）。

source /opt/Bigdata/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 如果当前集群未启用Kerberos认证，执行如下命令登录执行操作的用户，该用户需要属于supergroup组，创建用户可参考创建用户，将XXXX替换成用户名。

mkdir /home/XXXX

chown XXXX /home/XXXX

su - XXXX

步骤5 访问OBS。无需再配置AK、SK和endpoint。OBS路径格式：obs://buck_name/XXX。

例如：**hadoop fs -ls "obs://obs-example/job/hadoop-mapreduce-examples-3.1.2.jar"**

说明

- 如需使用hadoop fs删除OBS上文件，请使用**hadoop fs -rm -skipTrash**来删除文件。
- spark-sql、spark-beeline在创建表时，若不涉及数据导入，则不会访问OBS。即若在一个无权限的OBS目录下创建表，CREATE TABLE仍会成功，但插入数据会报403 AccessDeniedException。

----结束

在 IAM 服务创建策略及委托

步骤1 在IAM服务创建策略。

1. 登录IAM服务控制台。
2. 单击“权限 > 创建自定义策略”。
3. 参考**表11-55**填写参数。常用的OBS自定义策略样例请参考。

表 11-55 策略参数

参数	说明
策略名称	只能包含如下字符：大小写字母、中文、数字、空格和特殊字符（-_.,）。
作用范围	选择全局级服务，OBS为全局服务。
配置策略方式	选择可视化视图。
策略内容	1. “允许”选择“允许”。 2. “云服务”选择“对象存储服务(OBS)”。 3. “操作”勾选所有“写”、“列表”和“只读”权限。 4. “特定资源”选择： a. “object”选择“通过资源路径指定”，并单击“添加资源路径”分别输入路径***obs_bucket_name*/tmp/**和***obs_bucket_name*/tmp/***。此处以**/tmp**目录为例，如需其他目录权限请参考该步骤添加对应目录及该目录下所有对象的资源路径。 b. “bucket”选择“通过资源路径指定”，并单击“添加资源路径”输入路径***obs_bucket_name***。 5. （可选）请求条件，暂不添加。
策略描述	可选，对策略的描述。

说明

各个组件的写数据操作若通过rename的方式实现时，写数据时要配置删除对象的权限。

4. 单击“确定”保存策略。

步骤2 在IAM服务创建委托。

1. 登录IAM服务控制台。

2. 单击“委托 > 创建委托”。
3. 参考**表11-56**填写参数。

表 11-56 委托参数

参数	说明
委托名称	只能包含如下字符：大小写字母、中文、数字、空格和特殊字符（ -_., ）。
委托类型	选择普通帐号。
委托的帐号	填写本用户的云帐号，即使用手机号开通的帐号，不能是联邦用户或者IAM用户。
持续时间	请根据需要选择。
描述	可选，对委托的描述。
权限选择	1. 在“项目”列对应的“对象存储服务”行，单击“操作”列的“修改”。 2. 勾选**步骤1**中创建的策略，使之出现在“已选择策略中”。 3. 单击“确定”。

4. 单击“确定”保存委托。

说明

当使用该委托访问过OBS后，再修改该委托及其绑定的策略时，最长需要等待15分钟，修改的内容才能生效。

----结束

11.14 补丁操作指导

11.14.1 补丁操作指导

当您通过如下途径获知集群版本补丁信息，请根据您的实际需求进行补丁升级操作。

- 通过消息中心服务推送的消息获知MapReduce服务发布了补丁信息。
- 进入现有集群，查看“补丁信息”页面，呈现补丁信息。

安装补丁前准备

- 请参见**执行健康检查**检查集群状态，确认集群健康状态正常后再安装补丁。
- 您根据“补丁内容”中的补丁信息描述，确认将要安装的目标补丁。

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁信息”页面，在操作列表中单击“安装”，安装目标补丁。

说明

- 滚动补丁操作请参见**支持滚动补丁**。
- 对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁信息”页面，在操作列表中单击“卸载”，卸载目标补丁。

说明

- 滚动补丁操作请参见**支持滚动补丁**。
- 对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

11.14.2 支持滚动补丁

滚动补丁是指在补丁安装/卸载时，采用滚动重启服务（按批次重启服务或实例）的方式，在不中断或尽可能短地中断集群各个服务业务的前提下完成对集群中单个或多个服务的补丁安装/卸载操作。集群中的服务根据对滚动补丁的支持程度，分为三种：

- 支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的全部业务或部分业务（因服务而异，不同服务存在差别）不中断。
- 不支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的业务会中断。
- 部分角色支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的部分业务不中断。

当前MRS集群中，服务和实例是否支持滚动重启如**表11-57**所示。

表 11-57 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	

服务	实例	是否支持滚动重启
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	
	Logviewer	

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁信息”页面，在操作列表中单击“安装”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动安装补丁功能开启：补丁安装前不会停止服务，补丁安装后滚动重启服务来完成补丁安装，可以减少对集群业务的影响，但相比普通方式安装耗时更久。
- 滚动安装补丁功能关闭：补丁安装前会停止服务，补丁安装后再重新启动服务来完成补丁安装，会造成集群和服务暂时中断，但相比滚动方式安装补丁耗时更短。
- 少于2个Master节点和少于3个Core节点的集群不支持滚动方式安装补丁。

步骤5 单击“确定”，安装目标补丁。

步骤6 查看补丁安装进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁安装进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁信息”页面，在操作列表中单击“卸载”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动卸载补丁功能开启：补丁卸载前不会停止服务，补丁卸载后滚动重启服务来完成补丁卸载，可以减少对集群业务的影响，但相比普通方式卸载耗时更久。
- 滚动卸载补丁功能关闭：补丁卸载前会停止所有服务，补丁卸载后再重新启动所有服务来完成补丁卸载，会造成集群和服务暂时中断，但相比滚动方式卸载补丁耗时更短。
- 少于2个Master节点和少于3个Core节点的集群不支持滚动方式卸载补丁。

步骤5 单击“确定”，卸载目标补丁。

步骤6 查看补丁卸载进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁卸载进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

11.15 修复隔离主机补丁

若集群中存在主机被隔离的情况，集群补丁安装完成后，请参见本节操作对隔离主机进行补丁修复。修复完成后，被隔离的主机节点版本将与其他未被隔离的主机节点一致。

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“详情”。

步骤4 在补丁详情界面，选中“Status”是“Isolated”的主机节点。

步骤5 单击“Select and Restore”，修复被隔离的主机节点。

----结束

11.16 支持滚动重启

在修改了大数据组件的配置项后，需要重启对应的服务来使得配置生效，使用普通重启方式会并发重启所有服务或实例，可能引起业务断服。为了确保服务重启过程中，尽量减少或者不影响业务运行，可以通过滚动重启来按批次重启服务或实例（对于有主备状态的实例，会先重启备实例，再重启主实例）。滚动重启方式的重启时间比普通重启时间久。

当前MRS集群中，服务和实例是否支持滚动重启如**表11-58**所示。

表 11-58 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是

服务	实例	是否支持滚动重启
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	
	Logviewer	

使用限制

- 请在低业务负载时间段进行滚动重启操作。
  - 例如：在滚动重启kafka服务时候， 如果kafka服务业务吞吐量很高（100M/s以上的情况下），会出现kafka服务滚动重启失败的情况。
  - 例如：在滚动重启HBase服务时候，如果原生界面上每个RegionServer上每秒的请求数超过1W，需要增大handle数来预防重启过程中负载过大导致的RegionServer重启失败。
- 重启前需要观察当前hbase的负载请求数（原生界面上每个rs的请求数如果超过1W，需要增大handle数来预防到时候负载不过来）
- 在集群Core节点个数小于6个的情况下，可能会出现业务短时间受影响的情况。
- 请优先使用滚动重启操作来重启实例或服务，并勾选“仅重启配置过期的实例”。

滚动重启服务

步骤1 在MRS Manager，单击“服务管理”，选择需要滚动重启的服务，进入服务页面。

步骤2 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤3 输入管理员密码后，弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启服务。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启实例

步骤1 在MRS Manager，单击“服务管理”，选择需要滚动重启的服务，进入服务页面。

步骤2 在“实例”页签，勾选要重启的实例，单击“更多”，选择“滚动重启实例”。

步骤3 输入管理员密码后，弹出“滚动重启实例”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启实例。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启集群

步骤1 在MRS Manager，单击“服务管理”，进入服务管理页面。

步骤2 单击“更多”，选择“滚动重启集群”。

步骤3 输入管理员密码后，弹出“滚动重启集群”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启集群。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启参数说明

滚动重启参数说明如**表11-59**所示。

表 11-59 滚动重启参数说明

参数名称	描述
仅重启配置过期的实例	是否只重启集群内修改过配置的实例。
数据节点滚动重启并发数	采用分批并发滚动重启策略的数据节点实例每一个批次重启的实例数，默认为1，取值范围为1～20。只对数据节点有效。

参数名称	描述
批次时间间隔	滚动重启实例批次之间的间隔时间，默认为0，取值范围为0～2147483647，单位为秒。 说明：设置批次时间间隔参数可以增加滚动重启期间大数据组件进程的稳定性。建议设置该参数为非默认值，例如10。
批次容错阈值	滚动重启实例批次执行失败容错次数，默认为0，即表示任意一个批次的实例重启失败后，滚动重启任务终止。取值范围为0～214748364。

典型场景操作步骤

步骤1 在MRS Manager，单击“服务管理”，选择HBase，进入HBase服务页面。

步骤2 单击“服务配置”页签，修改HBase某个参数并保存配置，在出现如下弹窗后，单击“确定”进行保存。

说明

不要勾选“重新启动受影响的服务或实例”，该处重启是普通重启方式，会并发重启所有服务或实例，引起业务断服。

步骤3 保存配置完成后，单击“完成”。

步骤4 选择“服务状态”页签。

步骤5 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤6 输入管理员密码后，弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启。

步骤7 滚动重启任务完成后，单击“完成”。

----结束

12 MRS 集群组件操作指导

12.1 使用 Alluxio

12.1.1 配置底层存储系统

用户想要通过统一的客户端API和全局命名空间访问包括HDFS和OBS在内的持久化存储系统，从而实现了对计算和存储的分离时，可以在MRS Manager页面中配置Alluxio的底层存储系统来实现。集群创建后，默认的底层存储地址是hdfs://hacluster/，即将HDFS的根目录映射到Alluxio。

前提条件

- 已安装Alluxio服务的集群。
- 获取用户“admin”帐号密码。“admin”密码在创建MRS集群时由用户指定。

配置 HDFS 作为 Alluxio 的底层文件系统

说明

开启Kerberos认证的安全集群不支持该功能。

步骤1 请参考**修改集群服务配置参数**，进入Alluxio的“全部配置”页面。

步骤2 在左侧边栏中选择“Alluxio > 底层存储系统”，修改参数“alluxio.master.mount.table.root.ufs”的值为“hdfs://hacluster/XXX/”。

例如：若想将“HDFS根目录/alluxio/”作为alluxio的根目录，则修改参数“alluxio.master.mount.table.root.ufs”的值为“hdfs://hacluster/alluxio/”。

步骤3 单击“保存配置”，并在弹出窗口中勾选“重新启动受影响的服务和实例。”

步骤4 单击“确定”重启Alluxio服务。

----结束

12.1.2 通过数据应用访问 Alluxio

访问Alluxio文件系统的端口号是19998，即地址为alluxio://<alluxio的master节点ip>:19998/<PATH>，本节将通过示例介绍如何通过数据应用（Spark、Hive、Hadoop MapReduce和Presto）访问Alluxio。

使用 Alluxio 作为 Spark 应用程序的输入和输出

步骤1 以root用户登录集群的Master节点，密码为用户创建集群时设置的root密码。

步骤2 执行如下命令，配置环境变量。

source /opt/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行如下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 准备输入文件，将本地数据复制到Alluxio文件系统中。

如在本地/home目录下准备一个输入文件test_input.txt，然后执行如下命令，将test_input.txt文件放入Alluxio中。

alluxio fs copyFromLocal /home/test_input.txt /input

步骤5 执行如下命令启动spark-shell。

spark-shell

步骤6 在spark-shell中运行如下命令。

val s = sc.textFile("alluxio://<Alluxio的节点名称>:19998/input")

val double = s.map(line => line + line)

double.saveAsTextFile("alluxio://<Alluxio的节点名称>:19998/output")

说明

<Alluxio的节点名称>:19998，请根据实际情况替换为AlluxioMaster实例所在所有节点的节点名称与端口号，各个名称与端口之间以英文逗号间隔，例如：node-ana-coremspb.mrs-m0va.com:19998,node-master2kiww.mrs-m0va.com:19998,node-master1cqwv.mrs-m0va.com:19998

步骤7 使用“Ctrl + C”退出spark-shell。

步骤8 通过alluxio命令行**alluxio fs ls /**查看alluxio根目录下存在一个输出目录/output，其中包含了输入文件input的双倍内容。

----结束

在 Alluxio 上创建 Hive 表

步骤1 以root用户登录集群的Master节点，密码为用户创建集群时设置的root密码。

步骤2 执行如下命令，配置环境变量。

source /opt/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行如下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 准备输入文件，如在本地/home目录下准备一个输入文件hive_load.txt， 内容为

```
1, Alice, company A
2, Bob, company B
```

步骤5 执行如以下命令，将hive_load.txt文件放入Alluxio中。

alluxio fs copyFromLocal /home/hive_load.txt /hive_input

步骤6 执行如下命令启动hive beeline。

beeline

步骤7 在beeline中运行如下命令，根据Alluxio中的输入文件进行创表。

CREATE TABLE u_user(id INT, name STRING, company STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

LOAD DATA INPATH 'alluxio://<Alluxio的节点名称>:19998/hive_input' INTO TABLE u_user;

说明

<Alluxio的节点名称>:19998，请根据实际情况替换为AlluxioMaster实例所在所有节点的节点名称与端口号，各个名称与端口之间以英文逗号间隔，例如：node-ana-coremspb.mrs-m0va.com:19998,node-master2kiww.mrs-m0va.com:19998,node-master1cqwv.mrs-m0va.com:19998

步骤8 执行如下命令查看创建的表。

select * from u_user;

----结束

在 Alluxio 上运行 Hadoop Wordcount

步骤1 以root用户登录集群的Master节点，密码为用户创建集群时设置的root密码。

步骤2 执行如下命令，配置环境变量。

source /opt/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行如下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 准备输入文件，将本地数据复制到Alluxio文件系统中。

如在本地/home目录下准备一个输入文件test_input.txt，然后执行如下命令，将test_input.txt文件放入Alluxio中。

alluxio fs copyFromLocal /home/test_input.txt /input

步骤5 通过yarn jar执行wordcount作业。

yarn jar /opt/share/hadoop-mapreduce-examples-<hadoop版本号>-mrs-<mrs集群版本号>/hadoop-mapreduce-examples-<hadoop版本号>-mrs-<mrs集群版本号>.jar wordcount alluxio://<Alluxio的节点名称>:19998/input alluxio://<Alluxio的节点名称>:19998/output

说明

- <hadoop版本号>请根据实际情况替换。
- <mrs集群版本号>替换为MRS的大版本号，如MRS 1.9.2版本集群此处为**mrs-1.9.0**。
- <Alluxio的节点名称>:19998，请根据实际情况替换为AlluxioMaster实例所在所有节点的节点名称与端口号，各个名称与端口之间以英文逗号间隔，例如：node-ana-coremspb.mrs-m0va.com:19998,node-master2kiww.mrs-m0va.com:19998,node-master1cqwv.mrs-m0va.com:19998

步骤6 通过alluxio命令行**alluxio fs ls /**查看alluxio根目录下存在一个输出目录/output，包含了wordcount的结果。

----结束

使用 Presto 在 Alluxio 上查询表

步骤1 以root用户登录集群的Master节点，密码为用户创建集群时设置的root密码。

步骤2 执行如下命令，配置环境变量。

source /opt/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行如下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 启动hive beeline在alluxio上创建表。

beeline

CREATE TABLE u_user (id int, name string, company string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION 'alluxio://<Alluxio的节点名称>:19998/u_user';

insert into u_user values(1,'Alice','Company A'),(2, 'Bob', 'Company B');

说明

<Alluxio的节点名称>:19998，请根据实际情况替换为AlluxioMaster实例所在所有节点的节点名称与端口号，各个名称与端口之间以英文逗号间隔，例如：node-ana-coremspb.mrs-m0va.com:19998,node-master2kiww.mrs-m0va.com:19998,node-master1cqwv.mrs-m0va.com:19998

步骤5 启动Presto客户端，具体请参见**使用客户端执行查询语句**的**步骤2**~**步骤8**。

步骤6 在Presto客户端中执行查询语句**select * from hive.default.u_user;** 查询alluxio上创建表。

图 12-1 Presto 查询 alluxio 上创建的表

```
presto> select * from hive.default.u_user;
 id | name  |  company
----+-------+-----------
  1 | Alice | Company A
  2 | Bob   | Company B
(2 rows)
```

----结束

12.1.3 Alluxio 常用操作

前期准备

1. 创建安装Alluxio组件的集群。
2. 以root用户登录集群的主Master节点，密码为用户创建集群时设置的root密码。
3. 执行如下命令，配置环境变量。

   source /opt/client/bigdata_env

使用 Alluxio Shell

Alluxio shell包含多种与Alluxio交互的命令行操作。

- 要查看文件系统操作命令列表。

  alluxio fs
- 使用**ls**命令列出 Alluxio 里的文件。例如列出根目录下所有文件。

  alluxio fs ls /
- 使用copyFromLocal命令可以复制本地文件到 Alluxio 中。

  alluxio fs copyFromLocal /home/test_input.txt /test_input.txt

  命令执行后回显：

  ```
  Copied file:///home/test_input.txt to /test_input.txt
  ```
- 再次使用ls命令列出Alluxio中的文件，可以看到刚刚拷贝的test_input.txt文件。

  alluxio fs ls /

  命令执行后回显：

  ```
  12       PERSISTED 11-28-2019 17:10:17:449 100% /test_input.txt
  ```

  输出显示test_input.txt 文件在 Alluxio 中，各参数含义为文件的大小、是否被持久化、创建日期、Alluxio中这个文件的缓存占比、文件名。
- 使用cat命令打印文件的内容。

  alluxio fs cat /test_input.txt

  命令执行后回显：

  ```
  Test Alluxio
  ```

Alluxio 中的挂载功能

Alluxio 通过统一命名空间的特性统一了对存储系统的访问。详情请参考：**https://docs.alluxio.io/os/user/2.0/cn/advanced/Namespace-Management.html**

这个特性允许用户挂载不同的存储系统到Alluxio命名空间中并且通过Alluxio命名空间无缝地跨存储系统访问文件。

1. 在 Alluxio 中创建一个目录作为挂载点。

   alluxio fs mkdir /mnt

   ```
   Successfully created directory /mnt
   ```

2. 挂载一个已有的OBS文件系统到Alluxio（前提：给集群配置有OBS OperateAccess权限的委托）。此处以obs-mrstest文件系统为例，请根据实际情况替换文件系统名。

   alluxio fs mount /mnt/obs obs://obs-mrstest/data

   ```
   Mounted obs://obs-mrstest/data at /mnt/obs
   ```

3. 通过Alluxio命名空间列出OBS文件系统中的文件。使用**ls**命令列出OBS挂载目录下的文件。

   alluxio fs ls /mnt/obs

   ```
   38       PERSISTED 11-28-2019 17:42:54:554   0% /mnt/obs/hive_load.txt
   12       PERSISTED 11-28-2019 17:43:07:743   0% /mnt/obs/test_input.txt
   ```

   新挂载的文件和目录也可以通过Alluxio WebUI查看。

4. 挂载完成后，通过 Alluxio 统一命名空间，可以无缝地从不同存储系统中交互数据。例如，使用**ls -R**命令，递归地列举出一个目录下的所有文件。

   alluxio fs ls -R /

   ```
         0       PERSISTED 11-28-2019 11:15:19:719  DIR /app-logs
         1       PERSISTED 11-28-2019 11:18:36:885  DIR /apps
         1       PERSISTED 11-28-2019 11:18:40:209  DIR /apps/templeton
   239440292       PERSISTED 11-28-2019 11:18:40:209   0% /apps/templeton/hive.tar.gz
   .....
         1       PERSISTED 11-28-2019 19:00:23:879  DIR /mnt
         2       PERSISTED 11-28-2019 19:00:23:879  DIR /mnt/obs
        38       PERSISTED 11-28-2019 17:42:54:554   0% /mnt/obs/hive_load.txt
        12       PERSISTED 11-28-2019 17:43:07:743   0% /mnt/obs/test_input.txt
   .....
   ```

   输出显示了Alluxio文件系统根目录（默认值是HDFS的根目录，即hdfs://hacluster/）中来源于挂载存储系统的所有文件。/app-logs和/apps目录在HDFS文件系统中，/mnt/obs/目录在OBS 中。

用 Alluxio 加速数据访问

由于Alluxio利用内存存储数据，它可以加速数据的访问。例如：

1. 上传一个文件test_data.csv（文件是一份记录了食谱的样本）到obs-mrstest文件系统的/data目录下。通过**ls**命令显示文件状态：

   alluxio fs ls /mnt/obs/test_data.csv

   ```
   294520189       PERSISTED 11-28-2019 19:38:55:000   0% /mnt/obs/test_data.csv
   ```

   输出显示了该文件在Alluxio中缓存占比为0%，即不在Alluxio内存中。

2. 统计该文件中单词"milk"出现的次数，并计算耗时。

   time alluxio fs cat /mnt/obs/test_data.csv | grep -c milk

   ```
   52180

   real    0m10.765s
   user    0m5.540s
   sys     0m0.696s
   ```

3. 第一次读取数据后会将数据放在内存中，Alluxio再次读取时可以提高访问该数据的速度。例如：在通过**cat**命令获取文件后，用**ls**命令再查看文件的状态。

   alluxio fs ls /mnt/obs/test_data.csv

   ```
   294520189       PERSISTED 11-28-2019 19:38:55:000 100% /mnt/obs/test_data.csv
   ```

   输出显示文件已经 100% 被加载到 Alluxio 中。

4. 再次访问该文件，统计单词“eggs”出现的次数，并计算耗时。

   time alluxio fs cat /mnt/obs/test_data.csv | grep -c eggs

   ```
   59510

   real    0m5.777s
   user    0m5.992s
   sys     0m0.592s
   ```

   对比两次耗时可以看出存储在Alluxio内存中的数据，数据访问耗时明显缩短。

12.2 使用 CarbonData（MRS 3.x 之前版本）

12.2.1 从零开始使用 CarbonData

MRS 3.x之前版本参考本章节，MRS 3.x及后续版本请参考**使用CarbonData（MRS 3.x及之后版本）**。

本章节介绍使用Spark CarbonData的基本流程，所有任务场景基于spark-beeline环境。CarbonData快速入门包含以下任务：

1. 连接到Spark

   在对CarbonData进行任何操作之前，需要先连接到Spark。

2. 创建CarbonData表

   连接CarbonData之后，需要创建CarbonData Table，用于加载数据和执行查询操作。

3. 加载数据到CarbonData表

   用户从HDFS中的CSV文件加载数据到所创建的表中。

4. 在CarbonData中查询数据

   在CarbonData表加载数据之后，用户可以执行所需的查询操作，例如groupby或者where等。

前提条件

已安装客户端，具体参见**使用MRS客户端**。

操作步骤

步骤1 连接到Spark CarbonData。

1. 根据业务情况，准备好客户端，使用**root**用户登录安装客户端的节点。

   例如在Master2节点更新客户端，则在该节点登录客户端，具体参见**使用MRS客户端**。

2. 切换用户与配置环境变量。

   sudo su - omm

   source /opt/client/bigdata_env

3. 启用Kerberos认证的集群，执行以下命令认证用户身份。未启用Kerberos认证集群无需执行。

   kinit *Spark组件用户名*

说明

用户需要加入用户组hadoop、hive，主组hadoop。

4. 执行以下命令，连接到Spark运行环境：

   spark-beeline

步骤2 执行命令创建CarbonData表。

CarbonData表可用于加载数据和执行查询操作，例如执行以下命令创建CarbonData表：

CREATE TABLE x1 (imei string, deviceInformationId int, mac string, productdate timestamp, updatetime timestamp, gamePointId double, contractNumber double)

STORED BY 'org.apache.carbondata.format'

TBLPROPERTIES ('DICTIONARY_EXCLUDE'='mac','DICTIONARY_INCLUDE'='deviceInformationId');

命令执行结果如下：

```
+---------+--+
| result  |
+---------+--+
+---------+--+
No rows selected (1.551 seconds)
```

步骤3 从CSV文件加载数据到CarbonData表。

根据所要求的参数运行命令从CSV文件加载数据，且仅支持CSV文件。**LOAD**命令中配置的CSV列名，需要和CarbonData表列名相同，顺序也要对应。CSV文件中的数据的列数，以及数据格式需要和CarbonData表匹配。

文件需要保存在HDFS中。用户可以将文件上传到OBS，并在MRS管理控制台“文件管理”将文件从OBS导入HDFS。

如果集群启用了Kerberos认证，则需要在工作环境准备CSV文件，然后可以使用开源HDFS命令，参考5将文件从工作环境导入HDFS，并设置Spark组件用户在HDFS中对文件有读取和执行的权限。

例如，HDFS的“tmp”目录有一个文件“data.csv”，内容如下：

```
x123,111,dd,2017-04-20 08:51:27,2017-04-20 07:56:51,2222,33333
```

执行导入命令：

LOAD DATA inpath 'hdfs://hacluster/tmp/data.csv' into table x1 options('DELIMITER'=',','QUOTECHAR'='"','FILEHEADER'='imei, deviceinformationid,mac,productdate,updatetime,gamepointid,contractnumber');

命令执行结果如下：

```
+---------+--+
| Result  |
+---------+--+
+---------+--+
No rows selected (3.039 seconds)
```

步骤4 在CarbonData中查询数据。

- **获取记录数**

  为了获取在CarbonData table中的记录数，可以执行以下命令。

  select count(*) from x1;

- **使用Groupby查询**

  为了获取不重复的“deviceinformationid”记录数，可以执行以下命令。

  select deviceinformationid,count (distinct deviceinformationid) from x1 group by deviceinformationid;

- **使用条件查询**

  为了获取特定**deviceinformationid**的记录，可以执行以下命令。

  select * from x1 where deviceinformationid='111';

步骤5 执行以下命令退出Spark运行环境。

!quit

----结束

12.2.2 CarbonData 表简介

简介

CarbonData表与RDBMS中的表类似，RDBMS数据存储在由行和列构成的表中。CarbonData表存储的也是结构化的数据，具有固定列和数据类型。CarbonData中的数据存储在表实体文件中。

支持的数据类型

CarbonData表支持以下数据类型：

- Int
- String
- BigInt
- Decimal
- Double
- TimeStamp

表12-1对所支持的数据类型和对应的范围进行了详细说明。

表 12-1 CarbonData 数据类型

数据类型	描述
Int	4字节有符号整数，从-2,147,483,648到2,147,483,647。 **说明** 非字典列如果是Int类型，会在内部存储为BigInt类型。
String	最大支持字符长度为100000。

数据类型	描述
BigInt	使用64-bit存储数据，支持从-9,223,372,036,854,775,808到9,223,372,036,854,775,807。
Decimal	默认值是(10,0)，最大值是(38,38)。 **说明** 当进行带过滤条件的查询时，为了得到准确的结果，需要在数字后面加上BD。例如，**select * from carbon_table where num = 1234567890123456.22BD**。
Double	使用64-bit存储数据，从4.9E-324到1.7976931348623157E308。
TimeStamp	默认格式为“yyyy-MM-dd HH:mm:ss”。

说明

所有Integer类型度量均以BigInt类型进行处理与显示。

12.2.3 创建 CarbonData 表

操作场景

使用CarbonData前需先创建表，才可从表中加载数据和查询数据。

使用自定义列创建表

可通过指定各列及其数据类型来创建表。启用Kerberos认证的分析集群创建CarbonData表时，如果用户需要在默认数据库“default”以外的数据库创建新表，则需要在Hive角色管理中为用户绑定的角色添加指定数据库的“Create”权限。

命令示例：

```
CREATE TABLE IF NOT EXISTS productdb.productSalesTable (
productNumber Int,
productName String,
storeCity String,
storeProvince String,
revenue Int)
STORED BY 'org.apache.carbondata.format'
TBLPROPERTIES (
'table_blocksize'='128',
'DICTIONARY_EXCLUDE'='productName',
'DICTIONARY_INCLUDE'='productNumber');
```

上述命令所创建的表的详细信息如下：

表 12-2 表信息定义

参数	描述
productSalesTable	待创建的表的名称。该表用于加载数据进行分析。 表名由字母、数字、下划线组成。
productdb	数据库名称。该数据库将与其中的表保持逻辑连接以便于识别和管理。 数据库名称由字母、数字、下划线组成。
productNumber productName storeCity storeProvince revenue	表中的列，代表执行分析所需的业务实体。 列名（字段名）由字母、数字、下划线组成。 **说明** CarbonData暂不支持设置列是否允许为空、默认值以及主键。
table_blocksize	CarbonData表使用的数据文件的block大小，默认值为1024，取值范围为1～2048，单位为MB。 ● 如果“table_blocksize”值太小，数据加载时将生成过多的小数据文件，可能会影响HDFS的使用性能。 ● 如果“table_blocksize”值太大，数据查询时索引匹配的block数据量较大，导致读取并发度不高，从而降低查询性能。 一般情况下，建议根据数据量级别来选择大小。例如：GB级别用256，TB级别用512，PB级别用1024。
DICTIONARY_EXCLUDE	设置指定列不生成字典，适用于数值复杂度高的列。系统默认为String类型的列做字典编码，但是如果字典值过多，会导致字典转换操作增加造成性能下降。 一般情况下，列的数值复杂度高于5万，可以被认定为高复杂度，则需要排除掉字典编码，该参数为可选参数。 **说明** 在非字典列中，只支持String和Timestamp数据类型。
DICTIONARY_INCLUDE	设置指定列生成字典，适用于数值复杂度低的列，可以提升字典列上的**groupby**性能，为可选参数。一般情况下，字典列的复杂度不应该高于5万。

12.2.4 删除 CarbonData 表

操作场景

用户根据业务使用情况，可以删除不再使用的CarbonData表。删除表后，其所有的元数据以及表中已加载的数据都会被删除。

操作步骤

步骤1 运行如下命令删除表。

DROP TABLE *[IF EXISTS] [db_name.]table_name*;

“db_name”为可选参数。如果没有指定“db_name”，那么将会删除当前数据库下名为“table_name”的表。

例如执行命令，删除数据库“productdb”下的表“productSalesTable”：

DROP TABLE productdb.productSalesTable;

步骤2 执行以下命令查询表是否被删除：

SHOW TABLES;

----结束

12.3 使用 CarbonData（MRS 3.x 及之后版本）

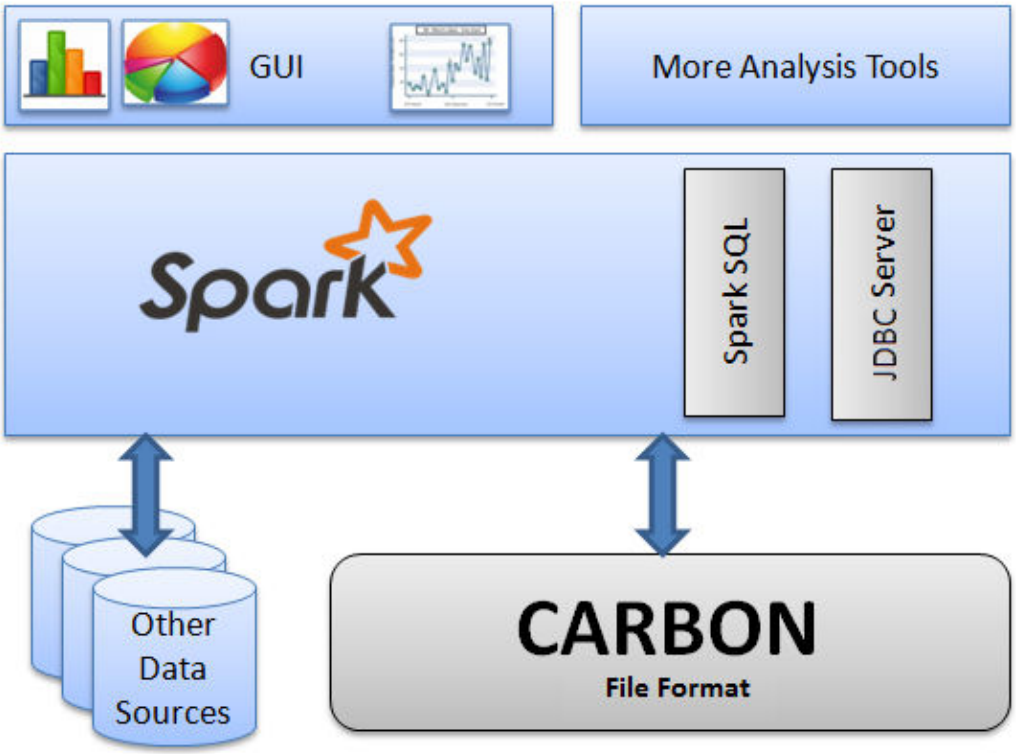
12.3.1 概述

MRS 3.x及后续版本参考本章节，MRS 3.x之前版本请参考**使用CarbonData（MRS 3.x之前版本）**。

12.3.1.1 CarbonData 简介

CarbonData是一种新型的Apache Hadoop本地文件格式，使用先进的列式存储、索引、压缩和编码技术，以提高计算效率，有助于加速超过PB数量级的数据查询，可用于更快的交互查询。同时，CarbonData也是一种将数据源与Spark集成的高性能分析引擎。

图 12-2 CarbonData 基本架构



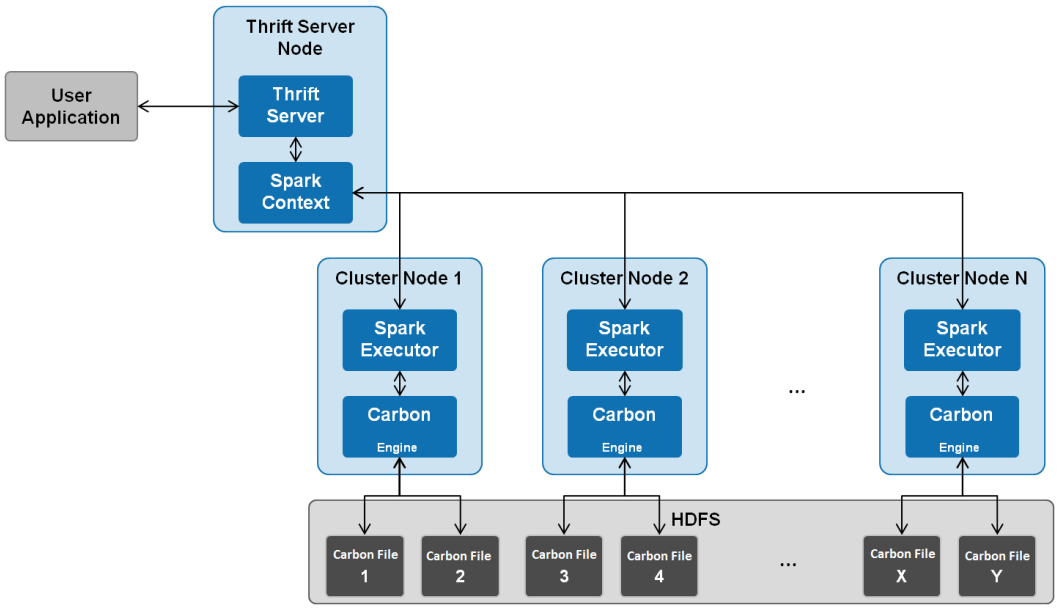
使用CarbonData的目的是对大数据即席查询提供超快速响应。从根本上说，CarbonData是一个OLAP引擎，采用类似于RDBMS中的表来存储数据。用户可将大量（10TB以上）的数据导入以CarbonData格式创建的表中，CarbonData将以压缩的多维索引列格式自动组织和存储数据。数据被加载到CarbonData后，就可以执行即席查询，CarbonData将对数据查询提供秒级响应。

CarbonData将数据源集成到Spark生态系统，用户可使用Spark SQL执行数据查询和分析。也可以使用Spark提供的第三方工具JDBCServer连接到Spark SQL。

CarbonData 结构

CarbonData作为Spark内部数据源运行，不需要额外启动集群节点中的其他进程，CarbonData Engine在Spark Executor进程之中运行。

图 12-3 CarbonData 结构



存储在CarbonData Table中的数据被分成若干个CarbonData数据文件，每一次数据查询时，CarbonData Engine模块负责执行数据集的读取、过滤等实际任务。CarbonData Engine作为Spark Executor进程的一部分运行，负责处理数据文件块的一个子集。

Table数据集数据存储在HDFS中。同一Spark集群内的节点可以作为HDFS的数据节点。

CarbonData 特性

- SQL功能：CarbonData与Spark SQL完全兼容，支持所有可以直接在Spark SQL上运行的SQL查询操作。
- 简单的Table数据集定义：CarbonData支持易于使用的DDL(数据定义语言)语句来定义和创建数据集。CarbonData DDL十分灵活、易于使用，并且足够强大，可以定义复杂类型的Table。
- 便捷的数据管理：CarbonData为数据加载和维护提供多种数据管理功能。CarbonData支持加载历史数据以及增量加载新数据。加载的数据可以基于加载时间进行删除，也可以撤销特定的数据加载操作。

- CarbonData文件格式是HDFS中的列式存储格式。该格式具有许多新型列存储文件的特性，例如，分割表和数据压缩。CarbonData具有以下独有的特点：
  - 伴随索引的数据存储：由于在查询中设置了过滤器，可以显著加快查询性能，减少I/O扫描次数和CPU资源占用。CarbonData索引由多个级别的索引组成，处理框架可以利用这个索引来减少需要安排和处理的任务，也可以通过在任务扫描中以更精细的单元（称为blocklet）进行skip扫描来代替对整个文件的扫描。
  - 可选择的数据编码：通过支持高效的数据压缩，可基于压缩/编码数据进行查询，在将结果返回给用户之前，才将编码转化为实际数据，这被称为“延迟物化”。
  - 支持一种数据格式应用于多种用例场景：例如，交互式OLAP-style查询，顺序访问（big scan），随机访问（narrow scan）。

CarbonData 关键技术和优势

- 快速查询响应：高性能查询是CarbonData关键技术的优势之一。CarbonData查询速度大约是Spark SQL查询的10倍。CarbonData使用的专用数据格式围绕高性能查询进行设计，其中包括多种索引技术和多次的Push down优化，从而对TB级数据查询进行最快响应。
- 高效率数据压缩：CarbonData使用轻量级压缩和重量级压缩的组合压缩算法压缩数据，可以减少60%~80%数据存储空间，很大程度上节省硬件存储成本。

12.3.1.2 CarbonData 主要规格

CarbonData 主要规格

表 12-3 CarbonData 主要规格

实体	测试值	测试环境
表数	10000	3个节点，每个executor 4个CPU核，20GB。Driver内存5GB，3个Executor。 总列数：107 String：75 Int：13 BigInt：7 Timestamp：6 Double：6
表的列数	2000	3个节点，每个executor4个CPU核，20GB。Driver内存5GB，3个Executor。
原始CSV文件大小的最大值	200GB	17个cluster节点，每个executor 150GB，25个CPU核。Driver内存10 GB，17个Executor。

实体	测试值	测试环境
每个文件夹的CSV文件数	100个文件夹，每个文件夹10个文件，每个文件大小50MB。	3个节点，每个executor4个CPU核，20GB。Driver内存5GB，3个Executor。
加载文件夹数	10000	3个节点，每个executor4个CPU核，20GB。Driver内存5GB，3个Executor。

数据加载所需的内存取决于以下因素：

- 列数
- 列值大小
- 并发（使用“carbon.number.of.cores.while.loading”进行配置）
- 在内存中排序的大小（使用“carbon.sort.size”进行配置）
- 中间缓存（使用“carbon.graph.rowset.size”进行配置）

加载包含1000万条记录和300列的8 GB CSV文件的数据，每行大小约为0.8KB的8GB CSV文件的数据，需要约为10GB的executor执行内存，也就是说，“carbon.sort.size” 配置为“100000”，所有其他前面的配置保留默认值。

二级索引表规格

表 12-4 二级索引表规格

实体	测试值
二级索引表数量	10
二级索引表中的组合列的列数	5
二级索引表中的列名长度（单位：字符）	120
二级索引表名长度（单位：字符）	120
表中所有二级索引表的表名+列名的累积长度*（单位：字符）	3800**

说明

- * Hive允许的上限值或可用资源的上限值。
- ** 二级索引表使用hive注册，并以json格式的值存储在HiveSERDEPROPERTIES中。由hive支持的SERDEPROPERTIES的最大字符数为4000个字符，无法更改。

12.3.2 配置参考

本章节介绍CarbonData所有配置的详细信息。

表 12-5 carbon.properties 中的系统配置

参数	默认值	描述
carbon.ddl.base.hdfs.url	hdfs://hacluster/opt/data	此属性用于从HDFS基本路径配置HDFS相对路径，在“fs.defaultFS”中进行配置。在“carbon.ddl.base.hdfs.url”中配置的路径将被追加到在“fs.defaultFS”中配置的HDFS路径中。如果配置了这个路径，则用户不需要通过完整路径加载数据。 例如：如果CSV文件的绝对路径是“hdfs://10.18.101.155:54310/data/cnbc/2016/xyz.csv”，其中，路径“hdfs://10.18.101.155:54310”来源于属性“fs.defaultFS”并且用户可以把“/data/cnbc/”作为“carbon.ddl.base.hdfs.url”配置。 当前，在数据加载时，用户可以指定CSV文件为“/2016/xyz.csv”。
carbon.badRecords.location	-	指定Bad records的存储路径。此路径为HDFS路径。默认值为Null。如果启用了bad records日志记录或者bad records操作重定向，则该路径必须由用户进行配置。
carbon.bad.records.action	fail	以下是bad records的四种行为类型： FORCE：通过将bad records存储为NULL来自动更正数据。 REDIRECT：Bad records被写入原始CSV文件而不是被加载。 IGNORE：Bad records既不被加载也不被写入原始CSV文件。 FAIL：如果找到任何bad records，则数据加载失败。
carbon.update.sync.folder	/tmp/carbondata	modifiedTime.mdt文件路径，可以设置为已有路径或新路径。 **说明** 如果设置为已有路径，需确保所有用户都可以访问该路径，且该路径具有777权限。

表 12-6 carbon.properties 中的性能配置

参数	默认值	描述
数据加载配置		

参数	默认值	描述
carbon.sort.file.write.buffer.size	16384	为了限制内存的使用，CarbonData会将数据排序并写入临时文件中。该参数控制读取和写入临时文件过程使用的缓存大小。单位：字节。 取值范围为：10240~10485760。
carbon.graph.rowset.size	100000	数据加载图步骤之间交换的行集大小。 最小值=500，最大值=1000000
carbon.number.of.cores.while.loading	6	数据加载时所使用的核数。配置的核数越大压缩性能越好。如果CPU资源充足可以增加此值。
carbon.sort.size	500000	内存排序的数据大小。
carbon.enableXXHash	true	用于hashkey计算的hashmap算法。
carbon.number.of.cores.block.sort	7	数据加载时块排序所使用的核数。
carbon.max.driver.lru.cache.size	-1	在driver端加载数据所达到的最大LRU缓存大小。以MB为单位，默认值为-1，表示缓存没有内存限制。只允许使用大于0的整数值。
carbon.max.executor.lru.cache.size	-1	在executor端加载数据所达到的最大LRU缓存大小。以MB为单位，默认值为-1，表示缓存没有内存限制。只允许使用大于0的整数值。如果未配置该参数，则将考虑参数“carbon.max.driver.lru.cache.size”的值。
carbon.merge.sort.prefetch	true	在数据加载过程中，从排序的临时文件中读取数据进行合并排序时，启用数据预取。
carbon.update.persist.enable	true	启用此参数将考虑持久化数据，减少UPDATE操作的执行时间。
enable.unsafe.sort	true	指定在数据加载期间是否使用非安全排序。非安全的排序减少了数据加载操作期间的垃圾回收（GC），从而提高了性能。默认值为“true”，表示启用非安全排序功能。
enable.offheap.sort	true	在数据加载期间启用堆排序。
offheap.sort.chunk.size.inmb	64	指定需要用于排序的数据块的大小。最小值为1MB，最大值为1024MB。

参数	默认值	描述
carbon.unsafe.working.memory.in.mb	512	指定非安全工作内存的大小。这将用于排序数据，存储列页面等。单位是MB。 数据加载所需内存： (“carbon.number.of.cores.while.loading”的值[默认值 = 6]) x 并行加载数据的表格 x (“offheap.sort.chunk.size.inmb”的值[默认值 = 64 MB] + “carbon.blockletgroup.size.in.mb”的值[默认值 = 64 MB] + 当前的压缩率[64 MB/3.5]) = ~900 MB 每表格 数据查询所需内存： (SPARK_EXECUTOR_INSTANCES. [默认值 = 2]) x (carbon.blockletgroup.size.in.mb [默认值 = 64 MB] +“carbon.blockletgroup.size.in.mb”解压内容[默认值 = 64 MB * 3.5]) x (每个执行器核数[默认值 = 1]) = ~ 600 MB
carbon.sort.inmemory.storage.size.in.mb	512	指定要存储在内存中的中间排序数据的大小。达到该指定的值，系统会将数据写入磁盘。单位是MB。
sort.inmemory.size.inmb	1024	指定要保存在内存中的中间排序数据的大小。达到该指定值后，系统会将数据写入磁盘。单位：MB。 如果配置了“carbon.unsafe.working.memory.in.mb”和“carbon.sort.inmemory.storage.size.in.mb”，则不需要配置该参数。如果此时也配置了该参数，那么这个内存的20%将用于工作内存“carbon.unsafe.working.memory.in.mb”，80%将用于排序存储内存“carbon.sort.inmemory.storage.size.in.mb”。 **说明** Spark配置参数“spark.yarn.executor.memoryOverhead”的值应该大于CarbonData配置参数“sort.inmemory.size.inmb”的值，否则如果堆外（off heap）访问超出配置的executor内存，则YARN可能会停止executor。
carbon.blockletgroup.size.in.mb	64	数据作为blocklet group被系统读入。该参数指定blocklet group的大小。较高的值会有更好的顺序IO访问性能。 最小值为16MB，任何小于16MB的值都将重置为默认值（64MB）。 单位：MB。
enable.inmemory.merge.sort	false	指定是否启用内存合并排序（inmemorymerge sort）。

参数	默认值	描述
use.offheap.in.query.processing	true	指定是否在查询处理中启用offheap。
carbon.load.sort.scope	local_sort	指定加载操作的排序范围。支持两种类型的排序，batch_sort和local_sort。选择batch_sort将提升加载性能，但会降低查询性能。
carbon.batch.sort.size.inmb	-	指定在数据加载期间为批处理排序而考虑的数据大小。推荐值为小于总排序数据的45%。该值以MB为单位。 **说明** 如果没有设置参数值，那么默认情况下其大约等于“sort.inmemory.size.inmb”参数值的45%。
enable.unsafe.columnpage	true	指定在数据加载或查询期间，是否将页面数据保留在堆内存中，以避免垃圾回收受阻。
carbon.use.local.dir	false	是否使用YARN本地目录加载多个磁盘的数据。设置为true，则使用YARN本地目录加载多个磁盘的数据，以提高数据加载性能。
carbon.use.multiple.temp.dir	false	是否使用多个临时目录存储临时文件以提高数据加载性能。
carbon.load.datamaps.parallel.db_name.table_name	NA	值为true或者false。可以设置数据库名和表名，使得该表的首次查询性能得到提升。
压缩配置		
carbon.number.of.cores.while.compacting	2	在压缩过程中用于写入数据所使用的核数。配置的核数越大压缩性能越好。如果CPU资源充足可以增加此值。
carbon.compaction.level.threshold	4,3	该属性用于Minor压缩，决定合并segment的数量。 例如：如果被设置为“2,3”，则将每2个segment触发一次Minor压缩。“3”是Level 1压缩的segment个数，这些segment将进一步被压缩为新的segment。 有效值为0-100。
carbon.major.compaction.size	1024	使用该参数配置Major压缩的大小。总数低于该阈值的segment将被合并。 单位为MB。

参数	默认值	描述
carbon.horizontal.compaction.enable	true	该参数用于配置打开/关闭水平压缩。在每个DELETE和UPDATE语句之后，如果增量（DELETE / UPDATE）文件超过指定的阈值，则可能发生水平压缩。默认情况下，该参数值设置为“true”，打开水平压缩功能，可将参数值设置为“false”来关闭水平压缩功能。
carbon.horizontal.update.compaction.threshold	1	该参数指定segment内的UPDATE增量文件数的阈值限制。在增量文件数量超过阈值的情况下，segment内的UPDATE增量文件变得适合水平压缩，并压缩为单个UPDATE增量文件。默认情况下，该参数值设置为1。可以设置为1到10000之间的值。
carbon.horizontal.delete.compaction.threshold	1	该参数指定segment的block中的DELETE增量文件数量的阈值限制。在增量文件数量超过阈值的情况下，segment特定block的DELETE增量文件变得适合水平压缩，并压缩为单个DELETE增量文件。 默认情况下，该参数值设置为1。可以设置为1到10000之间的值。
查询配置		
carbon.number.of.cores	4	查询时所使用的核数。
carbon.limit.block.distribution.enable	false	当查询语句中包含关键字limit时，启用或禁用CarbonData块分布。默认值为“false”，将对包含关键字limit的查询语句禁用块分布。此参数调优请参考**性能调优的相关配置**。
carbon.custom.block.distribution	false	指定是使用Spark还是CarbonData的块分配功能。默认情况下，其配置值为“false”，表明启用Spark块分配。若要使用CarbonData块分配，请将配置值更改为“true”。
carbon.infilter.subquery.pushdown.enable	false	如果启用此参数，并且用户在具有subquery的过滤器中触发Select查询，则执行子查询，并将输出作为IN过滤器广播到左表，否则将执行SortMergeSemiJoin。建议在IN过滤器子查询未返回太多记录时启用此参数。例如，IN子句子查询返回10k或更少的记录时，启用此参数将更快地给出查询结果。 示例：***select * from flow_carbon_256b where cus_no in (select cus_no from flow_carbon_256b where dt>='20260101' and dt<='20260701' and txn_bk='tk_1' and txn_br='tr_1') limit 1000;***
carbon.scheduler.minRegisteredResourcesRatio	0.8	启动块分布所需的最小资源（executor）比率。默认值为“0.8”，表示所请求资源的80%被分配用于启动块分布。
carbon.dynamicAllocation.schedulerTimeout	5	此参数值指示调度器等待executors处于活动状态的最长时间。默认值为“5”秒，允许的最大值为“15”秒。

参数	默认值	描述
enable.unsafe.in.query.processing	true	指定在查询操作期间是否使用非安全排序。非安全排序减少查询操作期间的垃圾回收（GC），从而提高性能。默认值为“true”，表示启用非安全排序功能。
carbon.enable.vector.reader	true	为结果收集（result collection）启用向量处理，以增强查询性能。
carbon.query.show.datamaps	true	**SHOW TABLES**会展示所有的表包含主表和datamap。如果需要过滤掉datamap，将该配置设置为false。
二级索引配置		
carbon.secondary.index.creation.threads	1	该参数用于配置启动二级索引创建期间并行处理segments的线程数。当表的segments数较多时，该参数有助于微调系统生成二级索引的速度。该参数值范围为1到50。
carbon.si.lookup.partialstring	true	• 当配置为true时，它包括开始，结尾和包含。 • 当配置为false时，它只包括从二级索引开始。
carbon.si.segment.merge	true	开启这个配置后会合并二级索引表segment内的**.carbondata**文件。合并发生在导入操作后，在二级索引表导入操作的最后，会检查小文件并合并他们。 **说明** Table Block Size会用作合并小文件的大小阈值。

表 12-7 carbon.properties 中的其它配置

参数	默认值	描述
数据加载配置		
carbon.lock.type	HDFSLOCK	该配置指定了表上并发操作过程中所要求的锁的类型。 有以下几种类型锁实现方式： • LOCALLOCK：基于本地文件系统的文件来创建的锁。该锁只适用于一台机器上只运行一个Spark Driver（或者JDBCServer）的情况。 • HDFSLOCK：基于HDFS文件系统上的文件来创建的锁。该锁适用于集群上有多个运行的Spark应用而且没有可用的ZooKeeper的情况。
carbon.sort.intermediate.files.limit	20	中间文件的最小数量。生成中间文件后开始排序合并。此参数调优请参考**性能调优的相关配置**。

参数	默认值	描述
carbon.csv.read.buffersize.byte	1048576	CSV读缓冲区大小。
carbon.merge.sort.reader.thread	3	用于读取中间文件进行最终合并的最大线程数。
carbon.concurrent.lock.retries	100	指定获取并发操作锁的最大重试次数。该参数用于并发加载。
carbon.concurrent.lock.retry.timeout.sec	1	指定获取并发操作的锁重试之间的间隔。
carbon.lock.retries	3	指定除导入操作外其他所有操作尝试获取锁的次数。
carbon.lock.retry.timeout.sec	5	指定除导入操作外其他所有操作尝试获取锁的时间间隔。
carbon.tempstore.location	/opt/Carbon/TempStoreLoc	临时存储位置。默认情况下，采用“System.getProperty("java.io.tmpdir")”方法获取。此参数调优请参考**性能调优的相关配置**中关于“carbon.use.local.dir”的描述。
carbon.load.log.counter	500000	数据加载记录计数日志。
SERIALIZATION_NULL_FORMAT	\N	指定需要替换为NULL的值。
carbon.skip.empty.line	false	设置此属性将在数据加载期间忽略CSV文件中的空行。
carbon.load.datamaps.parallel	false	该配置项将会开启对所有会话所有表的datamap并行加载。该配置项通过将导入datamap到内存的工作分发给所有的executor来缩短时间，进而提升查询性能。
合并配置		
carbon.numberof.preserve.segments	0	若用户希望从被合并的segment中保留一定数量的segment，可设置该属性参数。 例如：“carbon.numberof.preserve.segments”=“2”，那么合并的segement中将不包含最新的2个segment。 默认保留No segment的状态。

参数	默认值	描述
carbon.allowed.compaction.days	0	合并将合并在配置的指定天数中加载的segment。 例如：如果配置值为“2”，那么只有在2天时间框架中加载的segment被合并。2天以外被加载的segment不会被合并。 该参数默认为禁用。
carbon.enable.auto.load.merge	false	在数据加载时启用压缩。
carbon.merge.index.in.segment	true	如果设置，则Segment内的所有Carbon索引文件（.carbonindex）将合并为单个Carbon索引合并文件（.carbonindexmerge）。 这增强了首次查询性能
查询配置		
max.query.execution.time	60	单次查询允许的最大时间。 单位为分钟。
carbon.enableMinMax	true	MinMax用于提高查询性能。设置为false可禁用该功能。
carbon.lease.recovery.retry.count	5	需要为恢复文件租约所需的最大尝试次数。 最小值：1 最大值：50
carbon.lease.recovery.retry.interval	1000（ms）	尝试在文件上进行租约恢复之后的间隔（Interval）或暂停（Pause）时间。 最小值：1000（ms） 最大值：10000（ms）

表 12-8 spark-defaults.conf 中的 Spark 配置参考

参数	默认值	描述
spark.driver.memory	4G	指定用于driver端进程的内存，其中SparkContext已初始化。 **说明** 在客户端模式下，不要使用SparkConf在应用程序中设置该参数，因为驱动程序JVM已经启动。要配置该参数，请在--driver-memory命令行选项或默认属性文件中进行配置。
spark.executor.memory	4GB	指定每个执行程序进程使用的内存。

参数	默认值	描述
spark.sql.crossJoin.enabled	true	如果查询包含交叉连接，请启用此属性，以便不会抛出错误，此时使用交叉连接而不是连接，可实现更好的性能。

在Spark Driver端的“spark-defaults.conf”文件中配置以下参数。

- 在spark-sql模式下配置：

  表 12-9 spark-sql 模式下的配置参数

参数	配置值	描述
spark.driver.extraJavaOptions	-Dlog4j.configuration=file:/opt/client/Spark2x/spark/conf/log4j.properties -Djetty.version=x.y.z -Dzookeeper.server.principal=zookeeper/hadoop.*<系统域名>* -Djava.security.krb5.conf=/opt/client/KrbClient/kerberos/var/krb5kdc/krb5.conf -Djava.security.auth.login.config=/opt/client/Spark2x/spark/conf/jaas.conf -Dorg.xerial.snappy.tempdir=/opt/client/Spark2x/tmp -Dcarbon.properties.filepath=/opt/client/Spark2x/spark/conf/carbon.properties -Djava.io.tmpdir=/opt/client/Spark2x/tmp	默认值中“/opt/client/Spark2x/spark”为客户端的CLIENT_HOME，且该默认值是追加到参数“spark.driver.extraJavaOptions”其他值之后的，此参数用于指定Driver端的“carbon.properties”文件路径。 **说明** 请注意“=”两边不要有空格。
spark.sql.session.state.builder	org.apache.spark.sql.hive.FIHiveACLSessionStateBuilder	指定会话状态构造器。
spark.carbon.sqlastbuilder.classname	org.apache.spark.sql.hive.CarbonInternalSqlAstBuilder	指定AST构造器。
spark.sql.catalog.class	org.apache.spark.sql.hive.HiveACLExternalCatalog	指定Hive的外部目录实现。启用Spark ACL时必须提供。
spark.sql.hive.implementation	org.apache.spark.sql.hive.HiveACLClientImpl	指定Hive客户端调用的实现。启用Spark ACL时必须提供。

参数	配置值	描述
spark.sql.hiveClient.isolation.enabled	false	启用Spark ACL时必须提供。

- 在JDBCServer服务中配置：

表 12-10 JDBCServer 服务中的配置参数

参数	配置值	描述
spark.driver.extraJavaOptions	-Xloggc:${SPARK_LOG_DIR}/indexserver-omm-%p-gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:MaxDirectMemorySize=512M -XX:MaxMetaspaceSize=512M -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=10M -XX:OnOutOfMemoryError='kill -9 %p' -Djetty.version=x.y.z -Dorg.xerial.snappy.tempdir=${BIGDATA_HOME}/tmp/spark2x/JDBCServer/snappy_tmp -Djava.io.tmpdir=${BIGDATA_HOME}/tmp/spark2x/JDBCServer/io_tmp -Dcarbon.properties.filepath=${SPARK_CONF_DIR}/carbon.properties -Djdk.tls.ephemeralDHKeySize=20	默认值中${SPARK_CONF_DIR}需视具体的集群而定，且该默认值是追加到参数“spark.driver.extraJavaOptions”其他值之后的，此参数用于指定Driver端的“carbon.properties”文件路径。 **说明** 请注意“=”两边不要有空格。

参数	配置值	描述
	48 -Dspark.ssl.keyStore=${SPARK_CONF_DIR}/child.keystore#{java_stack_prefer}	
spark.sql.session.state.builder	org.apache.spark.sql.hive.FIHiveACLSessionStateBuilder	指定会话状态构造器。
spark.carbon.sqlastbuilder.classname	org.apache.spark.sql.hive.CarbonInternalSqlAstBuilder	指定AST构造器。
spark.sql.catalog.class	org.apache.spark.sql.hive.HiveACLExternalCatalog	指定Hive的外部目录实现。启用Spark ACL时必须提供。
spark.sql.hive.implementation	org.apache.spark.sql.hive.HiveACLClientImpl	指定Hive客户端调用的实现。启用Spark ACL时必须提供。
spark.sql.hiveClient.isolation.enabled	false	启用Spark ACL时必须提供。

12.3.3 CarbonData 操作指导

12.3.3.1 CarbonData 快速入门

本章节介绍创建CarbonData table、加载数据，以及查询数据的快速入门流程。该快速入门提供基于Spark Beeline客户端的操作。如果使用Spark shell，需将查询命令写在spark.sql()的括号中。

本操作以从CSV文件加载数据到CarbonData Table为例

表 12-11 CarbonData 快速入门

操作	说明
准备CSV文件	准备加载到CarbonData Table的CSV文件。
连接到CarbonData	在对CarbonData进行任何一种操作之前，首先需要连接到CarbonData。

操作	说明
创建CarbonData Table	连接到CarbonData之后，需要创建CarbonData table用于加载数据和执行查询操作。
加载数据到CarbonData Table	创建CarbonData table之后，可以从CSV文件加载数据到所创建的table中。
在CarbonData中查询数据	创建CarbonData table并加载数据之后，可以执行所需的查询操作，例如filters，groupby等。

准备 CSV 文件

1. 在本地准备CSV文件，文件名为：test.csv，样例如下：

```
13418592122,1001,MAC地址,2017-10-23 15:32:30,2017-10-24 15:32:30,62.50,74.56
13418592123,1002,MAC地址,2017-10-23 16:32:30,2017-10-24 16:32:30,17.80,76.28
13418592124,1003,MAC地址,2017-10-23 17:32:30,2017-10-24 17:32:30,20.40,92.94
13418592125,1004,MAC地址,2017-10-23 18:32:30,2017-10-24 18:32:30,73.84,8.58
13418592126,1005,MAC地址,2017-10-23 19:32:30,2017-10-24 19:32:30,80.50,88.02
13418592127,1006,MAC地址,2017-10-23 20:32:30,2017-10-24 20:32:30,65.77,71.24
13418592128,1007,MAC地址,2017-10-23 21:32:30,2017-10-24 21:32:30,75.21,76.04
13418592129,1008,MAC地址,2017-10-23 22:32:30,2017-10-24 22:32:30,63.30,94.40
13418592130,1009,MAC地址,2017-10-23 23:32:30,2017-10-24 23:32:30,95.51,50.17
13418592131,1010,MAC地址,2017-10-24 00:32:30,2017-10-25 00:32:30,39.62,99.13
```

2. 使用WinSCP工具将CSV文件导入客户端节点，例如“/opt”目录下。
3. 登录FusionInsight Manager页面，选择“系统 > 权限 > 用户”，添加人机用户sparkuser，用户组（hadoop、hive），主组（hadoop）。
4. 进入客户端目录，加载环境变量并认证用户：

   cd /*客户端安装目录*

   source ./bigdata_env

   source ./Spark2x/component_env

   kinit sparkuser
5. 上传CSV中的文件到HDFS的“/data”目录：

   hdfs dfs -put /opt/test.csv /data/

连接到 CarbonData

- 使用Spark SQL或Spark shell连接到Spark并执行Spark SQL命令。
- 开启JDBCServer并使用JDBC客户端（例如，Spark Beeline）连接。

  执行如下命令：

  cd ./Spark2x/spark/bin

  ./spark-beeline

创建 CarbonData Table

在Spark Beeline被连接到JDBCServer之后，需要创建一个CarbonData table用于加载数据和执行查询操作。下面是创建一个简单的表的命令。

create table x1 (imei string, deviceInformationId int, mac string, productdate timestamp, updatetime timestamp, gamePointId double, contractNumber

double) STORED AS carbondata TBLPROPERTIES ('SORT_COLUMNS'='imei,mac');

命令执行结果如下：

```
+---------+
| Result  |
+---------+
+---------+
No rows selected (1.093 seconds)
```

加载数据到 CarbonData Table

创建CarbonData table之后，可以从CSV文件加载数据到所创建的表中。

用所要求的参数运行以下命令从CSV文件加载数据。该表的列名需要与CSV文件的列名匹配。

LOAD DATA inpath 'hdfs://hacluster/data/*test.csv*' into table *x1* options('DELIMITER'=',', 'QUOTECHAR'='"','FILEHEADER'='imei, deviceinformationid,mac, productdate,updatetime, gamepointid,contractnumber');

其中，“test.csv”为**准备CSV文件**的CSV文件，“x1”为示例的表名。

CSV样例内容如下：

```
13418592122,1001,MAC地址,2017-10-23 15:32:30,2017-10-24 15:32:30,62.50,74.56
13418592123,1002,MAC地址,2017-10-23 16:32:30,2017-10-24 16:32:30,17.80,76.28
13418592124,1003,MAC地址,2017-10-23 17:32:30,2017-10-24 17:32:30,20.40,92.94
13418592125,1004,MAC地址,2017-10-23 18:32:30,2017-10-24 18:32:30,73.84,8.58
13418592126,1005,MAC地址,2017-10-23 19:32:30,2017-10-24 19:32:30,80.50,88.02
13418592127,1006,MAC地址,2017-10-23 20:32:30,2017-10-24 20:32:30,65.77,71.24
13418592128,1007,MAC地址,2017-10-23 21:32:30,2017-10-24 21:32:30,75.21,76.04
13418592129,1008,MAC地址,2017-10-23 22:32:30,2017-10-24 22:32:30,63.30,94.40
13418592130,1009,MAC地址,2017-10-23 23:32:30,2017-10-24 23:32:30,95.51,50.17
13418592131,1010,MAC地址,2017-10-24 00:32:30,2017-10-25 00:32:30,39.62,99.13
```

命令执行结果如下：

```
+-------------+
|Segment ID  |
+-------------+
|0           |
+-------------+
No rows selected (3.039 seconds)
```

在 CarbonData 中查询数据

创建CarbonData table并加载数据之后，可以执行所需的数据查询操作。以下为一些查询操作举例。

- **获取记录数**

  为了获取在CarbonData table中的记录数，可以运行以下命令。

  select count(*) from x1;

- **使用Groupby查询**

  为了获取不重复的deviceinformationid记录数，可以运行以下命令。

  select deviceinformationid,count (distinct deviceinformationid) from x1 group by deviceinformationid;

- **用Filter查询**

  为了获取特定deviceinformationid的记录，可以运行以下命令。

  select * from x1 where deviceinformationid='1010';

在 Spark-shell 上使用 CarbonData

用户若需要在Spark-shell上使用CarbonData，需通过如下方式创建CarbonData Table，加载数据到CarbonData Table和在CarbonData中查询数据的操作。

```
spark.sql("CREATE TABLE x2(imei string, deviceInformationId int, mac string, productdate timestamp,
updatetime timestamp, gamePointId double, contractNumber double) STORED AS carbondata")
spark.sql("LOAD DATA inpath 'hdfs://hacluster/data/x1_without_header.csv' into table x2
options('DELIMITER'=',', 'QUOTECHAR'='\"','FILEHEADER'='imei, deviceinformationid,mac,
productdate,updatetime, gamepointid,contractnumber')")
spark.sql("SELECT * FROM x2").show()
```

12.3.3.2 管理 CarbonData Table

12.3.3.2.1 CarbonData Table 简介

简介

CarbonData中的数据存储在table实体中。CarbonData table与RDBMS中的表类似。RDBMS数据存储在由行和列构成的表中。CarbonData table存储的也是结构化的数据，拥有固定列和数据类型。

支持数据类型

CarbonData支持以下数据类型：

- Int
- String
- BigInt
- Smallint
- Char
- Varchar
- Boolean
- Decimal
- Double
- TimeStamp
- Date
- Array
- Struct
- Map

下表对所支持的数据类型及其各自的范围进行了详细说明。

表 12-12 CarbonData 数据类型

数据类型	范围
Int	4字节有符号整数，从-2,147,483,648到2,147,483,647 **说明** 非字典列如果是Int类型，会在内部存储为BigInt类型。
String	100000字符 **说明** 如果在**CREATE TABLE**中使用Char或Varchar数据类型，则这两种数据类型将自动转换为String数据类型。 如果存在字符长度超过32000的列，需要在建表时，将该列加入到tblproperties的LONG_STRING_COLUMNS属性里。
BigInt	64-bit，从-9,223,372,036,854,775,808到9,223,372,036,854,775,807
SmallInt	范围-32,768到32,767
Char	范围A到Z&a到z
Varchar	范围A到Z&a到z&0到9
Boolean	范围true或者false
Decimal	默认值是(10,0) ，最大值是(38,38) **说明** 当进行带过滤条件的查询时，为了得到准确的结果，需要在数字后面加上BD。例如，*select * from carbon_table where num = 1234567890123456.22BD*.
Double	64-bit，从4.9E-324到1.7976931348623157E308
TimeStamp	NA，默认格式为“yyyy-MM-dd HH:mm:ss”。
Date	DATE数据类型用于存储日历日期。 默认格式为“yyyy-MM-dd”。
Array<data_type>	NA **说明** 现仅支持2层复杂类型的嵌套。
Struct<col_name: data_type COMMENT col_comment, ...>	
Map<primitive_type, data_type>	

12.3.3.2.2 新建 CarbonData Table

操作场景

使用CarbonData前需先创建表，才可在其中加载数据和查询数据。可通过**Create Table**命令来创建表。该命令支持使用自定义列创建表。

使用自定义列创建表

可通过指定各列及其数据类型来创建表。

命令示例：

CREATE TABLE *IF NOT EXISTS productdb.productSalesTable (*

productNumber Int,

productName String,

storeCity String,

storeProvince String,

productCategory String,

productBatch String,

saleQuantity Int,

revenue Int)

STORED AS *carbondata*

TBLPROPERTIES (

'table_blocksize'='128');

上述命令所创建的表的详细信息如下：

表 12-13 表信息定义

参数	描述
productSalesTable	待创建的表的名称。该表用于加载数据进行分析。 表名由字母、数字、下划线组成。
productdb	数据库名称。该数据库将与其中的表保持逻辑连接以便于识别和管理。 数据库名称由字母、数字、下划线组成。
productName storeCity storeProvince procuctCategory productBatch saleQuantity revenue	表中的列，代表执行分析所需的业务实体。 列名（字段名）由字母、数字、下划线组成。

参数	描述
table_blocksize	CarbonData表使用的数据文件的block大小，默认值为1024，最小值为1，最大值为2048，单位为MB。 如果“table_blocksize”值太小，数据加载时，生成过多的小数据文件，可能会影响HDFS的使用性能。 如果“table_blocksize”值太大，数据查询时，索引匹配的block数据量较大，某些block会包含较多的blocklet，导致读取并发度不高，从而降低查询性能。 一般情况下，建议根据数据量级别来选择大小。例如：GB级别用256，TB级别用512，PB级别用1024。

说明

- 所有Integer类型度量均以BigInt类型进行处理与显示。
- CarbonData遵循严格解析，因此任何不可解析的数据都会被保存为null。例如，在BigInt列中加载double值（3.14），将会保存为null。
- 在Create Table中使用的Short和Long数据类型在DESCRIBE命令中分别显示为Smallint和Bigint。
- 可以使用DESCRIBE格式化命令查看表数据大小和表索引大小。

操作结果

根据命令创建表。

12.3.3.2.3 删除 CarbonData Table

操作场景

可使用**DROP TABLE**命令删除表。删除表后，所有metadata以及表中已加载的数据都会被删除。

操作步骤

运行如下命令删除表。

命令：

DROP TABLE *[IF EXISTS] [db_name.]table_name;*

一旦执行该命令，将会从系统中删除表。命令中的“db_name”为可选参数。如果没有指定“db_name”，那么将会删除当前数据库下名为“table_name”的表。

示例：

DROP TABLE *productdb.productSalesTable;*

通过上述命令，删除数据库“productdb”下的表“productSalesTable”。

操作结果

从系统中删除命令中指定的表。删除完成后，可通过**SHOW TABLES**命令进行查询，确认所需删除的表是否成功被删除，详见**SHOW TABLES**。

12.3.3.2.4 修改 CarbonData Table

SET 和 UNSET

当使用set命令时，所有新set的属性将会覆盖已存在的旧的属性。

- SORT SCOPE

  SET SORT SCOPE命令示例：

  ALTER TABLE *tablename* **SET TBLPROPERTIES('SORT_SCOPE'**='*no_sort*')

  当UNSET SORT SCOPE后，会使用默认值NO_SORT。

  UNSET SORT SCOPE命令示例：

  ALTER TABLE *tablename* **UNSET TBLPROPERTIES('SORT_SCOPE'**)

- SORT COLUMNS

  SET SORT COLUMNS命令示例：

  ALTER TABLE *tablename* **SET TBLPROPERTIES('SORT_COLUMNS'**='*column1*')

  在执行该命令后，新的导入会使用新的SORT_COLUMNS配置值。 用户可以根据查询的情况来调整SORT_COLUMNS，但是不会直接影响旧的数据。所以对历史的segments的查询性能不会受到影响，因为历史的segments不是按照新的SORT_COLUMNS。

  不支持UNSET命令，但是可以使用set SORT_COLUMNS等于空字符串来代替UNSET命令。

  ALTER TABLE *tablename* **SET TBLPROPERTIES('SORT_COLUMNS'**='')

  说明

  - 后续版本会加强自定义合并来对旧的segment重新排序。
  - 流式表不支持修改SORT_COLUMNS。
  - 如果inverted index的列从SORT_COLUMNS里面移除了，该列不会再创建inverted index。但是旧的INVERTED_INDEX配置值不会变化。

12.3.3.3 管理 CarbonData Table 数据

12.3.3.3.1 加载数据

操作场景

CarbonData table创建成功后，可使用**LOAD DATA**命令在表中加载数据，并可供查询。触发数据加载后，数据以CarbonData格式进行编码，并将多维列式存储格式文件压缩后复制到存储CarbonData文件的HDFS路径下供快速分析查询使用。HDFS路径可以配置在carbon.properties文件中。具体请参考**配置参考**。

12.3.3.3.2 删除 Segments

操作场景

如果用户将错误数据加载到表中，或者数据加载后出现许多错误记录，用户希望修改并重新加载数据时，可删除对应的segment。可使用segment ID来删除segment，也可以使用加载数据的时间来删除segment。

说明

删除segment操作只能删除未合并的segment，已合并的segment可以通过**CLEAN FILES**命令清除segment。

通过 Segment ID 删除

每个Segment都有与其关联的唯一Segment ID。使用这个Segment ID可以删除该Segment。

步骤1 运行如下命令获取Segment ID。

命令：

SHOW SEGMENTS FOR Table *dbname.tablename LIMIT number_of_loads;*

示例：

SHOW SEGMENTS FOR TABLE *carbonTable;*

上述命令可显示tablename为carbonTable的表的所有Segment信息。

SHOW SEGMENTS FOR TABLE *carbonTable LIMIT 2;*

上述命令可显示*number_of_loads*规定条数的Segment信息。

输出结果如下：

```
+-----+----------+--------------------------+------------------+------------+------------+-------------+--------------+--
+
| ID  |  Status  |     Load Start Time      | Load Time Taken  | Partition  | Data Size  | Index Size  | File Format  |
+-----+----------+--------------------------+------------------+------------+------------+-------------+--------------+--
+
| 3   | Success  | 2020-09-28 22:53:26.336  | 3.726S           | {}         | 6.47KB     | 3.30KB      | columnar_v3  |
| 2   | Success  | 2020-09-28 22:53:01.702  | 6.688S           | {}         | 6.47KB     | 3.30KB      | columnar_v3  |
+-----+----------+--------------------------+------------------+------------+------------+-------------+--------------+--
+
```

说明

SHOW SEGMENTS命令输出包括ID、Status、Load Start Time、Load Time Taken、Partition、Data Size、Index Size、File Format。最新的加载信息在输出中第一行显示。

步骤2 获取到需要删除的Segment的Segment ID后，执行如下命令删除对应Segment：

命令：

DELETE FROM TABLE tableName WHERE SEGMENT.ID IN (load_sequence_id1, load_sequence_id2, ...);

示例：

DELETE FROM TABLE carbonTable WHERE SEGMENT.ID IN (1,2,3);

详细信息，请参阅**DELETE SEGMENT by ID**。

----结束

通过加载数据的时间删除

用户可基于特定的加载时间删除数据。

命令：

DELETE FROM TABLE db_name.table_name WHERE SEGMENT.STARTTIME BEFORE date_value;

示例：

DELETE FROM TABLE carbonTable WHERE SEGMENT.STARTTIME BEFORE '2017-07-01 12:07:20';

上述命令可删除'2017-07-01 12:07:20'之前的所有segment。

有关详细信息，请参阅**DELETE SEGMENT by DATE**。

删除结果

数据对应的segment被删除，数据将不能再被访问。可通过**SHOW SEGMENTS**命令显示segment状态，查看是否成功删除。

说明

- 调用**DELETE SEGMENT**命令时，物理上而言，Segment并没有从文件系统中被删除。使用命令**SHOW SEGMENTS**查看Segment信息，可看见被删除的Segment的状态被标识为"Marked for Delete"。但使用**SELECT * FROM tablename**命令查询时，不会显示被删除的Segment的内容。
- 下一次加载数据且达到最大查询执行时间（由“max.query.execution.time”配置，默认为“60分钟”）后，Segment才会从文件系统中真正删除。
- 如果用户想要强制删除物理Segment文件，那么可以使用**CLEAN FILES**命令。

  示例：

  CLEAN FILES FOR TABLE table1;

  该命令将从物理上删除状态为“Marked for delete”的Segment文件。

  如果在“max.query.execution.time”规定的时间到达之前使用该命令，可能会导致查询失败。“max.query.execution.time”可在“carbon.properties”文件中设置，表示一次查询允许花费的最长时间。

12.3.3.3.3 合并 Segments

操作场景

频繁的数据获取导致在存储目录中产生许多零碎的CarbonData文件。由于数据排序只在每次加载时进行，所以，索引也只在每次加载时执行。这意味着，对于每次加载都会产生一个索引，随着数据加载数量的增加，索引的数量也随之增加。由于每个索引只在一次加载时工作，索引的性能被降低。CarbonData提供加载压缩。压缩过程通过合并排序各segment中的数据，将多个segment合并为一个大的segment。

前提条件

已经加载了多次数据。

操作描述

有Minor合并、Major合并和Custom合并三种类型。

- Minor合并：

  在Minor合并中，用户可指定合并数据加载的数量。如果设置了参数“carbon.enable.auto.load.merge”，每次数据加载都可触发Minor合并。如果任意segment均可合并，那么合并将于数据加载时并行进行。

  Minor合并有两个级别。

  - Level 1：合并未合并的segment。
  - Level 2：合并已合并的segment，以形成更大的segment。

- Major合并：

  在Major合并中，许多segment可以合并为一个大的segment。用户将指定合并尺寸，将对未达到该尺寸的segment进行合并。Major合并通常在非高峰时段进行。

- Custom合并：

  在Custom合并中，用户可以指定几个segment的id合并为一个大的segment。所有指定的segment的id必须存在并且有效，否则合并将会失败。Custom合并通常在非高峰时段进行。

具体的命令操作，请参考**ALTER TABLE COMPACTION**。

表 12-14 合并参数

参数	默认值	应用类型	描述
carbon.enable.auto.load.merge	false	Minor	数据加载时启用合并。 “true”：数据加载时自动触发segment合并。 “false”：数据加载时不触发segment合并。
carbon.compaction.level.threshold	4,3	Minor	对于Minor合并，该属性参数决定合并segment的数量。 例如，如果该参数设置为“2,3”，在Level 1，每2个segment触发一次Minor合并。在Level2，每3个Level 1合并的segment将被再次合并为新的segment。 合并策略根据实际的数据大小和可用资源决定。 有效值为0-100。

参数	默认值	应用类型	描述
carbon.major.compaction.size	1024mb	Major	通过配置该参数可配置Major合并。低于该阈值的segment之和将被合并。 例如，如果该阈值是1024MB，且有5个大小依次为300MB，400MB，500MB，200MB，100MB的segment用于Major合并，那么只有相加的总数小于阈值的segment会被合并，也就是300+400+200+100 = 1000MB的segment会被合并，而500MB的segment将会被跳过。
carbon.numberof.preserve.segments	0	Minor/Major	如果用户希望从被合并的segment中保留一定数量的segment，可通过该属性参数进行设置。 例如，“carbon.numberof.preserve.segments”=“2”，那么最新的2个segment将不会包含在合并中。 默认不保留任何segment。
carbon.allowed.compaction.days	0	Minor/Major	合并将合并在指定的配置天数中加载的segment。 例如，如果配置为“2”，那么只有在2天的时间框架中被加载的segment可以被合并。在2天以外被加载的segment将不被合并。 默认为禁用。
carbon.number.of.cores.while.compacting	2	Minor/Major	在合并过程中写入数据时所用的核数。配置的核数越大合并性能越好。如果CPU资源充足可以增加此值。
carbon.merge.index.in.segment	true	SEGMENT_INDEX	如果设置为true，则一个segment中所有Carbon索引文件（.carbonindex）将合并为单个Carbon索引合并文件（.carbonindexmerge）。 这增强了首次查询性能。

参考信息

建议避免对历史数据进行minor compaction，请参考**如何避免对历史数据进行minor compaction?**

12.3.3.4 迁移 CarbonData 数据

操作场景

如果用户需要快速从一个集群中将CarbonData的数据迁移到另外一个集群的CarbonData中，可以使用CarbonData的数据备份与恢复命令来完成该任务。使用此方法迁移数据，无需在新集群执行数据导入的过程，可以减少迁移的时间。

前提条件

两个集群已安装Spark2x客户端，例如安装目录为“/opt/client”。假设原始数据所在集群为A，需要迁移到集群B。

操作步骤

步骤1 使用客户端安装用户登录A集群客户端所在节点。

步骤2 使用客户端用户执行以下命令，配置环境变量。

source /opt/client/bigdata_env

source /opt/client/Spark2x/component_env

步骤3 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *carbondatauser*

carbondatauser需要为原始数据的使用者，即拥有表的读写权限。

说明

该用户需要加入hadoop，hive组，主组选择hadoop组，并关联角色System_administrator。

步骤4 执行以下命令连接数据库，并查看表的数据在HDFS保存的位置：

spark-beeline

desc formatted *原始数据的表名称*;

查看系统显示的信息中“Location”表示数据文件所在目录。

步骤5 使用客户端安装用户登录B集群客户端所在节点，并执行以下命令，配置环境变量。

source /opt/client/bigdata_env

source /opt/client/Spark2x/component_env

步骤6 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *carbondatauser2*

carbondatauser2需要为上传数据的用户。

说明

该用户需要加入hadoop，hive组，主组选择hadoop组，并关联角色System_administrator。

步骤7 执行**spark-beeline**命令连接数据库。

步骤8 原始数据对应的数据库是否存在?

- 是，执行**步骤9**。
- 否，执行命令**create database** *数据库名称*，创建一个同名的数据库，然后执行**步骤9**。

步骤9 将原始数据从集群A的HDFS目录中，拷贝数据到集群B的HDFS中。

在集群B上传数据时，上传目录中需要存在与原始目录有相同的数据库以及表名目录，且上传用户需要有在此目录写入数据的权限。上传后该用户将拥有数据的读写权限。

例如，原始数据保存在“/user/carboncadauser/warehouse/db1/tb1”，则在新集群中数据可以保存在“/user/carbondatauser2/warehouse/db1/tb1”中：

1. 下载原始数据到集群A的“/opt/backup”目录下：

   hdfs dfs -get /user/carboncadauser/warehouse/db1/tb1 /opt/backup
2. 拷贝集群A的原始数据到集群B客户端节点的“/opt/backup”目录下：

   scp /opt/backup root@*集群B客户端节点IP***:/opt/backup**
3. 上传拷贝到集群B的数据到HDFS中：

   hdfs dfs -put /opt/backup /user/carbondatauser2/warehouse/db1/tb1

步骤10 在集群B的客户端环境执行以下命令，在Hive中生成原始数据对应的表所关联的元数据：

REFRESH TABLE *$dbName.$tbName;*

$dbName和$tbName分别表示数据对应的数据库名称以及表名称。

步骤11 如果原表存在索引表，执行**步骤9**和**步骤10**，将集群A的索引表目录迁移到集群B。

步骤12 执行以下命令，为CarbonData表注册索引表（注意，如果原表没有创建索引表，则不需要执行此步骤）：

REGISTER INDEX TABLE *$tableName* ON *$maintable*;

$tableName和$maintable分别表示索引表名称和主表名称。

----结束

12.3.3.5 迁移 Spark1.5 的 Carbondata 数据到 Spark2x 的 Carbondata 中

迁移方案概览

本次迁移目标是将Spark1.5的CarbonData表数据迁移到Spark2x的CarbonData表中。

说明

执行本操作前需要将spark1.5的carbondata表入库业务中断，将数据一次性迁移至spark2x的carbondata表，完成迁移后使用spark2x进行业务操作。

迁移思路：

1. 先通过Spark1.5将历史数据迁移至中间表。
2. 再通过Spark2x将中间表的数据迁移至目标表，然后将目标表名修改为原表名。

3. 迁移完成后使用Spark2x操作CarbonData表中的数据。

具体迁移方案和命令

历史数据迁移

步骤1 中断CarbonData的数据入库业务，用Spark1.5的spark-beeline，查看CarbonData表当前最新的Segment的ID和时间，并记录此Segment的ID。

show segments for table dbname.tablename;

步骤2 用创建原CarbonData表的用户，执行Spark1.5的spark-beeline，创建ORC（或PARQUET）格式的中间表，然后将原CarbonData表中的数据导入该中间表，导入完成后即可恢复CarbonData表的业务。

创建ORC表：

CREATE TABLE dbname.mid_tablename_orc STORED AS ORC as select * from dbname.tablename;

创建PARQUET表：

CREATE TABLE dbname.mid_tablename_parq STORED AS PARQUET as select * from dbname.tablename;

其中dbname指数据库名，tablename指原CarbonData表名。

步骤3 用创建原CarbonData表的用户，执行Spark2x的spark-beeline；然后通过旧表的建表语句，创建新的CarbonData表。

说明

创建新表的语句中，字段顺序和类型必须和旧表的完全一致，如此才能保留原表的索引列等结构，且可以避免后续插入数据时因用到了 select * 而出错。

通过Spark1.5的spark-beeline命令查看旧表的建表语句：**SHOW CREATE TABLE dbname.tablename**;

创建新CarbonData表，名称为：**dbname.new_tablename**

步骤4 用创建原CarbonData表的用户，执行Spark2x的spark-beeline，将**步骤2**中创建的ORC（或PARQUET）格式的中间表的数据，加载到**步骤3**创建的新表中。此步骤可能耗时较长（200G数据耗时约2小时）。加载数据的命令以ORC格式的中间表举例：

insert into dbname.new_tablename select *

from dbname. mid_tablename_orc;

步骤5 用创建原CarbonData表的用户，执行Spark2x的spark-beeline，对新表的数据进行查询检验，确认数据无误后，将原CarbonData表修改为其他名称，再将新CarbonData表修改为原CarbonData表的名称。

ALTER TABLE dbname.tablename RENAME TO dbname.old_tablename;

ALTER TABLE dbname.new_tablename RENAME TO dbname.tablename;

步骤6 迁移完成。此时即可通过Spark2x对新表进行查询、重建二级索引等操作。

----结束

12.3.4 CarbonData 性能调优

12.3.4.1 调优指导

查询性能调优

CarbonData可以通过调整各种参数来提高查询性能。大部分参数聚焦于增加并行性处理和更好地使用系统资源。

- Spark Executor数量：Executor是Spark并行性的基础实体。通过增加Executor数量，集群中的并行数量也会增加。关于如何配置Executor数量，请参考Spark资料。
- Executor核：每个Executor内，并行任务数受Executor核的配置控制。通过增加Executor核数，可增加并行任务数，从而提高性能。
- HDFS block容量：CarbonData通过给不同的处理器分配不同的block来分配查询任务。所以一个HDFS block是一个分区单元。另外，CarbonData在Spark驱动器中，支持全局block级索引，这有助于减少需要被扫描的查询block的数量。设置较大的block容量，可提高I/O效率，但是会降低全局索引效率；设置较小的block容量，意味着更多的block数量，会降低I/O效率，但是会提高全局索引效率，同时，对于索引查询会要求更多的内存。
- 扫描线程数量：扫描仪（Scanner）线程控制每个任务中并行处理的数据块的数量。通过增加扫描仪线程数，可增加并行处理的数据块的数量，从而提高性能。可使用“carbon.properties”文件中的“carbon.number.of.cores”属性来配置扫描仪线程数。例如，“carbon.number.of.cores = 4”。
- B-Tree缓存：为了获得更好的查询特性，可以通过B-tree LRU（least recently used，最近最少使用）缓存来优化缓存内存。在driver中，B-Tree LRU缓存配置将有助于通过释放未被访问或未使用的表segments来释放缓存。类似地，在executor中，B-Tree LRU缓存配置将有助于释放未被访问或未使用的表blocks。具体可参考**表12-6**中的参数“carbon.max.driver.lru.cache.size”和“carbon.max.executor.lru.cache.size”的详细描述。

CarbonData 查询流程

当CarbonData首次收到对某个表（例如表A）的查询任务时，系统会加载表A的索引数据到内存中，执行查询流程。当CarbonData再次收到对表A的查询任务时，系统则不需要再加载其索引数据。

在CarbonData中执行查询时，查询任务会被分成几个扫描任务。即，基于CarbonData数据存储的HDFS block对扫描任务进行分割。扫描任务由集群中的执行器执行。扫描任务可以并行、部分并行，或顺序处理，具体采用的方式取决于执行器的数量以及配置的执行器核数。

查询任务的某些部分可在独立的任务级上处理，例如select和filter。查询任务的某些部分可在独立的任务级上进行部分处理，例如group-by、count、distinct count等。

某些操作无法在任务级上处理，例如Having Clause（分组后的过滤），sort等。这些无法在任务级上处理，或只能在任务级上部分处理的操作需要在集群内跨执行器来传输数据（部分结果）。这个传送操作被称为shuffle。

任务数量越多，需要shuffle的数据就越多，会对查询性能产生不利影响。

由于任务数量取决于HDFS block的数量，而HDFS block的数量取决于每个block的大小，因此合理选择HDFS block的大小很重要，需要在提高并行性，进行shuffle操作的数据量和聚合表的大小之间达到平衡。

分割和 Executors 的关系

如果分割数小于等于Executor数乘以Executor核数，那么任务将以并行方式运行。否则，某些任务只有在其他任务完成之后才能开始。因此，要确保Executor数乘以Executor核数大于等于分割数。同时，还要确保有足够的分割数，这样一个查询任务可被分为足够多的子任务，从而确保并行性。

配置扫描仪线程

扫描仪线程属性决定了每个分割的数据被划分的可并行处理的数据块的数量。如果数量过多，会产生很多小数据块，性能会受到影响。如果数量过少，并行性不佳，性能也会受到影响。因此，决定扫描仪线程数时，需要考虑一个分割内的平均数据大小，选择一个使数据块不会很小的值。经验法则是将单个块大小（MB）除以250得到的值作为扫描仪线程数。

增加并行性还需考虑的重要一点是集群中实际可用的CPU核数，确保并行计算数不超过实际CPU核数的75%至80%。

CPU核数约等于：

并行任务数x扫描仪线程数。其中并行任务数为分割数和执行器数x执行器核数两者之间的较小值。

数据加载性能调优

数据加载性能调优与查询性能调优差异很大。跟查询性能一样，数据加载性能也取决于可达到的并行性。在数据加载情况下，工作线程的数量决定并行的单元。因此，更多的执行器就意味着更多的执行器核数，每个执行器都可以提高数据加载性能。

同时，为了得到更好的性能，可在HDFS中配置如下参数。

表 12-15 HDFS 配置

参数	建议值
dfs.datanode.drop.cache.behind.reads	false
dfs.datanode.drop.cache.behind.writes	false
dfs.datanode.sync.behind.writes	true

压缩调优

CarbonData结合少数轻量级压缩算法和重量级压缩算法来压缩数据。虽然这些算法可处理任何类型的数据，但如果数据经过排序，相似值在一起出现时，就会获得更好的压缩率。

CarbonData数据加载过程中，数据基于Table中的列顺序进行排序，从而确保相似值在一起出现，以获得更好的压缩率。

由于CarbonData按照Table中定义的列顺序将数据进行排序，因此列顺序对于压缩效率起重要作用。如果低cardinality维度位于左边，那么排序后的数据分区范围较小，压缩效率较高。如果高cardinality维度位于左边，那么排序后的数据分区范围较大，压缩效率较低。

内存调优

CarbonData为内存调优提供了一个机制，其中数据加载会依赖于查询中需要的列。不论何时，接收到一个查询命令，将会获取到该查询中的列，并确保内存中这些列有数据加载。在该操作期间，如果达到内存的阈值，为了给查询需要的列提供内存空间，最少使用加载级别的文件将会被删除。

12.3.4.2 创建 CarbonData Table 的建议

操作场景

本章节根据超过50个测试用例总结得出建议，帮助用户创建拥有更高查询性能的CarbonData表。

表 12-16 CarbonData 表中的列

Column name	Data type	Cardinality	Attribution
msisdn	String	3千万	dimension
BEGIN_TIME	bigint	1万	dimension
host	String	1百万	dimension
dime_1	String	1千	dimension
dime_2	String	500	dimension
dime_3	String	800	dimension
counter_1	numeric(20,0)	NA	measure
...	...	NA	measure
counter_100	numeric(20,0)	NA	measure

操作步骤

- 如果待创建的表有一个常用于过滤的列 ，例如80%以上的场景使用此列过滤。

  针对此类场景，调优方法如下：

  将常用于过滤的列放在sort_columns第一列。

  例如，msisdn作为过滤条件在查询中使用的最多，则将其放在第一列。创建表的命令如下，其中采用msisdn作为过滤条件的查询性能将会很好。

  ```
  create table carbondata_table(
     msisdn String,
     ...
     )STORED AS carbondata TBLPROPERTIES ('SORT_COLUMS'='msisdn');
  ```

- 如果待创建的表有多个常用于过滤的列。

针对此类场景，调优方法如下：

为常用的过滤列创建索引。

例如，如果msisdn，host和dime_1是过滤经常使用的列，根据cardinality，sort_columns列的顺序是dime_1-> host-> msisdn…。创建表命令如下，以下命令可提高dime_1，host和msisdn上的过滤性能。

```
create table carbondata_table(
    dime_1 String,
    host String,
    msisdn String,
    dime_2 String,
    dime_3 String,
    ...
    )STORED AS carbondata
TBLPROPERTIES ('SORT_COLUMS'='dime_1,host,msisdn');
```

- 如果每个用于过滤的列的频率相当。

  针对此类场景，调优方法如下：

  sort_columns按照cardinality从低到高的顺序排列。

  创建表的命令如下：

  ```
  create table carbondata_table(
      Dime_1 String,
      BEGIN_TIME bigint,
      HOST String,
      MSISDN String,
      ...
      )STORED AS carbondata
  TBLPROPERTIES ('SORT_COLUMS'='dime_2,dime_3,dime_1, BEGIN_TIME,host,msisdn');
  ```

- 按照维度的cardinality从低到高创建表后，再为高Cardinality列创建SECONDARY INDEX。创建索引的语句如下：

  ```
  create index carbondata_table_index_msidn on tablecarbondata_table (
  MSISDN String) as 'carbondata' PROPERTIES ('table_blocksize'='128');
  create index carbondata_table_index_host on tablecarbondata_table (
  host String) as 'carbondata' PROPERTIES ('table_blocksize'='128');
  ```

- 对于不需要高精度的度量，无需使用numeric (20,0)数据类型，建议使用double数据类型来替换numeric (20,0)数据类型，以提高查询性能。

  在一个测试用例中，使用double来替换numeric (20, 0)，查询时间从15秒降低到3秒，查询速度提高了5倍。创建表命令如下：

  ```
  create table carbondata_table(
      Dime_1 String,
      BEGIN_TIME bigint,
      HOST String,
      MSISDN String,
      counter_1 double,
      counter_2 double,
      ...
      counter_100 double,
      )STORED AS carbondata
  ;
  ```

- 如果列值总是递增的，如start_time。

  例如，每天将数据加载到CarbonData，start_time是每次加载的增量。对于这种情况，建议将start_time列放在sort_columns的最后，因为总是递增的值可以始终使用最小/最大索引。创建表命令如下：

  ```
  create table carbondata_table(
      Dime_1 String,
      HOST String,
      MSISDN String,
      counter_1 double,
      counter_2 double,
      BEGIN_TIME bigint,
  ```

```
...
counter_100 double,
)STORED AS carbondata
TBLPROPERTIES ( 'SORT_COLUMS'='dime_2,dime_3,dime_1..BEGIN_TIME');
```

12.3.4.3 性能调优的相关配置

操作场景

CarbonData的性能与配置参数相关，本章节提供了能够提升性能的相关配置介绍。

操作步骤

用于CarbonData查询的配置介绍，详情请参见表12-17和表12-18。

表 12-17 Shuffle 过程中，启动 Task 的个数

参数	spark.sql.shuffle.partitions
所属配置文件	spark-defaults.conf
适用于	数据查询
场景描述	Spark shuffle时启动的Task个数。
如何调优	一般建议将该参数值设置为执行器核数的1到2倍。例如，在聚合场景中，将task个数从200减少到32，有些查询的性能可提升2倍。

表 12-18 设置用于 CarbonData 查询的 Executor 个数、CPU 核数以及内存大小

参数	spark.executor.cores spark.executor.instances spark.executor.memory
所属配置文件	spark-defaults.conf
适用于	数据查询
场景描述	设置用于CarbonData查询的Executor个数、CPU核数以及内存大小。
如何调优	在银行方案中，为每个执行器提供4个CPU内核和15GB内存，可以获得良好的性能。这2个值并不意味着越多越好，在资源有限的情况下，需要正确配置。例如，在银行方案中，每个节点有足够的32个CPU核，而只有64GB的内存，这个内存是不够的。例如，当每个执行器有4个内核和12GB内存，有时在查询期间发生垃圾收集（GC），会导致查询时间从3秒增加到超过15秒。在这种情况下需要增加内存或减少CPU内核。

用于CarbonData数据加载的配置参数，详情请参见表12-19、表12-20和表12-21。

表 12-19 设置数据加载使用的 CPU core 数量

参数	carbon.number.of.cores.while.loading
所属配置文件	carbon.properties
适用于	数据加载
场景描述	数据加载过程中，设置处理数据使用的CPU core数量。
如何调优	如果有更多的CPU个数，那么可以增加CPU值来提高性能。例如，将该参数值从2增加到4，那么CSV文件读取性能可以增加大约1倍。

表 12-20 是否使用 YARN 本地目录进行多磁盘数据加载

参数	carbon.use.local.dir
所属配置文件	carbon.properties
适用于	数据加载
场景描述	是否使用YARN本地目录进行多磁盘数据加载。
如何调优	如果将该参数值设置为“true”，CarbonData将使用YARN本地目录进行多表加载磁盘负载平衡，以提高数据加载性能。

表 12-21 加载时是否使用多路径

参数	carbon.use.multiple.temp.dir
所属配置文件	carbon.properties
适用于	数据加载
场景描述	是否使用多个临时目录存储sort临时文件。
如何调优	设置为true，则数据加载时使用多个临时目录存储sort临时文件。此配置能提高数据加载性能并避免磁盘单点故障。

用于CarbonData数据加载和数据查询的配置参数，详情请参见**表12-22**。

表 12-22 设置数据加载和查询使用的 CPU core 数量

参数	carbon.compaction.level.threshold
所属配置文件	carbon.properties
适用于	数据加载和查询

场景描述	对于minor压缩，在阶段1中要合并的segment数量和在阶段2中要合并的已压缩的segment数量。
如何调优	每次CarbonData加载创建一个segment，如果每次加载的数据量较小，将在一段时间内生成许多小文件，影响查询性能。配置该参数将小的segment合并为一个大的segment，然后对数据进行排序，可提高查询性能。 压缩的策略根据实际的数据大小和可用资源决定。如某银行1天加载一次数据，且加载数据选择在晚上无查询时进行，有足够的资源，压缩策略可选择为6、5。

表 12-23 使用索引缓存服务器时是否开启数据预加载

参数	carbon.indexserver.enable.prepriming
所属配置文件	carbon.properties
适用于	数据加载
场景描述	使用索引缓存服务器过程中开启数据预加载可以提升首次查询的性能。
如何调优	用户可以将该参数设置为true来开启预加载。默认情况，该参数为false。

12.3.5 CarbonData 访问控制

下表提供了对CarbonData Table执行相应操作所需的Hive ACL特权的详细信息。

前提条件

已经设置了表12-9或表12-10中Carbon相关参数。

Hive ACL 权限

表 12-24 CarbonData 表级操作所需的 Hive ACL 权限

场景	所需权限
DESCRIBE TABLE	SELECT (of table)
SELECT	SELECT (of table)
EXPLAIN	SELECT (of table)
CREATE TABLE	CREATE (of database)
CREATE TABLE As SELECT	CREATE (on database), INSERT (on table), RW on data file, and SELECT (on table)

场景	所需权限
LOAD	INSERT (of table) RW on data file
DROP TABLE	OWNER (of table)
DELETE SEGMENTS	DELETE (of table)
SHOW SEGMENTS	SELECT (of table)
CLEAN FILES	DELETE (of table)
INSERT OVERWRITE / INSERT INTO	INSERT (of table) RW on data file and SELECT (of table)
CREATE INDEX	OWNER (of table)
DROP INDEX	OWNER (of table)
SHOW INDEXES	SELECT (of table)
ALTER TABLE ADD COLUMN	OWNER (of table)
ALTER TABLE DROP COLUMN	OWNER (of table)
ALTER TABLE CHANGE DATATYPE	OWNER (of table)
ALTER TABLE RENAME	OWNER (of table)
ALTER TABLE COMPACTION	INSERT (on table)
FINISH STREAMING	OWNER (of table)
ALTER TABLE SET STREAMING PROPERTIES	OWNER (of table)
ALTER TABLE SET TABLE PROPERTIES	OWNER (of table)
UPDATE CARBON TABLE	UPDATE (of table)
DELETE RECORDS	DELETE (of table)
REFRESH TABLE	OWNER (of main table)
REGISTER INDEX TABLE	OWNER (of table)
SHOW PARTITIONS	SELECT (on table)
ALTER TABLE ADD PARTITION	OWNER (of table)
ALTER TABLE DROP PARTITION	OWNER (of table)

说明

- 如果数据库下的表由多个用户创建，那么执行Drop database命令会失败，即使执行的用户是数据库的拥有者。
- 在二级索引中，当父表（parent table）触发时，insert和compaction将在索引表上触发。如果选择具有过滤条件匹配索引表列的查询，用户应该为父表和索引表提供选择权限。
- LockFiles文件夹和LockFiles文件夹中创建的锁定文件将具有完全权限，因为LockFiles文件夹不包含任何敏感数据。
- 如果使用ACL，确保不要为DDL或DML配置任何被其他进程使用中的路径，建议创建新路径。

  以下配置项需要配置路径：

  1) carbon.badRecords.location

  2) 创建数据库时Db_Path及其他。
- 对于非安全集群中的Carbon ACL权限，hive-site.xml中的参数hive.server2.enable.doAs必须设置为false。 将此属性设置为false，查询将以hiveserver2进程运行的用户身份运行。

12.3.6 CarbonData 语法参考

12.3.6.1 DDL

12.3.6.1.1 CREATE TABLE

命令功能

CREATE TABLE命令通过指定带有表属性的字段列表来创建CarbonData Table。

命令格式

CREATE TABLE *[IF NOT EXISTS] [db_name.]table_name*

[(col_name data_type, ...)]

STORED AS *carbondata*

[TBLPROPERTIES (property_name=property_value, ...)];

所有表的附加属性都会放到TBLPROPERTIES中来定义。

参数描述

表 12-25 CREATE TABLE 参数描述

参数	描述
db_name	Database名称，由字母、数字和下划线（_）组成。
col_name data_type	以逗号分隔的带数据类型的列表。列名由字母、数字和下划线（_）组成。 **说明** 在CarbonData表创建过程中，不允许使用tupleId，PositionId和PositionReference为列命名，因为具有这些名称的列由二级索引命令在内部使用。

参数	描述
table_name	Database中的表名，由字母、数字和下划线（ _ ）组成。
STORED AS	参数carbondata，定义和创建CarbonData table。
TBLPROPERTIES	CarbonData table属性列表。

注意事项

以下是表格属性的使用。

- Block大小

  单个表的数据文件block大小可以通过TBLPROPERTIES进行定义，系统会选择数据文件实际大小和设置的blocksize大小中的较大值，作为该数据文件在HDFS上存储的实际blocksize大小。单位为MB，默认值为1024MB，范围为1MB~2048MB。若设置值不在[1, 2048]之间，系统将会报错。

  一旦block大小达到配置值，写入程序将启动新的CarbonData数据的block。数据以页面大小（32000个记录）的倍数写入，因此边界在字节级别上不严格。 如果新页面跨越配置block的边界，则不会将其写入当前block，而是写入新的block。

  TBLPROPERTIES('table_blocksize'='128')

  说明

  - 当在CarbonData表中配置了较小的blocksize，而加载的数据生成的数据文件比较大时，在HDFS上显示的blocksize会与设置值不同。这是因为，对于每一个本地block文件的首次写入，即使待写入数据的大小大于blocksize的配置值，也直接将待写入数据写入此block。所以，HDFS上blocksize的实际值为待写入数据大小与blocksize配置值中的较大值。
  - 当CarbonData表中的数据文件block.num小于任务并行度（parellelism）时，CarbonData数据文件的block会被切为新的block，使得blocks.num大于parellelism，这样所有core均可被使用。这种优化称为block distribution。

- SORT_SCOPE：指定表创建时的排序范围。如下为四种排序范围。
  - GLOBAL_SORT：它提高了查询性能，特别是点查询。
    TBLPROPERTIES('SORT_SCOPE'='GLOBAL_SORT')
  - LOCAL_SORT：数据会本地排序（任务级别排序）。
  - NO_SORT：默认排序。它将以不排序的方式加载数据，这将显着提升加载性能。

- SORT_COLUMNS

  此表属性指定排序列的顺序。

  TBLPROPERTIES('SORT_COLUMNS'='column1, column3')

  说明

  - 如果未指定此属性，则默认情况下，没有列会被排序。
  - 如果指定了此属性，但具有空参数，则表将被加载而不进行排序。例如，*('SORT_COLUMNS'='')*。
  - SORT_COLUMNS将接受string，date，timestamp，short，int，long, byte和boolean数据类型。

- RANGE_COLUMN

此表属性指定一列，该列将会按照一个范围值来对输入的数据进行分区。仅可配置一列。在数据导入过程中，可以使用“global_sort_partitions”或者“scale_factor”来避免生成小文件。

TBLPROPERTIES('RANGE_COLUMN'='column1')

- LONG_STRING_COLUMNS

  普通String类型的长度不能超过32000字符，如果需要存储超过32000字符的字符串，指定LONG_STRING_COLUMNS配置为该列。

  TBLPROPERTIES('LONG_STRING_COLUMNS'='column1, column3')

  说明

  LONG_STRING_COLUMNS仅可以设置string/char/varchar类型的列，并且不能为SORT_COLUMNS和复杂列。

使用场景

通过指定列创建表

CREATE TABLE命令与Hive DDL相同。CarbonData的额外配置将作为表格属性给出。

CREATE TABLE *[IF NOT EXISTS] [db_name.]table_name*

[(col_name data_type , ...)]

STORED AS *carbondata*

[TBLPROPERTIES (property_name=property_value, ...)];

示例

CREATE TABLE *IF NOT EXISTS productdb.productSalesTable (*

productNumber Int,

productName String,

storeCity String,

storeProvince String,

productCategory String,

productBatch String,

saleQuantity Int,

revenue Int)

STORED AS carbondata

TBLPROPERTIES (

'table_blocksize'='128',

'SORT_COLUMNS'='productBatch, productName')

系统响应

Table创建成功，创建成功的消息将被记录在系统日志中。

12.3.6.1.2 CREATE TABLE As SELECT

命令功能

CREATE TABLE As SELECT命令通过指定带有表属性的字段列表来创建CarbonData Table。

命令格式

CREATE TABLE *[IF NOT EXISTS] [db_name.]table_name* ***STORED AS carbondata*** *[TBLPROPERTIES (key1=val1, key2=val2, ...)] AS select_statement;*

参数描述

表 12-26 CREATE TABLE 参数描述

参数	描述
db_name	Database名称，由字母、数字和下划线（_）组成。
table_name	Database中的表名，由字母、数字和下划线（_）组成。
STORED AS	使用CarbonData数据格式存储数据。
TBLPROPERTIES	CarbonData table属性列表。详细信息，见**注意事项**。

注意事项

NA

示例

CREATE TABLE ctas_select_parquet ***STORED AS*** carbondata as select * from parquet_ctas_test;

系统响应

该命令会从Parquet表上创建一个Carbon表，同时导入所有Parquet表的数据。

12.3.6.1.3 DROP TABLE

命令功能

DROP TABLE的功能是用来删除已存在的Table。

命令格式

DROP TABLE *[IF EXISTS] [db_name.]table_name;*

参数描述

表 12-27 DROP TABLE 参数描述

参数	描述
db_name	Database名称。如果未指定，将选择当前database。
table_name	需要删除的Table名称。

注意事项

在该命令中，IF EXISTS和db_name是可选配置。

示例

DROP TABLE IF EXISTS productDatabase.productSalesTable;

系统响应

Table将被删除。

12.3.6.1.4 SHOW TABLES

命令功能

SHOW TABLES命令用于显示所有在当前database中的table，或所有指定database的table。

命令格式

SHOW TABLES *[IN db_name];*

参数描述

表 12-28 SHOW TABLES 参数描述

参数	描述
IN db_name	Database名称，仅当需要显示指定Database的所有Table时配置。

注意事项

IN db_Name为可选配置。

示例

SHOW TABLES IN ProductDatabase;

系统响应

显示所有Table。

12.3.6.1.5 ALTER TABLE COMPACTION

命令功能

ALTER TABLE COMPACTION命令将合并指定数量的segment为一个segment。这将提高该表的查询性能。

命令格式

ALTER TABLE *[db_name.]table_name COMPACT 'MINOR/MAJOR/SEGMENT_INDEX';*

ALTER TABLE *[db_name.]table_name COMPACT 'CUSTOM' WHERE SEGMENT.ID IN (id1, id2, ...);*

参数描述

表 12-29 ALTER TABLE COMPACTION 参数描述

Parameter	Description
db_name	数据库名。若未指定，则选择当前数据库。
table_name	表名。
MINOR	Minor合并，详见**合并Segments**。
MAJOR	Major合并，详见**合并Segments**。
SEGMENT_INDEX	这会将一个segment内的所有Carbon索引文件（.carbonindex）合并为一个Carbon索引合并文件（.carbonindexmerge）。 这增强了首次查询性能。详见**表12-14**。
CUSTOM	Custom合并，详见**合并Segments**。

注意事项

NA

示例

ALTER TABLE ProductDatabase COMPACT 'MINOR';

ALTER TABLE ProductDatabase COMPACT 'MAJOR';

ALTER TABLE ProductDatabase COMPACT 'SEGMENT_INDEX';

ALTER TABLE ProductDatabase COMPACT 'CUSTOM' WHERE SEGMENT.ID IN (0, 1);

系统响应

由于为后台运行，**ALTER TABLE COMPACTION**命令不会显示压缩响应。

如果想要查看MINOR合并和MAJOR合并的响应结果，用户可以检查日志或运行***SHOW SEGMENTS***命令查看。

示例：

```
+------+------------+--------------------------+-----------------+------------+------------+-------------+--------------
+--+
|  ID  |   Status   |     Load Start Time      | Load Time Taken  | Partition  | Data Size  | Index Size  | File
Format  |
+------+------------+--------------------------+-----------------+------------+------------+-------------+--------------
+--+
| 3    | Success    | 2020-09-28 22:53:26.336  | 3.726S           | {}         | 6.47KB     | 3.30KB      | columnar_v3  |
| 2    | Success    | 2020-09-28 22:53:01.702  | 6.688S           | {}         | 6.47KB     | 3.30KB      | columnar_v3  |
| 1    | Compacted  | 2020-09-28 22:51:15.242  | 5.82S            | {}         | 6.50KB     | 3.43KB      |
columnar_v3  |
| 0.1  | Success    | 2020-10-30 20:49:24.561  | 16.66S           | {}         | 12.87KB    | 6.91KB      | columnar_v3
|
| 0    | Compacted  | 2020-09-28 22:51:02.6    | 6.819S           | {}         | 6.50KB     | 3.43KB      | columnar_v3
|
+------+------------+--------------------------+-----------------+------------+------------+-------------+--------------
+--+
```

其中，

- Compacted表示该数据已被合并。
- 0.1表示segment0与segment1合并之后的结果。

数据合并前后的其他操作没有差别。

被合并的segments（例如segment0和segment1）即成为无用的segments，会占用空间，因此建议合并之后使用***CLEAN FILES***命令进行彻底删除，再进行其他操作。**CLEAN FILES**命令的使用方法可参考**CLEAN FILES**。

12.3.6.1.6 TABLE RENAME

命令功能

RENAME命令用于重命名现有表。

命令语法

ALTER TABLE *[db_name.]table_name* **RENAME TO** *new_table_name*;

参数描述

表 12-30 RENAME 参数描述

参数	描述
db_name	数据库名。若未指定，则选择当前数据库。
table_name	现有表名。
new_table_name	现有表名的新表名。

注意事项

- 并行运行的查询（需要使用表名获取路径，以读取CarbonData存储文件）可能会在此操作期间失败。
- 不允许二级索引表重命名。

示例

ALTER TABLE *carbon* **RENAME TO** *carbondata*;

ALTER TABLE *test_db.carbon* **RENAME TO** *test_db.carbondata*;

系统响应

CarbonData库中的文件夹将显示新表名称，可以通过运行SHOW TABLES显示新表名称。

12.3.6.1.7 ADD COLUMNS

命令功能

ADD COLUMNS命令用于为现有表添加新列。

命令语法

ALTER TABLE *[db_name.]table_name* **ADD COLUMNS** *(col_name data_type,...)* **TBLPROPERTIES** *(''COLUMNPROPERTIES.columnName.shared_column'='sharedFolder.sharedColumnName,...', 'DEFAULT.VALUE.COLUMN_NAME'='default_value')*;

参数描述

表 12-31 ADD COLUMNS 参数描述

参数	描述
db_name	数据库名。若未指定，则选择当前数据库。
table_name	表名。
col_name data_type	带数据类型且用逗号分隔的列的名称。列名称包含字母，数字和下划线（_）。 **说明** 创建CarbonData表时，不要将列名命名为tupleId，PositionId和PositionReference，因为将在UPDATE，DELETE和二级索引命令内部使用这些名称。

注意事项

- 除了shared_column和default_value之外，将不会读取其他属性。如果指定了任何其他属性名称，则不会抛出错误，其他属性将被忽略。

- 如果未指定默认值，则新列的默认值将被视为null。
- 如果在该列上应用filter，则在排序期间不会考虑新增列，新增列可能会影响查询性能。

示例

- **ALTER TABLE** *carbon* **ADD COLUMNS** *(a1 INT, b1 STRING)*;
- **ALTER TABLE** *carbon* **ADD COLUMNS** *(a1 INT, b1 STRING)* **TBLPROPERTIES** *('COLUMNPROPERTIES.b1.shared_column'='sharedFolder.b1')*;
- ALTER TABLE *carbon* **ADD COLUMNS** *(a1 INT, b1 STRING)* **TBLPROPERTIES** *('DEFAULT.VALUE.a1'='10')*;

系统响应

通过运行DESCRIBE命令，可显示新添加的列。

12.3.6.1.8 DROP COLUMNS

命令功能

DROP COLUMNS命令用于删除表中现有的列或多个列。

命令语法

ALTER TABLE *[db_name.]table_name* **DROP COLUMNS** *(col_name, ...)*;

参数描述

表 12-32 DROP COLUMNS 参数描述

参数	描述
db_name	数据库名。若未指定，则选择当前数据库。
table_name	表名。
col_name	表中的列名称。支持多列。列名称包含字母，数字和下划线（_）。

注意事项

对于删除列操作，至少要有一个key列在删除操作后存在于schema中，否则将显示出错信息，删除列操作将失败。

示例

假设表包含4个列，分别命名为a1，b1，c1和d1。

- 删除单个列：
  ALTER TABLE *carbon* **DROP COLUMNS** *(b1)*;

ALTER TABLE *test_db.carbon* **DROP COLUMNS** *(b1)*;

- 删除多个列：
  ALTER TABLE *carbon* **DROP COLUMNS** *(b1,c1)*;
  ALTER TABLE *test_db.carbon* **DROP COLUMNS** *(b1,c1)*;

系统响应

运行DESCRIBE命令，将不会显示已删除的列。

12.3.6.1.9 CHANGE DATA TYPE

命令功能

CHANGE命令用于将数据类型从INT更改为BIGINT或将Decimal精度从低精度改为高精度。

命令语法

ALTER TABLE *[db_name.]table_name* **CHANGE** *col_name col_name changed_column_type*;

参数描述

表 12-33 CHANGE DATA TYPE 参数描述

参数	描述
db_name	数据库名。若未指定，则选择当前数据库。
table_name	表名。
col_name	表中的列名称。列名称包含字母，数字和下划线（_）。
changed_column_type	所要更改为的新数据类型。

注意事项

- 仅在没有数据丢失的情况下支持将Decimal数据类型从较低精度更改为较高精度
  例如：
  - 无效场景：将Decimal数据精度从（10,2）更改为（10,5）无效，因为在这种情况下，只有scale增加，但总位数保持不变。
  - 有效场景：将Decimal数据精度从（10,2）更改为（12,3）有效，因为总位数增加2，但是scale仅增加1，这不会导致任何数据丢失。
- 将Decimal数据类型从较低精度更改为较高精度，其允许的最大精度(precision, scale)范围为(38,38)，并且只适用于不会导致数据丢失的有效提升精度的场景。

示例

- 将列a1的数据类型从INT更改为BIGINT。
  ALTER TABLE *test_db.carbon* **CHANGE** *a1 a1 BIGINT*;

- 将列a1的精度从10更改为18。

  ALTER TABLE *test_db.carbon* **CHANGE** *a1 a1 DECIMAL(18,2)*;

系统响应

通过运行DESCRIBE命令，将显示被修改列变更后的数据类型。

12.3.6.1.10 REFRESH TABLE

命令功能

REFRESH TABLE命令用于将已有的Carbon表数据注册到Hive元数据库中。

命令语法

REFRESH TABLE *db_name.table_name*;

参数描述

表 12-34 REFRESH TABLE 参数描述

参数	描述
db_name	数据库名。若未指定，则选择当前数据库。
table_name	表名。

注意事项

- 在执行此命令之前，应将旧表的表结构定义schema和数据复制到新数据库位置。
- 对于旧版本仓库，源集群和目的集群的时区应该相同。
- 新的数据库和旧数据库的名字应该相同。
- 执行命令前，旧表的表结构定义schema和数据应该复制到新的数据库位置。
- 如果表是聚合表，则应将所有聚合表复制到新的数据库位置。
- 如果旧集群使用HIVE元数据库来存储表结构，则刷新将不起作用，因为文件系统中不存在表结构定义schema文件。

示例

REFRESH TABLE *dbcarbon.productSalesTable*;

系统响应

通过运行该命令，已有的Carbon表数据会被注册到Hive元数据库中。

12.3.6.1.11 REGISTER INDEX TABLE

命令功能

REGISTER INDEX TABLE命令用于将索引表注册到主表。

命令语法

REGISTER INDEX TABLE *indextable_name* ON *db_name.maintable_name*;

参数描述

表 12-35 REFRESH INDEX TABLE 参数描述

参数	描述
db_name	数据库名。若未指定，则选择当前数据库。
indextable_name	索引表名。
maintable_name	主表名。

注意事项

在执行此命令之前，使用REFRESH TABLE将主表和二级所索引表都注册到Hive元数据中。

示例

create database productdb;

use productdb;

CREATE TABLE productSalesTable(a int,b string,c string) stored as carbondata;

create index productNameIndexTable on table productSalesTable(c) as 'carbondata';

insert into table productSalesTable select 1,'a','aaa';

create database productdb2;

使用hdfs命令将productdb数据库下的productSalesTable和productNameIndexTable拷贝到productdb2。

refresh table productdb2.productSalesTable ;

refresh table productdb2.productNameIndexTable ;

explain select * from productdb2.productSalesTable where c = 'aaa'; //可以发现该查询命令没有使用索引表

REGISTER INDEX TABLE productNameIndexTable ON productdb2.productSalesTable;

explain select * from productdb2.productSalesTable where c = 'aaa'; //可以发现该查询命令使用了索引表

系统响应

通过运行该命令，索引表会被注册到主表。

12.3.6.2 DML

12.3.6.2.1 LOAD DATA

命令功能

LOAD DATA命令以CarbonData特定的数据存储类型加载原始的用户数据，这样，CarbonData可以在查询数据时提供良好的性能。

说明

仅支持加载位于HDFS上的原始数据。

命令格式

LOAD DATA *INPATH 'folder_path' INTO TABLE [db_name.]table_name OPTIONS(property_name=property_value, ...);*

参数描述

表 12-36 LOAD DATA 参数描述

参数	描述
folder_path	原始CSV数据文件夹或者文件的路径。
db_name	Database名称。若未指定，则使用当前database。
table_name	所提供的database中的表的名称。

注意事项

以下是可以在加载数据时使用的配置选项：

- DELIMITER：可以在加载命令中提供分隔符和引号字符。默认值为**,**。
  OPTIONS('DELIMITER'=',' , 'QUOTECHAR'='"')
  可使用'DELIMITER'='\t'来表示用制表符tab对CSV数据进行分隔。
  OPTIONS('DELIMITER'='\t')
  CarbonData也支持\001和\017作为分隔符。

  说明

  对于CSV数据，分隔符为单引号（'）时，单引号必须在双引号（" "）内。例如：'DELIMITER'= "'"。

- QUOTECHAR：可以在加载命令中提供分隔符和引号字符。默认值为"。
  OPTIONS('DELIMITER'=',' , 'QUOTECHAR'='"')
- COMMENTCHAR：可以在加载命令中提供注释字符。在加载操作期间，如果在行的开头遇到注释字符，那么该行将被视为注释，并且不会被加载。默认值为 # 。
  OPTIONS('COMMENTCHAR'='#')
- FILEHEADER：如果源文件中没有表头，可在LOAD DATA命令中提供表头。
  OPTIONS('FILEHEADER'='column1,column2')
- ESCAPECHAR：如果用户想在CSV上对Escape字符进行严格验证，可以提供Escape字符。默认值为\。
  OPTIONS('ESCAPECHAR'='\')

  说明

  如果在CSV数据中输入ESCAPECHAR，该ESCAPECHAR必须在双引号（ " " ）内。例如："a \b"。
- Bad Records处理：
  为了使数据处理应用程序为用户增值，不可避免地需要对数据进行某种程度的集成。在大多数情况下，数据质量问题源于生成源数据的上游（主要）系统。
  有两种完全不同的方式处理Bad Data：
  - 按照数据原本的样子加载所有数据，之后进行除错处理。
  - 在进入数据源的过程中，可以清理或擦除Bad Data，或者在发现Bad Data时让数据加载失败。

  有多个选项可用于在CarbonData数据加载过程中清除源数据。对于CarbonData数据中的Bad Records管理，请参见表12-37。

表 12-37 Bad Records Logger

配置项	默认值	描述
BAD_RECORDS_LOGGER_ENABLE	false	若设置为true，则将创建Bad Records日志文件，其中包含Bad Records的详细信息。
BAD_RECORDS_ACTION	FAIL	以下为Bad Records的四种操作类型： ● FORCE：通过将Bad Records存储为NULL来自动校正数据。 ● REDIRECT：无法加载Bad Records，并将其写入原始CSV文件。 ● IGNORE：既不加载Bad Records也不将其写入原始CSV文件。 ● FAIL：如果发现存在Bad Records，数据加载将会失败。 **说明** 在加载数据时，如果所有记录都是Bad Records，则参数BAD_RECORDS_ACTION将不起作用，加载数据操作将会失败。

配置项	默认值	描述
IS_EMPTY_DATA_BAD_RECORD	false	如果设置为“false”，则空（""或''或,,）数据将不被视为Bad Records，如果设置为“true”，则空数据将被视为Bad Records。
BAD_RECORD_PATH	-	指定存储Bad Records的HDFS路径。默认值为Null。 如果启用了Bad Records日志记录或者Bad Records操作重定向，则该路径必须由用户进行配置。

示例：

LOAD DATA INPATH *'filepath.csv'* **INTO TABLE** *tablename*
OPTIONS('BAD_RECORDS_LOGGER_ENABLE'='true',
'BAD_RECORD_PATH'='hdfs://hacluster/tmp/carbon',
'BAD_RECORDS_ACTION'='REDIRECT',
'IS_EMPTY_DATA_BAD_RECORD'='false');

说明

使用“REDIRECT”选项，CarbonData会将所有的Bad Records添加到单独的CSV文件中，但是该文件内容不能用于后续的数据加载，因为其内容可能无法与源记录完全匹配。用户必须清理原始源记录以便于进一步的数据提取。该选项的目的只是让用户知道哪些记录被视为Bad Records。

- MAXCOLUMNS：该可选参数指定了在一行中，由CSV解析器解析的最大列数。

  OPTIONS('MAXCOLUMNS'='400')

表 12-38 MAXCOLUMNS

可选参数名称	默认值	最大值
MAXCOLUMNS	2000	20000

表 12-39 MAXCOLUMNS 可选参数的行为图

MAXCOLUMNS值	在文件Header选项中的列数	考虑的最终值
在加载项中未指定	5	2000
在加载项中未指定	6000	6000
40	7	文件header列数与MAXCOLUMNS值，两者中的最大值
22000	40	20000

MAXCOLUMNS值	在文件Header选项中的列数	考虑的最终值
60	在加载项中未指定	CSV文件第一行的列数与MAXCOLUMNS值，两者中的最大值

说明

对于设置MAXCOLUMNS Option的最大值，要求executor具有足够的内存，否则，数据加载会由于内存不足的错误而失败。

- 如果在创建表期间将SORT_SCOPE定义为GLOBAL_SORT，则可以指定在对数据进行排序时要使用的分区数。如果未配置或配置小于1，则将使用map任务的数量作为reduce任务的数量。 建议每个reduce任务处理512MB - 1GB数据。

  OPTIONS('GLOBAL_SORT_PARTITIONS'='2')

  说明

  增加分区数可能需要增加“spark.driver.maxResultSize”，因为在driver中收集的采样数据随着分区的增加而增加。

- DATEFORMAT：此选项用于指定表的日期格式。

  OPTIONS('DATEFORMAT'='dateFormat')

  说明

  日期格式由日期模式字符串指定。 Carbon中的日期模式字母与JAVA中的日期模式字母相同。

- TIMESTAMPFORMAT：此选项用于指定表的时间戳格式。
- *OPTIONS('TIMESTAMPFORMAT'='timestampFormat')*
- SKIP_EMPTY_LINE：数据加载期间，此选项将忽略CSV文件中的空行。

  OPTIONS('SKIP_EMPTY_LINE'='TRUE/FALSE')

- **可选:** SCALE_FACTOR：针对RANGE_COLUMN，SCALE_FACTOR用来控制分区的数量，根据如下公式：

  ```
  splitSize = max(blocklet_size, (block_size - blocklet_size)) * scale_factor
  numPartitions = total size of input data / splitSize
  ```

  默认值为3，range的范围为[1, 300]。

  OPTIONS('SCALE_FACTOR'='10')

  说明

  - 如果GLOBAL_SORT_PARTITIONS和SCALE_FACTOR同时使用，只有GLOBAL_SORT_PARTITIONS生效。
  - RANGE_COLUMN合并默认使用LOCAL_SORT。

使用场景

可使用下列语句从CSV文件加载CarbonData table。

LOAD DATA *INPATH 'folder path' INTO TABLE tablename OPTIONS(property_name=property_value, ...);*

示例

data.csv源文件数据如下所示：

```
ID,date,country,name,phonetype,serialname,salary
4,2014-01-21 00:00:00,xxx,aaa4,phone2435,ASD66902,15003
5,2014-01-22 00:00:00,xxx,aaa5,phone2441,ASD90633,15004
6,2014-03-07 00:00:00,xxx,aaa6,phone294,ASD59961,15005
```

CREATE TABLE carbontable(ID int, date Timestamp, country String, name String, phonetype String, serialname String,salary int) STORED AS carbondata;

LOAD DATA *inpath 'hdfs://hacluster/tmp/data.csv' INTO table carbontable*

options('DELIMITER'=',');

系统响应

可在driver日志中查看命令运行成功或失败。

12.3.6.2.2 UPDATE CARBON TABLE

命令功能

UPDATE命令根据列表达式和可选的过滤条件更新CarbonData表。

命令格式

- 格式1：

  UPDATE <CARBON TABLE> SET (column_name1, column_name2, ... column_name n) = (column1_expression , column2_expression , column3_expression ... column n_expression) [WHERE { <filter_condition> }];

- 格式2：

  UPDATE <CARBON TABLE> SET (column_name1, column_name2,) = (select sourceColumn1, sourceColumn2 from sourceTable [WHERE { <filter_condition> }]) [WHERE { <filter_condition> }];

参数描述

表 12-40 UPDATE 参数

参数	描述
CARBON TABLE	在其中执行更新操作的CarbonData表的名称。
column_name	待更新的目标列。
sourceColumn	需在目标表中更新的源表的列值。
sourceTable	将其记录更新到目标CarbonData表中的表。

注意事项

以下是使用UPDATE命令的条件：

- 如果源表中的多个输入行与目标表中的单行匹配，则UPDATE命令失败。
- 如果源表生成空记录，则UPDATE操作将在不更新表的情况下完成。
- 如果源表的行与目标表中任何已有的行不对应，则UPDATE操作将完成，不更新表。
- 具有二级索引的表不支持UPDATE命令。
- 在子查询中，如果源表和目标表相同，则UPDATE操作失败。
- 如果在UPDATE命令中使用的子查询包含聚合函数或group by子句，则UPDATE操作失败。

  例如，**update t_carbn01 a set (a.item_type_code, a.profit) = (select b.item_type_cd, sum(b.profit) from t_carbn01b b where item_type_cd =2 group by item_type_code);**

  其中，在子查询中使用聚合函数sum(b.profit)和group by子句，因此UPDATE操作失败。
- 如果查询的表设置了carbon.input.segments属性，则UPDATE操作失败。要解决该问题，在查询前执行以下语句。

  语法：

  SET carbon.input.segments. <database_name>. <table_name>=*;

示例

- 示例1：

  update carbonTable1 d set (d.column3,d.column5) = (select s.c33 ,s.c55 from sourceTable1 s where d.column1 = s.c11) where d.column1 = 'country' exists(select * from table3 o where o.c2 > 1);
- 示例2：

  update carbonTable1 d set (c3) = (select s.c33 from sourceTable1 s where d.column1 = s.c11) where exists(select * from iud.other o where o.c2 > 1);
- 示例3：

  update carbonTable1 set (c2, c5) = (c2 + 1, concat(c5 , "y"));
- 示例4：

  update carbonTable1 d set (c2, c5) = (c2 + 1, "xyx") where d.column1 = 'india';
- 示例5：

  update carbonTable1 d set (c2, c5) = (c2 + 1, "xyx") where d.column1 = 'india' and exists(select * from table3 o where o.column2 > 1);

系统响应

可在driver日志和客户端中查看命令运行成功或失败。

12.3.6.2.3 DELETE RECORDS from CARBON TABLE

命令功能

DELETE RECORDS命令从CarbonData表中删除记录。

命令格式

DELETE FROM CARBON_TABLE [WHERE expression];

参数描述

表 12-41 DELETE RECORDS 参数

参数	描述
CARBON TABLE	在其中执行删除操作的CarbonData表的名称。

注意事项

- 删除segment将删除相应segment的所有二级索引。
- 如果查询的表设置了carbon.input.segments属性，则DELETE操作失败。要解决该问题，在查询前执行以下语句。

  语法：

  SET carbon.input.segments. <database_name>.<table_name>=*;

示例

- 示例1：

  delete from columncarbonTable1 d where d.column1 = 'country';
- 示例2：

  delete from dest where column1 IN ('country1', 'country2');
- 示例3：

  delete from columncarbonTable1 where column1 IN (select column11 from sourceTable2);
- 示例4：

  delete from columncarbonTable1 where column1 IN (select column11 from sourceTable2 where column1 = '*');**
- 示例5：

  delete from columncarbonTable1 where column2 >= 4;

系统响应

可在driver日志和客户端中查看命令运行成功或失败。

12.3.6.2.4 INSERT INTO CARBON TABLE

命令功能

INSERT命令用于将SELECT查询结果加载到CarbonData表中。

命令格式

INSERT INTO [CARBON TABLE] [select query];

参数描述

表 12-42 INSERT INTO 参数

参数	描述
CARBON TABLE	需要执行INSERT命令的CarbonData表的名称。
select query	Source表上的SELECT查询（支持CarbonData、Hive和Parquet表）。

注意事项

- 表必须已经存在。
- 用户应属于数据加载组以执行数据加载操作。默认情况下，数据加载组被命名为“ficommon”。
- CarbonData表不支持Overwrite。
- 源表和目标表的数据类型应该相同，否则原表中的数据将被视为Bad Records。
- **INSERT INTO**命令不支持部分成功（partial success），如果存在Bad Records，该命令会失败。
- 在从源表插入数据到目标表的过程中，无法在源表中加载或更新数据。
  若要在INSERT操作期间启用数据加载或更新，请将以下参数配置为“true”。
  “carbon.insert.persist.enable”=“true”
  默认上述参数配置为“false”。

  说明

  启用该参数将降低INSERT操作的性能。

示例

create table carbon01(a int,b string,c string) stored as carbondata;

insert into table carbon01 values(1,'a','aa'),(2,'b','bb'),(3,'c','cc');

create table carbon02(a int,b string,c string) stored as carbondata;

INSERT INTO carbon02 select * from carbon01 where a > 1;

系统响应

可在driver日志中查看命令运行成功或失败。

12.3.6.2.5 DELETE SEGMENT by ID

命令功能

DELETE SEGMENT by ID命令是使用Segment ID来删除segment。

命令格式

DELETE FROM TABLE db_name.table_name WHERE SEGMENT.ID IN (segment_id1,segment_id2);

参数描述

表 12-43 DELETE LOAD 参数描述

参数	描述
segment_id	将要删除的Segment的ID。
db_name	Database名称，若未指定，则使用当前database。
table_name	在给定的database中的表名。

注意事项

流式表不支持删除segment。

示例

DELETE FROM TABLE CarbonDatabase.CarbonTable WHERE SEGMENT.ID IN (0);

DELETE FROM TABLE CarbonDatabase.CarbonTable WHERE SEGMENT.ID IN (0,5,8);

系统响应

操作成功或失败会在CarbonData日志中被记录。

12.3.6.2.6 DELETE SEGMENT by DATE

命令功能

DELETE SEGMENT by DATE命令用于通过加载日期删除CarbonData segment，在特定日期之前创建的segment将被删除。

命令格式

DELETE FROM TABLE db_name.table_name WHERE SEGMENT.STARTTIME BEFORE date_value;

参数描述

表 12-44 DELETE SEGMENT by DATE 参数描述

参数	描述
db_name	Database名称，若未指定，则使用当前database。
table_name	给定database中的表名。
date_value	有效Segment加载启动时间。在这个指定日期前的Segment将被删除。

注意事项

流式表不支持删除segment。

示例

DELETE FROM TABLE db_name.table_name WHERE SEGMENT.STARTTIME BEFORE '2017-07-01 12:07:20';

其中，STARTTIME是不同负载的加载启动时间。

系统响应

操作成功或失败会在CarbonData日志中被记录。

12.3.6.2.7 SHOW SEGMENTS

命令功能

SHOW SEGMENTS命令是用来向用户展示CarbonData table的Segment。

命令格式

SHOW SEGMENTS FOR TABLE *[db_name.]table_name* **LIMIT** *number_of_loads;*

参数描述

表 12-45 SHOW SEGMENTS FOR TABLE 参数描述

参数	描述
db_name	Database名，若未指定，则使用当前database。

参数	描述
table_name	在给定database中的表名。
number_of_loads	加载数的限制。

注意事项

无。

示例

create table carbon01(a int,b string,c string) stored as carbondata;

insert into table carbon01 select 1,'a','aa';

insert into table carbon01 select 2,'b','bb';

insert into table carbon01 select 3,'c','cc';

SHOW SEGMENTS FOR TABLE carbon01 LIMIT 2;

系统响应

```
+-----+----------+--------------------------+------------------+------------+------------+-------------+--------------+--
+
| ID  |  Status  |     Load Start Time      | Load Time Taken  | Partition  | Data Size  | Index Size  | File Format  |
+-----+----------+--------------------------+------------------+------------+------------+-------------+--------------+--
+
| 3   | Success  | 2020-09-28 22:53:26.336  | 3.726S           | {}         | 6.47KB     | 3.30KB      | columnar_v3  |
| 2   | Success  | 2020-09-28 22:53:01.702  | 6.688S           | {}         | 6.47KB     | 3.30KB      | columnar_v3  |
+-----+----------+--------------------------+------------------+------------+------------+-------------+--------------+--
+
```

12.3.6.2.8 CREATE SECONDARY INDEX

命令功能

该命令用于在CarbonData表中创建二级索引表。

命令格式

CREATE INDEX *index_name*

ON TABLE *[db_name.]table_name (col_name1, col_name2)*

AS *'carbondata'*

PROPERTIES *('table_blocksize'='256')*;

参数描述

表 12-46 CREATE SECONDARY INDEX 参数

参数	描述
index_name	索引表的名称。表名称应由字母数字字符和下划线（_）特殊字符组成。
db_name	数据库的名称。数据库名称应由字母数字字符和下划线（_）特殊字符组成。
table_name	数据库中的表名称。表名称应由字母数字字符和下划线（_）特殊字符组成。
col_name	表中的列名称。支持多列。列名称应由字母数字字符和下划线（_）特殊字符组成。
table_blocksize	数据文件的block大小。更多详细信息，请参考**•Block大小**。

注意事项

db_name为可选项。

示例

create table productdb.productSalesTable(id int,price int,productName string,city string) stored as carbondata;

CREATE INDEX productNameIndexTable on table productdb.productSalesTable (productName,city) as 'carbondata' ;

上述示例将创建名为“productdb.productNameIndexTable”的二级表并加载所提供列的索引信息。

系统响应

将创建二级索引表，加载与所提供的列相关的索引信息到二级索引表中，并将成功消息记录在系统日志中。

12.3.6.2.9 SHOW SECONDARY INDEXES

命令功能

该命令用于在所提供的CarbonData表中显示所有的二级索引表。

命令格式

SHOW INDEXES ON db_name.table_name;

参数描述

表 12-47 SHOW SECONDARY INDEXES 参数

参数	描述
db_name	数据库的名称。数据库名称应由字母数字字符和下划线（_）特殊字符组成
table_name	数据库中的表名称。表名称应由字母数字字符和下划线（_）特殊字符组成。

注意事项

db_name为可选项。

示例

create table productdb.productSalesTable(id int,price int,productName string,city string) stored as carbondata;

CREATE INDEX productNameIndexTable on table productdb.productSalesTable (productName,city) as 'carbondata' ;

SHOW INDEXES ON productdb.productSalesTable;

系统响应

显示列出给定CarbonData表中的所有索引表和相应的索引列。

12.3.6.2.10 DROP SECONDARY INDEX

命令功能

该命令用于删除给定表中存在的二级索引表。

命令格式

DROP INDEX [IF EXISTS] index_name ON [db_name.]table_name;

参数描述

表 12-48 DROP SECONDARY INDEX 参数

参数	描述
index_name	索引表的名称。表名称应由字母数字字符和下划线（_）特殊字符组成。
db_name	数据库的名称。若未指定，选择当前默认数据库。
table_name	需要删除的表的名称。

注意事项

该命令中IF EXISTS和db_name为可选项。

示例

DROP INDEX if exists productNameIndexTable ON productdb.productSalesTable;

系统响应

二级索引表将被删除，索引信息将在CarbonData表中被清除，删除成功的消息将记录在系统日志中。

12.3.6.2.11 CLEAN FILES

命令功能

DELETE SEGMENT命令会将被删除的segments标识为delete状态；segment合并后，旧的segments状态会变为compacted。这些segments的数据文件不会从物理上删除。如果用户希望强制删除这些文件，可以使用**CLEAN FILES**命令。

但是，使用该命令可能会导致查询命令执行失败。

命令格式

CLEAN FILES FOR TABLE *[db_name.]table_name* ;

参数描述

表 12-49 CLEAN FILES FOR TABLE 参数描述

参数	描述
db_name	数据库名称。数据库名称由字母，数字和下划线组成。
table_name	数据库中的表的名称。表名由字母，数字和下划线组成。

注意事项

无。

示例

添加carbon配置参数

```
carbon.clean.file.force.allowed = true
```

create table carbon01(a int,b string,c string) stored as carbondata;

insert into table carbon01 select 1,'a','aa';

insert into table carbon01 select 2,'b','bb';

delete from table carbon01 where segment.id in (0);

show segments for table carbon01;

CLEAN FILES FOR TABLE carbon01 options('force'='true');

show segments for table carbon01;

上述命令将从物理上删除所有DELETE SEGMENT命令删除的segment和合并后的旧的segment。

系统响应

可在driver日志中查看命令运行成功或失败。

12.3.6.2.12 SET/RESET

命令功能

此命令用于动态Add，Update，Display或Reset CarbonData参数，而无需重新启动driver。

命令格式

- Add或Update参数值：
  SET *parameter_name=parameter_value*
  此命令用于添加或更新“parameter_name”的值。
- Display参数值：
  SET *parameter_name*
  此命令用于显示指定的“parameter_name”的值。
- Display会话参数：
  SET
  此命令显示所有支持的会话参数。
- Display会话参数以及使用细节：
  SET -v
  此命令显示所有支持的会话参数及其使用细节。
- Reset参数值：
  RESET
  此命令清除所有会话参数。

参数描述

表 12-50 SET 参数描述

参数	描述
parameter_name	其值需要被动态添加（add），更新（update）或显示（display）的参数名称。

参数	描述
parameter_value	将要设置的“parameter_name”的新值。

注意事项

以下为分别使用SET和RESET命令进行动态设置或清除操作的属性：

表 12-51 属性描述

属性	描述
carbon.options.bad.records.logger.enable	启用或禁用bad record日志记录。
carbon.options.bad.records.action	指定bad record操作，例如，强制（force），重定向（redirect），失败（fail）或忽略（ignore）。有关详细信息，请参阅**•Bad Records处理：**
carbon.options.is.empty.data.bad.record	指定空数据是否被视为bad record。有关详细信息，请参阅 **Bad Records处理：**
carbon.options.sort.scope	指定数据加载期间排序的范围。
carbon.options.bad.record.path	指定需要存储bad record的HDFS路径。
carbon.custom.block.distribution	指定是否使用Spark或CarbonData的块分布功能。
enable.unsafe.sort	指定在数据加载期间是否使用不安全的排序。不安全的排序可减少数据加载操作期间的垃圾回收，从而实现更好的性能。
carbon.si.lookup.partialstring	当参数设置为TRUE时，二级索引采用starts-with、ends-with、contains和LIKE分区条件字符串。 当参数设置为FALSE时，二级索引只采用starts-with分区条件字符串。

属性	描述
carbon.input.segments	指定要查询的段ID。 此属性允许您查询指定表的指定段。CarbonScan将仅从指定的段ID读取数据。 语法： “carbon.input.segments.<database_name>. <table_name> = < list of segment ids >” 如果用户想在多线程模式下查询指定段，可使用**CarbonSession.threadSet**代替**SET**语句。 语法： “CarbonSession.threadSet ("carbon.input.segments.<database_name>. <table_name>","< list of segment ids >");” **说明** 不建议在**carbon.properties**文件中设置该属性，因为所有会话都包含段列表，除非发生会话级或线程级覆盖。

示例

- 添加（Add）或更新（Update）：

  SET *enable.unsafe.sort=true*
- 显示（Display）属性值：

  SET *enable.unsafe.sort*
- 显示段ID列表，段状态和其他所需详细信息的示例，然后指定要读取的段列表：

  SHOW SEGMENTS FOR *TABLE carbontable1;*

  SET *carbon.input.segments.db.carbontable1 = 1, 3, 9;*
- 多线程模式查询指定段示例如下：

  CarbonSession.threadSet (*"carbon.input.segments.default.carbon_table_MulTI_THread", "1,3"*);
- 在多线程环境中使用**CarbonSession.threadSet**查询段示例如下（以Scala代码为例）：

```
def main(args: Array[String]) {
 Future
{              CarbonSession.threadSet("carbon.input.segments.default.carbon_table_MulTI_THread", "1")
     spark.sql("select count(empno) from carbon_table_MulTI_THread").show()
   }
}
```

- 重置（Reset）：

  RESET

系统响应

- 若运行成功，将记录在driver日志中。
- 若出现故障，将显示在用户界面（UI）中。

12.3.6.3 操作并发

DDL和DML中的操作，执行前，需要获取对应的锁，各操作需要获取锁的情况见表1 操作获取锁一览表，√表示需要获取该锁，一个操作仅在获取到所有需要获取的锁后，才能继续执行。

任意两个操作是否可以并发执行，可以通过如下方法确定：表12-52两行代表两个操作，这两行没有任意一列都标记√，即不存在某一列两行全为√。

表 12-52 操作获取锁一览表

操作	METADATA_LOCK	COMPACTION_LOCK	DROP_TABLE_LOCK	DELETE_SEGMENT_LOCK	CLEAN_FILES_LOCK	ALTER_PARTITION_LOCK	UPDATE_LOCK	STREAMING_LOCK	CONCURRENT_LOAD_LOCK	SEGMENT_LOCK
CREATE TABLE	-	-	-	-	-	-	-	-	-	-
CREATE TABLE As SELECT	-	-	-	-	-	-	-	-	-	-
DROP TABLE	√	-	√	-	-	-	-	√	-	-
ALTER TABLE COMPACTION	-	√	-	-	-	-	√	-	-	-
TABLE RENAME	-	-	-	-	-	-	-	-	-	-

操作	METADATA_LOCK	COMPACTION_LOCK	DROP_TABLE_LOCK	DELETE_SEGMENT_LOCK	CLEAN_FILES_LOCK	ALTER_PARTITION_LOCK	UPDATE_LOCK	STREAMING_LOCK	CONCURRENT_LOAD_LOCK	SEGMENT_LOCK
ADD COLUMNS	√	√	-	-	-	-	-	-	-	-
DROP COLUMNS	√	√	-	-	-	-	-	-	-	-
CHANGE DATA TYPE	√	√	-	-	-	-	-	-	-	-
REFRESH TABLE	-	-	-	-	-	-	-	-	-	-
REGISTER INDEX TABLE	√	-	-	-	-	-	-	-	-	-
REFRESH INDEX	-	√	-	-	-	-	-	-	-	-
LOAD DATA/ INSERT INTO	-	-	-	-	-	-	-	-	√	√
UPDATE CARBON TABLE	√	√	-	-	-	-	√	-	-	-

操作	METADATA_LOCK	COMPACTION_LOCK	DROP_TABLE_LOCK	DELETE_SEGMENT_LOCK	CLEAN_FILES_LOCK	ALTER_PARTITION_LOCK	UPDATE_LOCK	STREAMING_LOCK	CONCURRENT_LOAD_LOCK	SEGMENT_LOCK
DELETE RECORDS from CARBON TABLE	√	√	-	-	-	-	√	-	-	-
DELETE SEGMENT by ID	-	-	-	√	√	-	-	-	-	-
DELETE SEGMENT by DATE	-	-	-	√	√	-	-	-	-	-
SHOW SEGMENTS	-	-	-	-	-	-	-	-	-	-
CREATE SECONDARY INDEX	√	√	-	√	-	-	-	-	-	-
SHOW SECONDARY INDEXES	-	-	-	-	-	-	-	-	-	-

操作	METADATA_LOCK	COMPACTION_LOCK	DROP_TABLE_LOCK	DELETE_SEGMENT_LOCK	CLEAN_FILES_LOCK	ALTER_PARTITION_LOCK	UPDATE_LOCK	STREAMING_LOCK	CONCURRENT_LOAD_LOCK	SEGMENT_LOCK
DROP SECONDARY INDEX	√	-	√	-	-	-	-	-	-	-
CLEAN FILES	-	-	-	-	-	-	-	-	-	-
SET/RESET	-	-	-	-	-	-	-	-	-	-
Add Hive Partition	-	-	-	-	-	-	-	-	-	-
Drop Hive Partition	√	√	√	√	√	√	-	-	-	-
Drop Partition	√	√	√	√	√	√	-	-	-	-
Alter table set	√	√	-	-	-	-	-	-	-	-

12.3.6.4 API

本章节描述Segment的API以及使用方法，所有方法在org.apache.spark.util.CarbonSegmentUtil类中。

如下方法已废弃：

```
/**
* Returns the valid segments for the query based on the filter condition
* present in carbonScanRdd.
*
* @param carbonScanRdd
* @return Array of valid segments
*/
@deprecated def getFilteredSegments(carbonScanRdd: CarbonScanRDD[InternalRow]): Array[String];
```

使用方法

使用如下方法从查询语句中获得CarbonScanRDD：

```
val df=carbon.sql("select * from table where age='12'")
val myscan=df.queryExecution.sparkPlan.collect {
case scan: CarbonDataSourceScan if scan.rdd.isInstanceOf[CarbonScanRDD[InternalRow]] => scan.rdd
case scan: RowDataSourceScanExec if scan.rdd.isInstanceOf[CarbonScanRDD[InternalRow]] => scan.rdd
}.head
val carbonrdd=myscan.asInstanceOf[CarbonScanRDD[InternalRow]]
```

例子：

```
CarbonSegmentUtil.getFilteredSegments(carbonrdd)
```

可以通过传入sql语句来获取过滤后的segment：

```
/**
* Returns an array of valid segment numbers based on the filter condition provided in the sql
* NOTE: This API is supported only for SELECT Sql (insert into,ctas,.., is not supported)
*
* @param sql
* @param sparkSession
* @return Array of valid segments
* @throws UnsupportedOperationException because Get Filter Segments API supports if and only
* if only one carbon main table is present in query.
*/
def getFilteredSegments(sql: String, sparkSession: SparkSession): Array[String];
```

例子：

```
CarbonSegmentUtil.getFilteredSegments("select * from table where age='12'”, sparkSession)
```

传入数据库名和表名，获取会被合并的segment列表，得到的segment列表可以当做getMergedLoadName函数的参数传入：

```
/**
* Identifies all segments which can be merged with MAJOR compaction type.
* NOTE: This result can be passed to getMergedLoadName API to get the merged load name.
*
* @param sparkSession
* @param tableName
* @param dbName
* @return list of LoadMetadataDetails
*/
def identifySegmentsToBeMerged(sparkSession: SparkSession,
tableName: String,
dbName: String) : util.List[LoadMetadataDetails];
```

例子：

```
CarbonSegmentUtil.identifySegmentsToBeMerged(sparkSession, "table_test","default")
```

传入数据库名、表名和自定义的segment列表，获取自定义合并操作会被合并的segment列表，得到的segment列表可以当做getMergedLoadName函数的参数传入：

```
/**
* Identifies all segments which can be merged with CUSTOM compaction type.
*  NOTE: This result can be passed to getMergedLoadName API to get the merged load name.
*
* @param sparkSession
* @param tableName
* @param dbName
* @param customSegments
* @return list of LoadMetadataDetails
* @throws UnsupportedOperationException if customSegments is null or empty.
* @throws MalformedCarbonCommandException if segment does not exist or is not valid
*/
def identifySegmentsToBeMergedCustom(sparkSession: SparkSession,
tableName: String,
```

```
dbName: String,
customSegments: util.List[String]): util.List[LoadMetadataDetails];
```

例子：

```
val customSegments = new util.ArrayList[String]()
customSegments.add("1")
customSegments.add("2")
CarbonSegmentUtil.identifySegmentsToBeMergedCustom(sparkSession, "table_test","default",
customSegments)
```

给定segment列表，返回合并后新的导入名称：

```
/**
* Returns the Merged Load Name for given list of segments
*
* @param list of segments
* @return Merged Load Name
* @throws UnsupportedOperationException if list of segments is less than 1
*/
def getMergedLoadName(list: util.List[LoadMetadataDetails]): String;
```

例子：

```
val carbonTable = CarbonEnv.getCarbonTable(Option(databaseName), tableName)(sparkSession)
val loadMetadataDetails = SegmentStatusManager.readLoadMetadata(carbonTable.getMetadataPath)
CarbonSegmentUtil.getMergedLoadName(loadMetadataDetails.toList.asJava)
```

12.3.6.5 空间索引

快速示例

```
create table IF NOT EXISTS carbonTable
(
COLUMN1    BIGINT,
LONGITUDE    BIGINT,
LATITUDE    BIGINT,
COLUMN2    BIGINT,
COLUMN3    BIGINT
)
STORED AS carbondata
TBLPROPERTIES
('SPATIAL_INDEX.mygeohash.type'='geohash','SPATIAL_INDEX.mygeohash.sourcecolumns'='longitude,
latitude','SPATIAL_INDEX.mygeohash.originLatitude'='39.850713','SPATIAL_INDEX.mygeohash.gridSize'='50','S
PATIAL_INDEX.mygeohash.minLongitude'='115.828503','SPATIAL_INDEX.mygeohash.maxLongitude'='720.000
000','SPATIAL_INDEX.mygeohash.minLatitude'='39.850713','SPATIAL_INDEX.mygeohash.maxLatitude'='720.0
00000','SPATIAL_INDEX'='mygeohash','SPATIAL_INDEX.mygeohash.conversionRatio'='1000000','SORT_COLU
MNS'='column1,column2,column3,latitude,longitude');
```

空间索引介绍

空间数据包括多维点、线、矩形、立方体、多边形和其他几何对象。空间数据对象占据空间的某一区域，称为空间范围，通过其位置和边界描述。空间数据可以是点数据，也可以是区域数据。

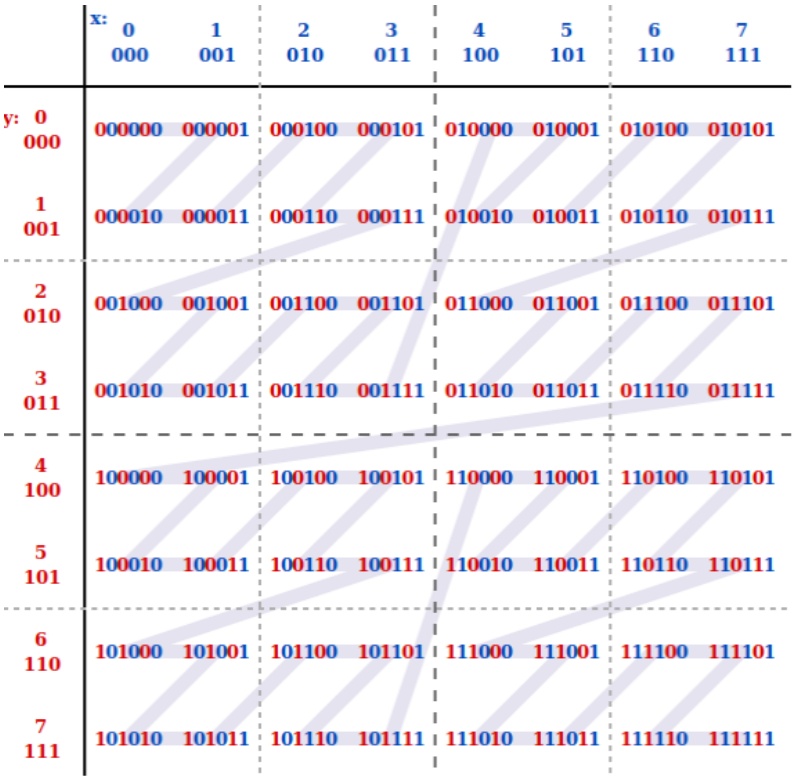
- 点数据：一个点具有一个空间范围，仅通过其位置描述。它不占用空间，没有相关的边界。点数据由二维空间中的点的集合组成。点可以存储为一对经纬度。
- 区域数据：一个区域有空间范围，有位置和边界。位置可以看作是一个定点在区域内的位置，例如它的质心。在二维中，边界可以可视化为一条线（有限区域，闭环）。区域数据包含一系列区域。

目前仅限于支持点数据，存储点数据。

经纬度可以编码为唯一的GeoID。Geohash是Gustavo Niemeyer发明的公共域地理编码系统，它将地理位置编码为一串由字母和数字组成的短字符串。它是一种分层的空

间数据结构，把空间细分为网格形状的桶，是被称为Z阶曲线和通常称为空间填充曲线的许多应用之一。

点在多维中的Z值是简单地通过交织其坐标值的二进制表示来计算的，如下图所示。使用Geohash创建GeoID时，数据按照GeoID排序，而不是按照经纬度排序，数据按照空间就近性排序存储。



建表

GeoHash编码：

```
create table IF NOT EXISTS carbonTable
(
...
`LONGITUDE`     BIGINT,
`LATITUDE`      BIGINT,
...
)
STORED AS carbondata
TBLPROPERTIES
('SPATIAL_INDEX.mygeohash.type'='geohash','SPATIAL_INDEX.mygeohash.sourcecolumns'='longitude,
latitude','SPATIAL_INDEX.mygeohash.originLatitude'='xx.xxxxxx','SPATIAL_INDEX.mygeohash.gridSize'='xx','SP
ATIAL_INDEX.mygeohash.minLongitude'='xxx.xxxxxx','SPATIAL_INDEX.mygeohash.maxLongitude'='xxx.xxxxxx'
,'SPATIAL_INDEX.mygeohash.minLatitude'='xx.xxxxxx','SPATIAL_INDEX.mygeohash.maxLatitude'='xxx.xxxxxx','
SPATIAL_INDEX'='mygeohash','SPATIAL_INDEX.mygeohash.conversionRatio'='1000000','SORT_COLUMNS'='co
lumn1,column2,column3,latitude,longitude');
```

SPATIAL_INDEX：自定义索引处理器。此处理程序允许用户从表结构列集合中创建新的列。新创建的列名与处理程序名相同。处理程序的type和sourcecolumns属性是必需的属性。目前，type属性只支持“geohash”。Carbon提供一个简单的默认实现类。用户可以通过扩展默认实现类来挂载geohash的自定义实现类。该默认处理程序还需提供以下的表属性：

- SPATIAL_INDEX.xxx.originLatitude：Double类型，坐标原点纬度
- SPATIAL_INDEX.xxx.gridSize：Int类型，栅格长度（米）
- SPATIAL_INDEX.xxx.minLongitude：Double类型，最小经度
- SPATIAL_INDEX.xxx.maxLongitude：Double类型，最大经度
- SPATIAL_INDEX.xxx.minLatitude：Double类型，最小纬度
- SPATIAL_INDEX.xxx.maxLatitude：Double类型，最大纬度
- SPATIAL_INDEX.xxx.conversionRatio：Int类型，将经纬度小数值转换为整型值

用户可以按照上述格式为处理程序添加自己的表属性，并在自定义实现类中访问它们。originLatitude，gridSize及conversionRatio是必选参数，其余属性在Carbon中都是可选的。可以使用“SPATIAL_INDEX.xxx.class”属性指定它们的实现类。

默认实现类可以为每一行的sourcecolumns生成handler列值，并且支持基于sourcecolumns的过滤条件查询。生成的handler列对用户不可见。除SORT_COLUMNS表属性外，任何DDL命令和属性都不允许包含handler列。

说明

- 生成的handler列默认被视为排序列。如果SORT_COLUMNS不包含任何sourcecolumns，则将handler列追加到现有的SORT_COLUMNS最后。如果在SORT_COLUMNS中已经指定了该handler列，则它在SORT_COLUMNS的顺序将保持不变。
- 如果SORT_COLUMNS包含任意的sourcecolumns，但是没有包含handler列，则handler列将自动插入到SORT_COLUMNS中的sourcecolumns之前。
- 如果SORT_COLUMNS需要包含任意的sourcecolumns，那么需要保证handler列出现在sourcecolumns之前，这样handler列才能在排序中生效。

GeoSOT编码：

```
CREATE TABLE carbontable(
…
longitude DOUBLE,
latitude DOUBLE,
…)
STORED AS carbondata
TBLPROPERTIES ('SPATIAL_INDEX'='xxx',
'SPATIAL_INDEX.xxx.type'='geosot',
'SPATIAL_INDEX.xxx.sourcecolumns'='longitude, latitude',
'SPATIAL_INDEX.xxx.level'='21',
'SPATIAL_INDEX.xxx.class'='org.apache.carbondata.geo.GeoSOTIndex')
```

表 12-53 参数说明

参数	说明
SPATIAL_INDEX	指定表属性”SPATIAL_INDEX”，空间索引列，列名与该属性的值相同。
SPATIAL_INDEX.xxx.type	必填参数，值为geosot。
SPATIAL_INDEX.xxx.sourcecolumns	必填参数，空间索引列属性，指定计算空间索引的源数据列，需为2个存在的列，且类型为double。
SPATIAL_INDEX.xxx.level	可选参数，用于计算空间索引列。默认值为17，因为该值可以计算出足够精确的结果，同时拥有良好的性能。

参数	说明
SPATIAL_INDEX.xxx.class	可选参数，用于指定geo的实现类，默认为“org.apache.carbondata.geo.GeoSOTIndex”。

使用示例：

```
create table geosot(
timevalue bigint,
longitude double,
latitude double)
stored as carbondata
TBLPROPERTIES ('SPATIAL_INDEX'='mygeosot',
'SPATIAL_INDEX.mygeosot.type'='geosot',
'SPATIAL_INDEX.mygeosot.level'='21', 'SPATIAL_INDEX.mygeosot.sourcecolumns'='longitude, latitude');
```

准备数据

- 准备数据文件1：geosotdata.csv

```
timevalue,longitude,latitude
1575428400000,116.285807,40.084087
1575428400000,116.372142,40.129503
1575428400000,116.187332,39.979316
1575428400000,116.337069,39.951887
1575428400000,116.359102,40.154684
1575428400000,116.736367,39.970323
1575428400000,116.720179,40.009893
1575428400000,116.346961,40.13355
1575428400000,116.302895,39.930753
1575428400000,116.288955,39.999101
1575428400000,116.17609,40.129953
1575428400000,116.725575,39.981115
1575428400000,116.266922,40.179415
1575428400000,116.353706,40.156483
1575428400000,116.362699,39.942444
1575428400000,116.325378,39.963129
```

- 准备数据文件2：geosotdata2.csv

```
timevalue,longitude,latitude
1575428400000,120.17708,30.326882
1575428400000,120.180685,30.326327
1575428400000,120.184976,30.327105
1575428400000,120.189311,30.327549
1575428400000,120.19446,30.329698
1575428400000,120.186965,30.329133
1575428400000,120.177481,30.328911
1575428400000,120.169713,30.325614
1575428400000,120.164563,30.322243
1575428400000,120.171558,30.319613
1575428400000,120.176365,30.320687
1575428400000,120.179669,30.323688
1575428400000,120.181001,30.320761
1575428400000,120.187094,30.32354
1575428400000,120.193574,30.323651
1575428400000,120.186192,30.320132
1575428400000,120.190055,30.317464
1575428400000,120.195376,30.318094
1575428400000,120.160786,30.317094
1575428400000,120.168211,30.318057
1575428400000,120.173618,30.316612
1575428400000,120.181001,30.317316
1575428400000,120.185162,30.315908
1575428400000,120.192415,30.315871
1575428400000,120.161902,30.325614
1575428400000,120.164306,30.328096
```

```
1575428400000,120.197093,30.325985
1575428400000,120.19602,30.321651
1575428400000,120.198638,30.32354
1575428400000,120.165421,30.314834
```

导入数据

GeoHash默认实现类扩展自定义索引抽象类。如果没有配置handler属性为自定义的实现类，则使用默认的实现类。用户可以通过扩展默认实现类来挂载geohash的自定义实现类。自定义索引抽象类方法包括：

- Init方法，用来提取、验证和存储handler属性。在失败时抛出异常，并显示错误信息。
- Generate方法，用来生成索引。它为每行数据生成一个索引数据。
- Query方法，用来对给定输入生成索引值范围列表。

导入命令同普通Carbon表：

LOAD DATA inpath '/tmp/*geosotdata.csv*' INTO TABLE geosot OPTIONS ('DELIMITER'= ',');

LOAD DATA inpath '/tmp/*geosotdata2.csv*' INTO TABLE geosot OPTIONS ('DELIMITER'= ',');

说明

geosotdata.csv和geosotdata2.csv表请参考**准备数据**。

不规则空间集合的聚合查询

查询语句及Filter UDF

- 根据polygon过滤数据

  IN_POLYGON(pointList)

  UDF输入参数：

参数	类型	说明
pointList	String	将多个点输入为一个字符串，每个点以**longitude latitude**表示。经纬度间用空格分隔，每对经纬度用逗号分隔，字符串首尾经纬度一致。

  UDF输出参数：

参数	类型	说明
inOrNot	Boolean	判断数据是否在指定的polygon_list之内。

  使用示例：

```
select longitude, latitude from geosot where IN_POLYGON('116.321011 40.123503, 116.137676 39.947911, 116.560993 39.935276, 116.321011 40.123503');
```

- 根据polygon列表过滤数据。

  IN_POLYGON_LIST(polygonList, opType)

  UDF输入参数：

参数	类型	说明
polygonList	String	将多个polygon输入为一个字符串，每个polygon以**POLYGON ((longitude1 latitude1, longitude2 latitude2, …))**表示。注意“POLYGON”后有空格，经纬度间用空格分隔，每对经纬度用逗号分隔，一个polygon的首尾经纬度一致。IN_POLYGON_LIST必须输入2个以上polygon。 一个polygon示例： POLYGON ((116.137676 40.163503, 116.137676 39.935276, 116.560993 39.935276, 116.137676 40.163503))
opType	String	对多个polygon进行并交差操作。 目前支持的操作类型： ● OR：A ∪ B ∪ C (假设输入了三个POLYGON，A、B、C) ● AND：A ∩ B ∩ C

  UDF输出参数：

参数	类型	说明
inOrNot	Boolean	判断数据是否在指定的polygon_list之内。

  使用示例：

```
select longitude, latitude from geosot where IN_POLYGON_LIST('POLYGON ((120.176433 30.327431,120.171283 30.322245,120.181411 30.314540, 120.190509 30.321653,120.185188 30.329358,120.176433 30.327431)), POLYGON ((120.191603 30.328946,120.184179 30.327465,120.181819 30.321464, 120.190359 30.315388,120.199242 30.324464,120.191603 30.328946))', 'OR');
```

- 根据polyline列表过滤数据。

  IN_POLYLINE_LIST(polylineList, bufferInMeter)

  UDF输入参数：

参数	类型	说明
polylineList	String	将多个polyline输入为一个字符串，每个polyline以**LINESTRING (longitude1 latitude1, longitude2 latitude2, …)**表示。注意“LINESTRING”后有空格，经纬度间用空格分隔，每组经纬度用逗号分隔。 对多个polyline区域内的数据会输出并集结果。 一个polyline示例： LINESTRING (116.137676 40.163503, 116.137676 39.935276, 116.260993 39.935276)
bufferInMeter	Float	polyline的buffer距离，单位为米。末端使用直角创建缓冲区。

UDF输出参数：

参数	类型	说明
inOrNot	Boolean	判断数据是否在指定的polyline_list之内。

使用示例：

```
select longitude, latitude from geosot where IN_POLYLINE_LIST('LINESTRING (120.184179 30.327465, 120.191603 30.328946, 120.199242 30.324464, 120.190359 30.315388)', 65);
```

- 根据GeoId区间列表过滤数据。

  IN_POLYGON_RANGE_LIST(polygonRangeList, opType)

  UDF输入参数：

参数	类型	说明
polygonRangeList	String	将多个rangeList输入为一个字符串，每个rangeList以**RANGELIST (startGeoId1 endGeoId1, startGeoId2 endGeoId2, …)**表示。注意“RANGELIST”后有空格，首尾GeoId间用空格分隔，每组GeoId range用逗号分隔。 一个rangeList示例： RANGELIST (855279368848 855279368850, 855280799610 855280799612, 855282156300 855282157400)

参数	类型	说明
opType	String	对多个rangeList进行并交差操作。 目前支持的操作类型： ● OR：A ∪ B ∪ C (假设输入了三个RANGELIST，A、B、C) ● AND：A ∩ B ∩ C

UDF输出参数：

参数	类型	说明
inOrNot	Boolean	判断数据是否在指定的polyRange_list之内。

使用示例：

```
select mygeosot, longitude, latitude from geosot where IN_POLYGON_RANGE_LIST('RANGELIST
(526549722865860608 526549722865860618, 532555655580483584 532555655580483594)', 'OR');
```

- polygon连接查询

  IN_POLYGON_JOIN(GEO_HASH_INDEX_COLUMN, POLYGON_COLUMN)

  两张表做join查询，一张表为空间数据表（有经纬度列和GeoHashIndex列），另一张表为维度表，保存polygon数据。

  查询使用IN_POLYGON_JOIN UDF，参数GEO_HASH_INDEX_COLUMN和polygon表的POLYGON_COLUMN。Polygon_column列是一系列的点（经纬度列）。Polygon表的每一行的第一个点和最后一个点必须是相同的。Polygon表的每一行的所有点连接起来形成一个封闭的几何对象。

  UDF输入参数：

参数	类型	说明
GEO_HASH_INDEX_COLUMN	Long	空间数据表的GeoHashIndex列。
POLYGON_COLUMN	String	Polygon表的polygon列，数据为polygon的字符串表示。例如，一个polygon是POLYGON ((longitude1 latitude1, longitude2 latitude2, …))

使用示例：

```
CREATE TABLE polygonTable(
polygon string,
poiType string,
poiId String)
STORED AS carbondata;

insert into polygonTable select 'POLYGON ((120.176433 30.327431,120.171283 30.322245,
120.181411 30.314540,120.190509 30.321653,120.185188 30.329358,120.176433 30.327431))','abc','1';
```

```
insert into polygonTable select 'POLYGON ((120.191603 30.328946,120.184179 30.327465,
120.181819 30.321464,120.190359 30.315388,120.199242 30.324464,120.191603 30.328946))','abc','2';

select t1.longitude,t1.latitude from geosot t1
inner join
(select polygon,poiId from polygonTable where poitype='abc') t2
on in_polygon_join(t1.mygeosot,t2.polygon) group by t1.longitude,t1.latitude;
```

- range_list连接查询

  IN_POLYGON_JOIN_RANGE_LIST(GEO_HASH_INDEX_COLUMN, POLYGON_COLUMN)

  同IN_POLYGON_JOIN，使用IN_POLYGON_JOIN_RANGE_LIST UDF关联空间数据表和polygon维度表，关联基于Polygon_RangeList。直接使用range list可以避免polygon到range list的转换。

  UDF输入参数：

参数	类型	说明
GEO_HASH_INDEX_COLUMN	Long	空间数据表的GeoHashIndex列。
POLYGON_COLUMN	String	Polygon表的rangelist列，数据为rangeList的字符串。例如，一个rangelist是RANGELIST (startGeoId1 endGeoId1, startGeoId2 endGeoId2, …)

  使用示例：

```
CREATE TABLE polygonTable(
polygon string,
poiType string,
poiId String)
STORED AS carbondata;

insert into polygonTable select 'RANGELIST (526546455897309184 526546455897309284,
526549831217315840 526549831217315850, 532555655580483534 532555655580483584)','xyz','2';

select t1.*
from geosot t1
inner join
(select polygon,poiId from polygonTable where poitype='xyz') t2
on in_polygon_join_range_list(t1.mygeosot,t2.polygon);
```

空间索引工具类UDF

- GeoID转栅格行列号。

  GeoIdToGridXy(geoId)

  UDF输入参数：

参数	类型	说明
geoId	Long	根据GeoId计算栅格行列号。

  UDF输出参数：

参数	类型	说明
gridArray	Array[Int]	返回该geoid所包含的栅格行列号，以数组的方式返回，第一位为行，第二位为列。

使用示例：

```
select longitude, latitude, mygeohash, GeoIdToGridXy(mygeohash) as GridXY from geoTable;
```

- 经纬度转GeoID。

  LatLngToGeoId(latitude, longitude oriLatitude, gridSize)

  UDF输入参数：

参数	类型	说明
longitude	Long	经度，注：转换后的整数类型。
latitude	Long	纬度，注：转换后的整数类型。
oriLatitude	Double	原点纬度，计算GeoId需要参数。
gridSize	Int	栅格大小，计算GeoId需要参数。

  UDF输出参数：

参数	类型	说明
geoId	Long	通过编码获得一个表示经纬度的数。

  使用示例：

```
select longitude, latitude, mygeohash, LatLngToGeoId(latitude, longitude, 39.832277, 50) as geoId
from geoTable;
```

- GeoID转经纬度。

  GeoIdToLatLng(geoId, oriLatitude, gridSize)

  UDF输入参数：

参数	类型	说明
geoId	Long	根据GeoId计算经纬度。
oriLatitude	Double	原点纬度，计算经纬度需要参数。
gridSize	Int	栅格大小，计算经纬度需要参数。

说明

由于GeoId由栅格坐标生成，坐标为栅格中心点，则计算出的经纬度是栅格中心点经纬度，与生成该GeoId的经纬度可能有[0度~半个栅格度数]的误差。

UDF输出参数：

参数	类型	说明
latitudeAndLongitude	Array[Double]	返回该geoid所表示的栅格的中心点的经纬度坐标，以数组的方式返回，第一位为latitude，第二位为longitude。

使用示例：

```
select longitude, latitude, mygeohash, GeoIdToLatLng(mygeohash, 39.832277, 50) as LatitudeAndLongitude from geoTable;
```

- 计算金字塔模型向上汇聚一层的GeoID。

  ToUpperLayerGeoId(geoId)

  UDF输入参数：

参数	类型	说明
geoId	Long	根据输入GeoId计算金字塔模型上一层GeoId。

  UDF输出参数：

参数	类型	说明
geoId	Long	金字塔模型上一层GeoId。

  使用示例：

```
select longitude, latitude, mygeohash, ToUpperLayerGeoId(mygeohash) as upperLayerGeoId from geoTable;
```

- 输入polygon获得GeoID范围列表。

  ToRangeList(polygon, oriLatitude, gridSize)

  UDF输入参数：

参数	类型	说明
polygon	String	输入polygon字符串，用一组经纬度表示。 经纬度间用空格分隔，每对经纬度间用逗号分隔，首尾经纬度一致。
oriLatitude	Double	原点纬度，计算GeoId需要参数。
gridSize	Int	栅格大小，计算GeoId需要参数。

UDF输出参数：

参数	类型	说明
geoIdList	Buffer[Array[Long]]	将polygon转换为一串geoid的范围列表。

使用示例：

```
select ToRangeList('116.321011 40.123503, 116.137676 39.947911, 116.560993 39.935276,
116.321011 40.123503', 39.832277, 50) as rangeList from geoTable;
```

- 计算金字塔模型向上汇聚一层的longitude。

  ToUpperLongitude (longitude, gridSize, oriLat)

  UDF输入参数：

参数	类型	说明
longitude	Long	输入longitude，用一个长整型表示。
gridSize	Int	栅格大小，计算longitude需要参数。
oriLatitude	Double	原点纬度，计算longitude需要参数。

  UDF输出参数：

参数	类型	说明
longitude	Long	返回上一层的longitude。

  使用示例：

```
select ToUpperLongitude (-23575161504L, 50, 39.832277) as upperLongitude from geoTable;
```

- 计算金字塔模型向上汇聚一层的Latitude。

  ToUpperLatitude(Latitude, gridSize, oriLat)

  UDF输入参数：

参数	类型	说明
latitude	Long	输入latitude，用一个长整型表示。
gridSize	Int	栅格大小，计算latitude需要参数。
oriLatitude	Double	原点纬度，计算latitude需要参数。

  UDF输出参数：

参数	类型	说明
Latitude	Long	返回上一层的latitude。

使用示例：

```
select ToUpperLatitude (-23575161504L, 50, 39.832277) as upperLatitude from geoTable;
```

- 经纬度转GeoSOT

  LatLngToGridCode(latitude, longitude, level)

  UDF输入参数：

参数	类型	说明
latitude	Double	输入latitude。
longitude	Double	输入longitude。
level	Int	输入level，值区间[0-32]。

UDF输出参数：

参数	类型	说明
geoId	Long	通过GeoSOT编码获得一个表示经纬度的数。

使用示例：

```
select LatLngToGridCode(39.930753, 116.302895, 21) as geoId;
```

12.3.7 CarbonData 故障处理

12.3.7.1 当在 Filter 中使用 Big Double 类型数值时，过滤结果与 Hive 不一致

现象描述

当在filter中使用更高精度的double数据类型的数值时，过滤结果没有按照所使用的filter的要求返回正确的值。

可能原因

如果filter使用更高精度的double数据类型的数值，系统将会对该值四舍五入进行比较，因此在这种情况下，即使小数部分不同，系统仍然会认为double数据类型的值是相同的。

定位思路

无。

处理步骤

当需要高精度的数据比较时，可以使用Decimal数据类型的数值，例如，在财务应用程序中，equality和inequality检查，以及取整运算，均可使用Decimal数据类型的数值。

参考信息

无。

12.3.7.2 查询性能下降

现象描述

在不同的查询周期内运行查询功能，查询性能会有起伏。

可能原因

在处理数据加载时，为每个executor程序实例配置的内存不足，可能会产生更多的Java GC（垃圾收集）。当GC发生时，会发现查询性能下降。

定位思路

在Spark UI上，会发现某些executors的GC时间明显比其他executors高，或者所有的executors都表现出高GC时间。

处理步骤

登录Manager页面，选择“集群 > 服务 > Spark2x > 配置 > 全部配置”，在搜索框搜索“spark.executor.memory”，通过参数“spark.executor.memory”配置更高的内存值。

参考信息

无。

12.3.8 CarbonData FAQ

12.3.8.1 为什么对 decimal 数据类型进行带过滤条件的查询时会出现异常输出?

问题

当对decimal数据类型进行带过滤条件的查询时，输出结果不正确。

例如，

select * from carbon_table where num = 1234567890123456.22;

输出结果：

```
+------+----------------------+--+
| name |        num           |
+------+----------------------+--+
| IAA  | 1234567890123456.22 |
| IAA  | 1234567890123456.21 |
+------+----------------------+--+
```

回答

为了得到准确的输出结果，需在数字后面加上“BD”。

例如，

select * from carbon_table where num = 1234567890123456.22BD;

输出结果：

```
+------+----------------------+--+
| name |        num           |
+------+----------------------+--+
| IAA  | 1234567890123456.22 |
+------+----------------------+--+
```

12.3.8.2 如何避免对历史数据进行 minor compaction?

问题

如何避免对历史数据进行minor compaction?

回答

如果要先加载历史数据，后加载增量数据，则以下步骤可避免对历史数据进行minor compaction：

1. 加载所有历史数据。
2. 将major compaction大小配置为小于历史数据segment大小的值。
3. 对历史数据进行一次major compaction，之后将不会考虑这些segments进行minor compaction。
4. 加载增量数据。
5. 用户可以根据自己的需要配置minor compaction阈值。

配置示例和预期输出：

1. 用户将所有历史数据加载到CarbonData，此数据的一个segment的大小假定为500GB。
2. 用户设置major compaction参数的阈值：“carbon.major.compaction.size” = “491520（480gb * 1024）”。其中，*491520*可配置。
3. 运行major compaction。由于每个segment的大小超过配置值的大小，因此这些segments将会被压缩。
4. 加载增量负载。
5. 配置minor compaction参数的阈值：“compaction.level.threshold” = “6,6”。
6. 运行minor compaction。此时只考虑增量负载。

12.3.8.3 如何在 CarbonData 数据加载时修改默认的组名?

问题

如何在CarbonData数据加载时修改默认的组名?

回答

CarbonData数据加载时，默认的组名为“ficommon”。可以根据需要修改默认的组名。

1. 编辑“carbon.properties”文件。
2. 根据需要修改关键字“carbon.dataload.group.name”的值。其默认值为“ficommon”。

12.3.8.4 为什么 INSERT INTO CARBON TABLE 失败?

问题

为什么***INSERT INTO CARBON TABLE***命令无法在日志文件中记录以下信息?

```
Data load failed due to bad record
```

回答

在以下场景中，**INSERT INTO CARBON TABLE**命令会失败：

- 当源表和目标表的列数据类型不同时，源表中的数据将被视为Bad Records，则INSERT INTO命令会失败。
- 源列上的aggregartion函数的结果超过目标列的最大范围，则INSERT INTO命令会失败。

  解决方法：

  在进行插入操作时，可在对应的列上使用cast函数。

  示例：

  a. 使用DESCRIBE命令查询目标表和源表。

     DESCRIBE *newcarbontable*;

     结果：

     ```
     col1 int
     col2 bigint
     ```

     DESCRIBE *sourcetable*;

     结果：

     ```
     col1 int
     col2 int
     ```

  b. 添加cast函数以将BigInt类型数据转换为Integer类型数据。

     INSERT INTO *newcarbontable select col1, cast(col2 as integer) from sourcetable*;

12.3.8.5 为什么含转义字符的输入数据记录到 Bad Records 中的值与原始数据不同?

问题

为什么含转义字符的输入数据记录到Bad Records中的值与原始数据不同?

回答

转义字符以反斜线"\"开头，后跟一个或几个字符。如果输入记录包含类似\t，\b，\n，\r，\f，\'，\"，\\的转义字符，Java将把转义符'\'和它后面的字符一起处理得到转义后的值。

例如：如果CSV数据类似“2010\\10,test”，将这两列插入“String,int”类型时，因为“test”无法转换为int类型，表会将这条记录重定向到Bad Records中。但记录到Bad Records中的值为“2010\10”，Java会将原始数据中的“\\”转义为“\”。

12.3.8.6 为什么 Bad Records 导致数据加载性能降低?

问题

为什么Bad Records导致数据加载性能降低?

回答

如果数据中存在Bad Records，并且“BAD_RECORDS_LOGGER_ENABLE”参数值为“true”或“BAD_RECORDS_ACTION”参数值为“redirect”，则由于将失败原因写入日志文件中或将Bad Records重定向到原始CSV文件中导致的额外的I/O开销，数据加载性能就会降低 。

12.3.8.7 当初始 Executor 为 0 时，为什么 INSERT INTO/LOAD DATA 任务分配不正确，打开的 task 少于可用的 Executor?

问题

当初始Executor为0时，为什么INSERT INTO/LOAD DATA任务分配不正确，打开的task少于可用的Executor?

回答

在这种场景下，CarbonData会给每个节点分配一个INSERT INTO或LOAD DATA任务。如果Executor不是不同的节点分配的，CarbonData将会启动较少的task。

解决措施:

您可以适当增大Executor内存和Executor核数，以便YARN可以在每个节点上启动一个Executor。具体的配置方法如下:

1. 配置Executor核数。
   - 将“spark-defaults.conf”中的“spark.executor.cores”配置项或者“spark-env.sh”中的“SPARK_EXECUTOR_CORES”配置项设置为合适大小。
   - 在使用spark-submit命令时，添加“--executor-cores NUM”参数设置核数。

2. 配置Executor内存。
   - 将“spark-defaults.conf”中的“spark.executor.memory”配置项或者“spark-env.sh”中的“SPARK_EXECUTOR_MEMORY”配置项设置为合适大小。
   - 在使用spark-submit命令时，添加“--executor-memory MEM”参数设置内存。

12.3.8.8 为什么并行度大于待处理的 block 数目时，CarbonData 仍需要额外的 executor?

问题

为什么并行度大于待处理的block数目时，CarbonData仍需要额外的executor?

回答

CarbonData块分布对于数据处理进行了如下优化：

1. 优化数据处理并行度。
2. 优化了读取块数据的并行性。

为了优化并行数据处理及并行读取块数据，CarbonData根据块的局域性申请executor，因此CarbonData可获得所有节点上的executor。

为了优化并行数据处理及并行读取块数据，运用动态分配的用户需配置以下特性。

1. 使用参数“spark.dynamicAllocation.executorIdleTimeout”并将此参数值设置为15min（或平均查询时间）。
2. 正确配置参数“spark.dynamicAllocation.maxExecutors”，不推荐使用默认值（2048），否则CarbonData将申请最大数量的executor。
3. 对于更大的集群，配置参数“carbon.dynamicAllocation.schedulerTimeout”为10～15sec，默认值为5sec。
4. 配置参数“carbon.scheduler.minRegisteredResourcesRatio”为0.1～1.0，默认值为0.8。只要达到此参数值，块分布可启动。

12.3.8.9 为什么在 off heap 时数据加载失败?

问题

为什么在off heap时数据加载失败?

回答

YARN Resource Manager将（Java堆内存 +“spark.yarn.am.memoryOverhead”）作为内存限制，因此在off heap时，内存可能会超出此限制。您需配置参数“spark.yarn.am.memoryOverhead”以增加memory。

12.3.8.10 为什么创建 Hive 表失败?

问题

为什么创建Hive表失败?

回答

当源表或子查询具有大数据量的Partition时，创建Hive表失败。执行查询需要很多的task，此时输出的文件数就会很多，从而导致driver OOM。

可以在创建Hive表的语句中增加***distribute by***子句来解决这个问题，其中***distribute by***的字段要选取合适的cardinality（即distinct值的个数）。

distribute by子句限制了Hive表的Partition数量。增加***distribute by*** 子句后，最终的输出文件数取决于指定列的cardinality和“spark.sql.shuffle.partitions”参数值。但如果distribute by的字段的cardinality值很小，例如，“spark.sql.shuffle.partitions”参数值为200，但***distribute by***字段的cardinality只有100，则输出的200个文件中，只有其中100个文件有数据，剩下的100个文件为空文件。也就是说，如果选取的字段的cardinality过低，如1，则会造成严重的数据倾斜，从而严重影响查询性能。

因此，建议选取的distribute by字段的cardinality个数要大于“spark.sql.shuffle.partitions”参数，可大于2~3倍。

示例：

create table hivetable1 as select * from sourcetable1 distribute by col_age;

12.3.8.11 为什么在 V100R002C50RC1 版本中创建的 CarbonData 表不具有 Hive 特权为非所有者提供的特权?

问题

为什么在V100R002C50RC1版本中创建的CarbonData表不具有Hive特权为非所有者提供的特权?

回答

Hive ACL在V100R002C50RC1版本之后实现，因此无法体现Hive ACL特权。

为了支持在V100R002C50RC1中创建的CarbonData表的HIVE ACL特权，必须由表的所有者执行以下两个**ALTER TABLE**命令。

ALTER TABLE *$dbname.$tablename SET LOCATION '$carbon.store/$dbname/$tablename';*

ALTER TABLE *$dbname.$tablename SET SERDEPROPERTIES ('path'='$carbon.store/$dbname/$tablename');*

示例：

假设数据库名称为"carbondb"，表名为"carbontable"，Carbon的储存位置为“hdfs：//hacluster/user/hive/warehouse/carbon.store”，执行如下命令：

ALTER TABLE *carbondb.carbontable SET LOCATION 'hdfs:///hacluster/user/hive/warehouse/carbon.store/carbondb/carbontable';*

ALTER TABLE *carbondb.carbontable SET SERDEPROPERTIES ('path'='hdfs://hacluster/user/hive/warehouse/carbon.store/carbondb/carbontable');*

12.3.8.12 如何在不同的 namespaces 上逻辑地分割数据

问题

如何在不同的namespaces上逻辑地分割数据?

回答

- 配置:

  要在不同namespaces之间逻辑地分割数据，必须更新HDFS，Hive和Spark的“core-site.xml”文件中的以下配置。

  说明

  改变Hive组件将改变carbonstore的位置和warehouse的位置。

  - HDFS中的配置
    - fs.defaultFS - 默认文件系统的名称。URI模式必须设置为“viewfs”。当使用“viewfs”模式时，权限部分必须是“ClusterX”。
    - fs.viewfs.mountable.ClusterX.homedir - 主目录基本路径。每个用户都可以使用在“FileSystem/FileContext”中定义的getHomeDirectory()方法访问其主目录。
    - fs.viewfs.mountable.default.link.<dir_name> - ViewFS安装表。

    示例:

    ```
    <property>
    <name>fs.defaultFS</name>
    <value>viewfs://ClusterX/</value>
    </property>
    <property>
    <name>fs.viewfs.mounttable.ClusterX.link./folder1</name>
    <value>hdfs://NS1/folder1</value>
    </property>
    <property>
    <name>fs.viewfs.mounttable.ClusterX.link./folder2</name>
    <value>hdfs://NS2/folder2</value>
    </property>
    ```

  - Hive和Spark中的配置

    fs.defaultFS - 默认文件系统的名称。URI模式必须设置为“viewfs”。 当使用“viewfs”模式时，权限部分必须是“ClusterX”。

- 命令格式:

  LOAD DATA INPATH *'path to data' INTO TABLE table_name OPTIONS ('...');*

  说明

  每当Spark配置有viewFS文件系统时，当尝试从HDFS加载数据时，用户必须在***LOAD***语句中指定如“viewfs://”这样的路径或相对路径作为文件路径。

- 示例:
  - viewFS路径举例:

    LOAD DATA INPATH *'viewfs://ClusterX/dir/data.csv' INTO TABLE table_name OPTIONS ('...');*
  - 相对路径举例:

    LOAD DATA INPATH *'/apps/input_data1.txt'* ***INTO TABLE*** *table_name*;

12.3.8.13 为什么 drop 数据库抛出 Missing Privileges 异常?

问题

为什么drop数据库抛出以下异常?

```
Error: org.apache.spark.sql.AnalysisException: Missing Privileges;(State=,code=0)
```

回答

当数据库的所有者执行***drop database <database_name> cascade***命令（包含其他用户创建的表）时，会抛出此错误。

12.3.8.14 为什么在 Spark Shell 中不能执行更新命令?

问题

为什么在Spark Shell中不能执行更新命令?

回答

本文档中给出的语法和示例是关于Beeline的命令，而不是Spark Shell中的命令。

若要在Spark Shell中使用更新命令，可以使用以下语法。

- 语法1

  <carbon_context>.sql("UPDATE <CARBON TABLE> SET (column_name1, column_name2, ... column_name n) = (column1_expression , column2_expression , column3_expression ... column n_expression) [WHERE { <filter_condition> }];").show

- 语法2

  <carbon_context>.sql("UPDATE <CARBON TABLE> SET (column_name1, column_name2,) = (select sourceColumn1, sourceColumn2 from sourceTable [WHERE { <filter_condition> }]) [WHERE { <filter_condition> }];").show

示例:

如果CarbonData的context是“carbon”，那么更新命令如下:

carbon.sql("update carbonTable1 d set (d.column3,d.column5) = (select s.c33 ,s.c55 from sourceTable1 s where d.column1 = s.c11) where d.column1 = 'country' exists(select * from table3 o where o.c2 > 1);").show

12.3.8.15 如何在 CarbonData 中配置非安全内存?

问题

如何在CarbonData中配置非安全内存?

回答

在Spark配置中，“spark.yarn.executor.memoryOverhead”参数的值应大于CarbonData配置参数“sort.inmemory.size.inmb” 与“Netty offheapmemory

required”参数值的总和，或者“carbon.unsafe.working.memory.in.mb” 、“carbon.sort.inememory.storage.size.in.mb” 与 “Netty offheapmemory required”参数值的总和。否则，如果堆外（off heap）访问超出配置的executor内存，则YARN可能会停止executor。

“Netty offheapmemory required”说明：当“spark.shuffle.io.preferDirectBufs”设为true时，Spark中netty 传输服务从"spark.yarn.executor.memoryOverhead"中拿掉部分堆内存[~ 384 MB or 0.1 x 执行器内存]。

详细信息请参考常见**配置executor堆外内存大小**。

12.3.8.16 设置了 HDFS 存储目录的磁盘空间配额，CarbonData 为什么会发生异常?

问题

设置了HDFS存储目录的磁盘空间配额，CarbonData为什么会发生异常。

回答

创建、加载、更新表或进行其他操作时，数据会被写入HDFS。若HDFS目录的磁盘空间配额不足，则操作失败并抛出以下异常。

```
org.apache.hadoop.hdfs.protocol.DSQuotaExceededException: The DiskSpace quota of /user/tenant is exceeded: quota = 314572800 B = 300 MB but diskspace consumed = 402653184 B = 384 MB at org.apache.hadoop.hdfs.server.namenode.DirectoryWithQuotaFeature.verifyStoragespaceQuota(DirectoryWithQuotaFeature.java:211) at org.apache.hadoop.hdfs.server.namenode.DirectoryWithQuotaFeature.verifyQuota(DirectoryWithQuotaFeature.java:239) at org.apache.hadoop.hdfs.server.namenode.FSDirectory.verifyQuota(FSDirectory.java:941) at org.apache.hadoop.hdfs.server.namenode.FSDirectory.updateCount(FSDirectory.java:745)
```

若发生此异常，请为租户配置足够的磁盘空间配额。

例如：

需要的磁盘空间配置可以按照如下方法计算：

如果HDFS的副本数为3， HDFS默认的块大小为128MB，则最小需要384MB的磁盘空间用于写表的schema文件到HDFS上。计算公式：no. of block x block_size x replication_factor of the schema file = 1 x 128 x 3 = 384 MB

说明

数据加载时，由于默认块大小为1024MB，每个fact文件需要的最小空间为3072MB。

12.3.8.17 为什么数据查询/加载失败，且抛出“org.apache.carbondata.core.memory.MemoryException: Not enough memory”异常?

问题

为什么数据查询/加载失败，且抛出“org.apache.carbondata.core.memory.MemoryException: Not enough memory”异常?

回答

当执行器中此次数据查询和加载所需要的堆外内存不足时，便会抛出此异常。

在这种情况下,请增大“carbon.unsafe.working.memory.in.mb”和“spark.yarn.executor.memoryOverhead”的值。

详细信息请参考**如何在CarbonData中配置非安全内存?**

该内存被数据查询和加载共享。所以如果加载和查询需要同时进行，建议将“carbon.unsafe.working.memory.in.mb”和“spark.yarn.executor.memoryOverhead”的值配置为2048 MB以上。

可以使用以下公式进行估算：

数据加载所需内存：

(“carbon.number.of.cores.while.loading”的值[默认值 = 6]) x 并行加载数据的表格 x (“offheap.sort.chunk.size.inmb”的值[默认值 = 64 MB] + “carbon.blockletgroup.size.in.mb”的值[默认值 = 64 MB] + 当前的压缩率[64 MB/ 3.5])

= ~900 MB 每表格

数据查询所需内存：

(SPARK_EXECUTOR_INSTANCES. [默认值 = 2]) x (carbon.blockletgroup.size.in.mb [默认值 = 64 MB] +“carbon.blockletgroup.size.in.mb”解压内容[默认值 = 64 MB * 3.5]) x (每个执行器核数[默认值 = 1])

= ~ 600 MB

12.3.8.18 开启防误删下，为什么 Carbon 表没有执行 drop table 命令，回收站中也会存在该表的文件?

问题

开启防误删下，为什么Carbon表没有执行drop table命令，回收站中也会存在该表的文件?

回答

在Carbon适配防误删后，调用文件删除命令，会将删除的文件放入回收站中。在insert、load等命令中会有中间文件.carbonindex文件的删除，所以在未执行drop table命令的时候，回收站中也可能会存在该表的文件。如果这个时候再执行drop table命令，那么按照回收站机制，会生成一个带时间戳的该表目录，该目录中的文件是完整的。

12.4 使用 ClickHouse

12.4.1 从零开始使用 ClickHouse

ClickHouse是面向联机分析处理的列式数据库，支持SQL查询，且查询性能好，特别是基于大宽表的聚合分析查询性能非常优异，比其他分析型数据库速度快一个数量级。

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。在使用客户端前，需要先下载并更新客户端配置文件，确认Manager的主管理节点后才能使用客户端。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限。如果当前集群未启用Kerberos认证，则无需执行本步骤。

1. 如果是MRS 3.1.0版本集群，则需要先执行：**export CLICKHOUSE_SECURITY_ENABLED=true**
2. **kinit** *组件业务用户*

   例如，**kinit** clickhouseuser。

步骤5 执行ClickHouse组件的客户端命令。

执行**clickhouse -h**，查看ClickHouse组件命令帮助。

回显信息如下：

```
Use one of the following commands:
clickhouse local [args]
clickhouse client [args]
clickhouse benchmark [args]
clickhouse server [args]
clickhouse performance-test [args]
clickhouse extract-from-config [args]
clickhouse compressor [args]
clickhouse format [args]
clickhouse copier [args]
clickhouse obfuscator [args]
...
```

MRS 3.1.0及之后版本，使用**clickhouse client**命令连接ClickHouse服务端：

- 例如，当前集群未启用Kerberos认证，使用ssl安全方式登录：

  clickhouse client --host *ClickHouse的实例IP* **--user** *用户名* **--password** *密码* **--port** 9440 **--secure**
- 例如，当前集群已启用Kerberos认证，使用ssl安全方式登录。

  Kerberos集群场景下没有默认用户，必须在Manager上创建用户，详细参考**ClickHouse用户及权限管理**。

  使用kinit认证成功后，客户端登录时可以不携带--user和--password参数，即使用kinit认证的用户登录。

  clickhouse client --host *ClickHouse的实例IP* **--port** 9440 **--secure**

  相关参数使用说明如下表：

表 12-54 clickhouse client 命令行参数说明

参数名	参数说明
--host	服务端的host名称，默认是localhost。您可以选择使用ClickHouse实例所在节点主机名或者IP地址。 **说明** ClickHouse的实例IP地址可登录集群FusionInsight Manager，然后选择“集群 > 服务 > ClickHouse > 实例”，获取ClickHouseServer实例对应的业务IP地址。
--port	连接的端口。 • 如果使用ssl安全连接则默认端口为9440，并且需要携带参数--secure。具体的端口值可通过ClickHouseServer实例配置搜索“tcp_port_secure”参数获取。 • 如果使用非ssl安全连接则默认端口为9000，不需要携带参数--secure。具体的端口值可通过ClickHouseServer实例配置搜索“tcp_port”参数获取。
--user	用户名。 可以在Manager上创建该用户名并绑定对应的角色权限，具体可以参考**ClickHouse用户及权限管理**。 • 如果当前集群已启用Kerberos认证，使用kinit认证成功后，客户端登录时可以不携带--user和--password参数，即使用kinit认证的用户登录。Kerberos集群场景下没有默认用户，必须在Manager上创建该用户名。 • 如果当前集群未启用Kerberos认证，客户端登录时可以指定Manager上创建的用户和密码。不携带用户和密码参数时则默认使用**default**用户登录。
--password	密码。 默认值：空字符串。该参数和--user参数配套使用，可以在Manager上创建用户名时设置该密码。
--query	使用非交互模式查询。
--database	默认当前操作的数据库。默认值：服务端默认的配置（默认是default）。
--multiline	如果指定，允许多行语句查询（Enter仅代表换行，不代表查询语句完结）。
--multiquery	如果指定，允许处理用;号分隔的多个查询，只在非交互模式下生效。
--format	使用指定的默认格式输出结果。
--vertical	如果指定，默认情况下使用垂直格式输出结果。在这种格式中，每个值都在单独的行上打印，适用显示宽表的场景。
--time	如果指定，非交互模式下会打印查询执行的时间到stderr中。
--stacktrace	如果指定，如果出现异常，会打印堆栈跟踪信息。
--config-file	配置文件的名称。
--secure	如果指定，将通过ssl安全模式连接到服务器。

参数名	参数说明
--history_file	存放命令历史的文件的路径。
--param_<name>	带有参数的查询，并将值从客户端传递给服务器。具体用法详见 https://clickhouse.tech/docs/zh/interfaces/cli/#cli-queries-with-parameters。

----结束

12.4.2 ClickHouse 表引擎介绍

背景介绍

表引擎在ClickHouse中的作用十分关键，不同的表引擎决定了：

- 数据存储和读取的位置
- 支持哪些查询方式
- 能否并发式访问数据
- 能不能使用索引
- 是否可以执行多线程请求
- 数据复制使用的参数

其中MergeTree和Distributed是ClickHouse表引擎中最重要，也是最常使用的两个引擎，本文将重点进行介绍。

其他表引擎详细可以参考官网链接：**https://clickhouse.tech/docs/en/engines/table-engines**。

MergeTree 系列引擎

MergeTree用于高负载任务的最通用和功能最强大的表引擎，其主要有以下关键特征：

- 基于分区键（partitioning key）的数据分区分块存储
- 数据索引排序（基于primary key和order by）
- 支持数据复制（带Replicated前缀的表引擎）
- 支持数据抽样

在写入数据时，该系列引擎表会按照分区键将数据分成不同的文件夹，文件夹内每列数据为不同的独立文件，以及创建数据的序列化索引排序记录文件。该结构使得数据读取时能够减少数据检索时的数据量，极大的提高查询效率。

- MergeTree

  建表语法：

```
CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
(
    name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1] [TTL expr1],
    name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2] [TTL expr2],
    ...
    INDEX index_name1 expr1 TYPE type1(...) GRANULARITY value1,
```

```
    INDEX index_name2 expr2 TYPE type2(...) GRANULARITY value2
) ENGINE = MergeTree()
ORDER BY expr
[PARTITION BY expr]
[PRIMARY KEY expr]
[SAMPLE BY expr]
[TTL expr [DELETE|TO DISK 'xxx'|TO VOLUME 'xxx'], ...]
[SETTINGS name=value, ...]
```

使用示例：

```
CREATE TABLE default.test (
  name1 DateTime,
  name2 String,
  name3 String,
  name4 String,
  name5 Date,
  ...
) ENGINE = MergeTree()
PARTITION BY toYYYYMM(name5)
ORDER BY (name1, name2)
SETTINGS index_granularity = 8192
```

示例参数说明如下：

- **ENGINE = MergeTree()**：MergeTree表引擎。
- **PARTITION BY** toYYYYMM(name4)：分区，示例数据将以月份为分区，每个月份一个文件夹。
- **ORDER BY**：排序字段，支持多字段的索引排序，第一个相同的时候按照第二个排序依次类推。
- **index_granularity = 8192**：排序索引的颗粒度，每8192条数据记录一个排序索引值。

如果被查询的数据存在于分区或排序字段中，能极大降低数据查找时间。

- ReplacingMergeTree

  该引擎和MergeTree的不同之处在于它会删除排序键值相同的重复项。ReplacingMergeTree适合于清除重复数据节省存储空间，但是它不保证重复数据不出现，一般不建议使用。

  建表语法：

  ```
  CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
  (
    name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1],
    name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2],
    ...
  ) ENGINE = ReplacingMergeTree([ver])
  [PARTITION BY expr]
  [ORDER BY expr]
  [SAMPLE BY expr]
  [SETTINGS name=value, ...]
  ```

- SummingMergeTree

  当合并SummingMergeTree表的数据片段时，ClickHouse会把所有具有相同主键的行合并为一行，该行包含了被合并的行中具有数值数据类型的列的汇总值。如果主键的组合方式使得单个键值对应于大量的行，则可以显著的减少存储空间并加快数据查询的速度。

  建表语法：

  ```
  CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
  (
    name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1],
    name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2],
    ...
  ) ENGINE = SummingMergeTree([columns])
  [PARTITION BY expr]
  ```

```
[ORDER BY expr]
[SAMPLE BY expr]
[SETTINGS name=value, ...]
```

使用示例：

创建一个SummingMergeTree表testTable：

```
CREATE TABLE testTable
(
   id UInt32,
   value UInt32
)
ENGINE = SummingMergeTree()
ORDER BY id
```

插入表数据：

```
INSERT INTO testTable Values(5,9),(5,3),(4,6),(1,2),(2,5),(1,4),(3,8);
INSERT INTO testTable Values(88,5),(5,5),(3,7),(3,5),(1,6),(2,6),(4,7),(4,6),(43,5),(5,9),(3,6);
```

在未合并parts查询所有数据：

```
SELECT * FROM testTable
┌─id─┬─value─┐
│  1 │     6 │
│  2 │     5 │
│  3 │     8 │
│  4 │     6 │
│  5 │    12 │
└────┴───────┘
┌─id─┬─value─┐
│  1 │     6 │
│  2 │     6 │
│  3 │    18 │
│  4 │    13 │
│  5 │    14 │
│ 43 │     5 │
│ 88 │     5 │
└────┴───────┘
```

ClickHouse还没有汇总所有行，如果需要通过ID进行汇总聚合，需要用到sum和GROUP BY子句：

```
SELECT id, sum(value) FROM testTable GROUP BY id
┌─id─┬─sum(value)─┐
│  4 │         19 │
│  3 │         26 │
│ 88 │          5 │
│  2 │         11 │
│  5 │         26 │
│  1 │         12 │
│ 43 │          5 │
└────┴────────────┘
```

手工执行合并操作：

```
OPTIMIZE TABLE testTable
```

此时再查询testTable表数据：

```
SELECT * FROM testTable
┌─id─┬─value─┐
│  1 │    12 │
│  2 │    11 │
│  3 │    26 │
│  4 │    19 │
│  5 │    26 │
│ 43 │     5 │
│ 88 │     5 │
└────┴───────┘
```

SummingMergeTree根据ORDER BY排序键作为聚合数据的条件Key。即如果排序key是相同的，则会合并成一条数据，并对指定的合并字段进行聚合。

后台执行合并操作时才会进行数据的预先聚合，而合并操作的执行时机无法预测，所以可能存在部分数据已经被预先聚合、部分数据尚未被聚合的情况。因此，在执行聚合计算时，SQL中仍需要使用GROUP BY子句。

- AggregatingMergeTree

  AggregatingMergeTree是预先聚合引擎的一种，用于提升聚合计算的性能。AggregatingMergeTree引擎能够在合并分区时，按照预先定义的条件聚合数据，同时根据预先定义的聚合函数计算数据并通过二进制的格式存入表内。

  建表语法：

```
CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
(
    name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1],
    name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2],
    ...
) ENGINE = AggregatingMergeTree()
[PARTITION BY expr]
[ORDER BY expr]
[SAMPLE BY expr]
[TTL expr]
[SETTINGS name=value, ...]
```

  使用示例：

  AggregatingMergeTree无单独参数设置，在分区合并时，在每个数据分区内，会按照ORDER BY聚合，使用何种聚合函数，对哪些列字段计算，则是通过定义AggregateFunction函数类型实现，例如：

```
create table test_table (
name1 String,
name2 String,
name3 AggregateFunction(uniq,String),
name4 AggregateFunction(sum,Int),
name5 DateTime
) ENGINE = AggregatingMergeTree()
PARTITION BY toYYYYMM(name5)
ORDER BY (name1,name2)
PRIMARY KEY name1;
```

  AggregateFunction类型的数据在写入和查询时需要分别调用*state、*merge函数，*表示定义字段类型时使用的聚合函数。如上示例表test_table定义的name3、name4字段分别使用了uniq、sum函数，那么在写入数据时需要调用uniqState、sumState函数，并使用INSERT SELECT语法。

```
insert into test_table select '8','test1',uniqState('name1'),sumState(toInt32(100)),'2021-04-30 17:18:00';
insert into test_table select '8','test1',uniqState('name1'),sumState(toInt32(200)),'2021-04-30 17:18:00';
```

  在查询数据时也需要调用对应的函数uniqMerge、sumMerge：

```
select name1,name2,uniqMerge(name3),sumMerge(name4) from test_table group by name1,name2;
┌─name1─┬─name2─┬─uniqMerge(name3)─┬─sumMerge(name4)─┐
│ 8     │ test1 │                1 │             300 │
└───────┴───────┴──────────────────┴─────────────────┘
```

  AggregatingMergeTree更常用的方式是结合物化视图使用，物化视图即其它数据表上层的一种查询视图。详细可以参考：**https://clickhouse.tech/docs/en/engines/table-engines/mergetree-family/aggregatingmergetree/**

- CollapsingMergeTree

  CollapsingMergeTree它通过定义一个sign标记位字段记录数据行的状态。如果sign标记为1，则表示这是一行有效的数据；如果sign标记为-1，则表示这行数据需要被删除。

  建表语法：

```
CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
(
```

```
    name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1],
    name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2],
    ...
) ENGINE = CollapsingMergeTree(sign)
[PARTITION BY expr]
[ORDER BY expr]
[SAMPLE BY expr]
[SETTINGS name=value, ...]
```

使用示例：

具体的使用示例可以参考：**https://clickhouse.tech/docs/en/engines/table-engines/mergetree-family/collapsingmergetree/**。

- VersionedCollapsingMergeTree

  VersionedCollapsingMergeTree表引擎在建表语句中新增了一列version，用于在乱序情况下记录状态行与取消行的对应关系。主键相同，且Version相同、Sign相反的行，在Compaction时会被删除。

  建表语法：

  ```
  CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
  (
      name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1],
      name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2],
      ...
  ) ENGINE = VersionedCollapsingMergeTree(sign, version)
  [PARTITION BY expr]
  [ORDER BY expr]
  [SAMPLE BY expr]
  [SETTINGS name=value, ...]
  ```

  使用示例：

  具体的使用示例可以参考：**https://clickhouse.tech/docs/en/engines/table-engines/mergetree-family/versionedcollapsingmergetree/**。

- GraphiteMergeTree

  GraphiteMergeTree引擎用来存储时序数据库Graphite的数据。

  建表语法：

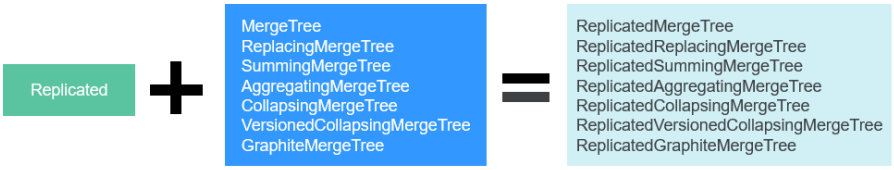
  ```
  CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
  (
      Path String,
      Time DateTime,
      Value <Numeric_type>,
      Version <Numeric_type>
      ...
  ) ENGINE = GraphiteMergeTree(config_section)
  [PARTITION BY expr]
  [ORDER BY expr]
  [SAMPLE BY expr]
  [SETTINGS name=value, ...]
  ```

  使用示例：

  具体的使用示例可以参考：**https://clickhouse.tech/docs/en/engines/table-engines/mergetree-family/graphitemergetree/**。

Replicated*MergeTree 引擎

ClickHouse中的所有MergeTree家族引擎前面加上Replicated就成了支持副本的合并树引擎。



Replicated系列引擎借助ZooKeeper实现数据的同步，创建Replicated复制表时通过注册到ZooKeeper上的信息实现同一个分片的所有副本数据进行同步。

Replicated表引擎的创建模板：

```
ENGINE = Replicated*MergeTree('ZooKeeper存储路径','副本名称', ...)
```

Replicated表引擎需指定两个参数：

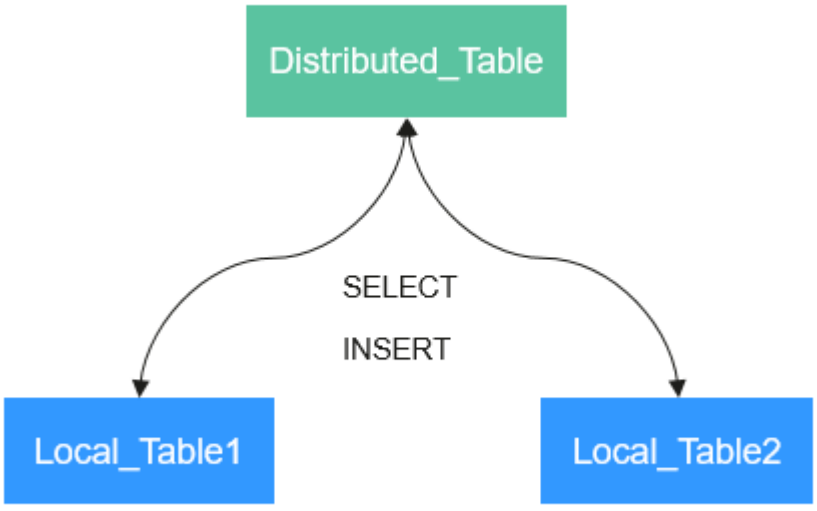
- ZooKeeper存储路径：ZooKeeper中该表相关数据的存储路径，建议规范化，如：*/clickhouse/tables/{shard}/数据库名/表名*。
- 副本名称，一般用**{replica}**即可。

Replicated表引擎使用示例可以参考：**ClickHouse表创建**。

Distributed 表引擎

Distributed表引擎本身不存储任何数据，而是作为数据分片的透明代理，能够自动路由数据到集群中的各个节点，分布式表需要和其他本地数据表一起协同工作。分布式表会将接收到的读写任务分发到各个本地表，而实际上数据的存储在各个节点的本地表中。

图 12-4 Distributed 原理图



Distributed表引擎的创建模板：

```
ENGINE = Distributed(cluster_name, database_name, table_name, [sharding_key])
```

Distributed表参数解析如下：

- cluster_name：集群名称，在对分布式表执行读写的过程中，使用集群的配置信息查找对应的ClickHouse实例节点。
- database_name：数据库名称。
- table_name：数据库下对应的本地表名称，用于将分布式表映射到本地表上。
- sharding_key：分片键（可选参数），分布式表会按照这个规则，将数据分发到各个本地表中。

Distributed表引擎使用示例：

```
--先创建一个表名为test的ReplicatedMergeTree本地表
CREATE TABLE default.test ON CLUSTER default_cluster_1
(
    `EventDate` DateTime,
    `id` UInt64
)
ENGINE = ReplicatedMergeTree('/clickhouse/tables/{shard}/default/test', '{replica}')
PARTITION BY toYYYYMM(EventDate)
ORDER BY id

--基于本地表test创建表名为test_all的Distributed表
CREATE TABLE default.test_all ON CLUSTER default_cluster_1
(
    `EventDate` DateTime,
    `id` UInt64
)
ENGINE = Distributed(default_cluster_1, default, test, rand())
```

分布式表创建规则：

- 创建Distributed表时需加上**on cluster** *cluster_name*，这样建表语句在某一个ClickHouse实例上执行一次即可分发到集群中所有实例上执行。
- 分布式表通常以本地表加“_all”命名。它与本地表形成一对多的映射关系，之后可以通过分布式表代理操作多张本地表。
- 分布式表的表结构尽量和本地表的结构一致。如果不一致，在建表时不会报错，但在查询或者插入时可能会抛出异常。

12.4.3 ClickHouse 表创建

ClickHouse依靠ReplicatedMergeTree引擎与ZooKeeper实现了复制表机制，用户在创建表时可以通过指定引擎选择该表是否高可用，每张表的分片与副本都是互相独立的。

同时ClickHouse依靠Distributed引擎实现了分布式表机制，在所有分片（本地表）上建立视图进行分布式查询，使用很方便。ClickHouse有数据分片（shard）的概念，这也是分布式存储的特点之一，即通过并行读写提高效率。

CPU架构为鲲鹏计算的ClickHouse集群表引擎不支持使用HDFS和Kafka。

查看 ClickHouse 服务 cluster 等环境参数信息

步骤1 参考**从零开始使用ClickHouse**使用ClickHouse客户端连接到ClickHouse服务端。

步骤2 查询集群标识符cluster等其他环境参数信息。

select cluster,shard_num,replica_num,host_name from system.clusters;

```
SELECT
    cluster,
```

```
    shard_num,
    replica_num,
    host_name
FROM system.clusters

┌─cluster─────────────┬─shard_num─┬─replica_num─┬─host_name──────────┐
│ default_cluster_1   │         1 │           1 │ node-master1dOnG   │
│ default_cluster_1   │         1 │           2 │ node-group-1tXED0001 │
│ default_cluster_1   │         2 │           1 │ node-master2OXQS   │
│ default_cluster_1   │         2 │           2 │ node-group-1tXED0002 │
│ default_cluster_1   │         3 │           1 │ node-master3QsRI   │
│ default_cluster_1   │         3 │           2 │ node-group-1tXED0003 │
└─────────────────────┴───────────┴─────────────┴────────────────────┘

6 rows in set. Elapsed: 0.001 sec.
```

步骤3 查询分片标识符shard和副本标识符replica。

select * from system.macros;

```
SELECT *
FROM system.macros

┌─macro───┬─substitution──────┐
│ id      │ 76                │
│ replica │ node-master3QsRI  │
│ shard   │ 3                 │
└─────────┴───────────────────┘

3 rows in set. Elapsed: 0.001 sec.
```

----结束

创建本地复制表和分布式表

步骤1 客户端登录ClickHouse节点，例如：**clickhouse client --host** *node-master3QsRI* **--multiline --port 9440 --secure;**

说明

node-master3QsRI 参数为**查看ClickHouse服务cluster等环境参数信息**中**步骤2**对应的host_name参数的值。

步骤2 使用ReplicatedMergeTree引擎创建复制表。

详细的语法说明请参考：**https://clickhouse.tech/docs/zh/engines/table-engines/mergetree-family/replication/#creating-replicated-tables**。

例如，如下在default_cluster_1集群节点上和default数据库下创建表名为test的ReplicatedMergeTree表：

CREATE TABLE *default.test* **ON CLUSTER** *default_cluster_1*

(

`EventDate` DateTime,

`id` UInt64

)

ENGINE = ReplicatedMergeTree('*/clickhouse/tables/{shard}/default/test***', '***{replica}***')**

PARTITION BY toYYYYMM(EventDate)

ORDER BY id;

参数说明如下：

- ON CLUSTER语法表示分布式DDL，即执行一次就可在集群所有实例上创建同样的本地表。
- default_cluster_1为**查看ClickHouse服务cluster等环境参数信息**中**步骤2**查询到的cluster集群标识符。

> **⚠ 注意**
>
> **ReplicatedMergeTree**引擎族接收两个参数：
>
> - ZooKeeper中该表相关数据的存储路径。
>
> **该路径必须在/clickhouse目录下，否则后续可能因为ZooKeeper配额不够导致数据插入失败。**
>
> **为了避免不同表在ZooKeeper上数据冲突，目录格式必须按照如下规范填写：**
>
> */clickhouse/tables/{shard}/default/test*，其中/clickhouse/tables/{shard}为固定值，default为数据库名，test为创建的表名。
> - 副本名称，一般用**{replica}**即可。

```
CREATE TABLE default.test ON CLUSTER default_cluster_1
(
   `EventDate` DateTime,
   `id` UInt64
)
ENGINE = ReplicatedMergeTree('/clickhouse/tables/{shard}/default/test', '{replica}')
PARTITION BY toYYYYMM(EventDate)
ORDER BY id

┌─host──────────────────────┬─port─┬─status─┬─error─┬─num_hosts_remaining─┬─num_hosts_active─┐
│ node-group-1tXED0002      │ 9000 │      0 │       │                   5 │                3 │
│ node-group-1tXED0003      │ 9000 │      0 │       │                   4 │                3 │
│ node-master1dOnG          │ 9000 │      0 │       │                   3 │                3 │
└───────────────────────────┴──────┴────────┴───────┴─────────────────────┴──────────────────┘
┌─host──────────────────────┬─port─┬─status─┬─error─┬─num_hosts_remaining─┬─num_hosts_active─┐
│ node-master3QsRI          │ 9000 │      0 │       │                   2 │                0 │
│ node-group-1tXED0001      │ 9000 │      0 │       │                   1 │                0 │
│ node-master2OXQS          │ 9000 │      0 │       │                   0 │                0 │
└───────────────────────────┴──────┴────────┴───────┴─────────────────────┴──────────────────┘

6 rows in set. Elapsed: 0.189 sec.
```

步骤3 使用Distributed引擎创建分布式表。

例如，以下将在default_cluster_1集群节点上和default数据库下创建名为test_all 的Distributed表：

CREATE TABLE *default.test_all* **ON CLUSTER** *default_cluster_1*

(

`EventDate` DateTime,

`id` UInt64

)

ENGINE = Distributed(*default_cluster_1, default, test, rand()*);

```
CREATE TABLE default.test_all ON CLUSTER default_cluster_1
(
   `EventDate` DateTime,
   `id` UInt64
)
ENGINE = Distributed(default_cluster_1, default, test, rand())

┌─host──────────────────────┬─port─┬─status─┬─error─┬─num_hosts_remaining─┬─num_hosts_active─┐
│ node-group-1tXED0002      │ 9000 │      0 │       │                   5 │                0 │
│ node-master3QsRI          │ 9000 │      0 │       │                   4 │                0 │
│ node-group-1tXED0003      │ 9000 │      0 │       │                   3 │                0 │
│ node-group-1tXED0001      │ 9000 │      0 │       │                   2 │                0 │
│ node-master1dOnG          │ 9000 │      0 │       │                   1 │                0 │
│ node-master2OXQS          │ 9000 │      0 │       │                   0 │                0 │
└───────────────────────────┴──────┴────────┴───────┴─────────────────────┴──────────────────┘

6 rows in set. Elapsed: 0.115 sec.
```

说明

Distributed引擎需要以下几个参数：

- default_cluster_1为**查看ClickHouse服务cluster等环境参数信息**中**步骤2**查询到的cluster集群标识符。
- default本地表所在的数据库名称。
- test为本地表名称，该例中为**步骤2**中创建的表名。
- （可选的）分片键（sharding key）

  该键与config.xml中配置的分片权重（weight）一同决定写入分布式表时的路由，即数据最终落到哪个物理表上。它可以是表中一列的原始数据（如site_id），也可以是函数调用的结果，如上面的SQL语句采用了随机值rand()。注意该键要尽量保证数据均匀分布，另外一个常用的操作是采用区分度较高的列的哈希值，如intHash64(user_id)。

----结束

ClickHouse 表数据操作

步骤1 客户端登录ClickHouse节点。例如：

clickhouse client --host *node-master3QsRI* **--multiline --port 9440 --secure;**

说明

node-master3QsRI 参数为**查看ClickHouse服务cluster等环境参数信息**中**步骤2**对应的host_name参数的值。

步骤2 参考**创建本地复制表和分布式表**创建表后，可以插入数据到本地表。

例如插入数据到本地表：test

insert into test values(toDateTime(now()), rand());

步骤3 查询本地表信息。

例如查询**步骤2**中的表test数据信息：

select * from test;

```
SELECT *
FROM test
```

```
┌─────────────EventDate─┬──────────id─┐
│ 2020-11-05 21:10:42   │ 1596238076  │
└───────────────────────┴─────────────┘

1 rows in set. Elapsed: 0.002 sec.
```

步骤4 查询Distributed分布式表。

例如**步骤3**中因为分布式表test_all基于test创建，所以test_all表也能查询到和test相同的数据。

select * from test_all;

```
SELECT *
FROM test_all

┌─────────────EventDate─┬──────────id─┐
│ 2020-11-05 21:10:42   │ 1596238076  │
└───────────────────────┴─────────────┘

1 rows in set. Elapsed: 0.004 sec.
```

步骤5 切换登录节点为相同shard_num的shard节点，并且查询当前表信息，能查询到相同的表数据。

例如，退出原有登录节点：**exit;**

切换到节点node-group-1tXED0003：

clickhouse client --host *node-group-1tXED0003* **--multiline --port 9440 --secure;**

说明

通过**步骤2**可以看到node-group-1tXED0003和node-master3QsRI的shard_num值相同。

show tables;

```
SHOW TABLES

┌─name─────┐
│ test     │
│ test_all │
└──────────┘
```

步骤6 查询本地表数据。例如在节点node-group-1tXED0003查询test表数据。

select * from test;

```
SELECT *
FROM test

┌─────────────EventDate─┬──────────id─┐
│ 2020-11-05 21:10:42   │ 1596238076  │
└───────────────────────┴─────────────┘

1 rows in set. Elapsed: 0.005 sec.
```

步骤7 切换到不同shard_num的shard节点，并且查询之前创建的表数据信息。

例如退出之前的登录节点node-group-1tXED0003：

exit;

切换到node-group-1tXED0001节点。通过**步骤2**可以看到node-group-1tXED0001和node-master3QsRI的shard_num值不相同。

clickhouse client --host *node-group-1tXED0001* **--multiline --port 9440 --secure;**

查询test本地表数据，因为test是本地表所以在不同分片节点上查询不到数据。

select * from test;

```
SELECT *
FROM test

Ok.
```

查询test_all分布式表数据，能正常查询到数据信息。

select * from test_all;

```
SELECT *
FROM test

┌─────────────EventDate─┬──────────id─┐
│ 2020-11-05 21:12:19   │ 3686805070  │
└───────────────────────┴─────────────┘

1 rows in set. Elapsed: 0.002 sec.
```

----结束

12.4.4 ClickHouse 常用 SQL 语法

12.4.4.1 CREATE DATABASE 创建数据库

本章节主要介绍ClickHouse创建数据库的SQL基本语法和使用说明。

基本语法

CREATE DATABASE [IF NOT EXISTS] *database_name* **[ON CLUSTER** *ClickHouse集群名***]**

说明

ON CLUSTER ClickHouse集群名的语法，使得该DDL语句执行一次即可在集群中所有实例上都执行。集群名信息可以使用以下语句的**cluster**字段获取：

select cluster,shard_num,replica_num,host_name from system.clusters;

使用示例

```
--创建数据库名为test的数据库
CREATE DATABASE test ON CLUSTER default_cluster;
--创建成功后，通过查询命令验证
show databases;
┌─name───┐
│ default │
│ system  │
│ test    │
└────────┘
```

12.4.4.2 CREATE TABLE 创建表

本章节主要介绍ClickHouse创建表的SQL基本语法和使用说明。

基本语法

- 方法一：在指定的“database_name”数据库中创建一个名为“table_name ”的表。

  如果建表语句中没有包含“database_name”，则默认使用客户端登录时选择的数据库作为数据库名称。

  CREATE TABLE [IF NOT EXISTS] *[database_name.]table_name* **[ON CLUSTER** *ClickHouse集群名*]

  (

  name1 **[***type1***] [DEFAULT|MATERIALIZED|ALIAS** *expr1***],**

  name2 **[***type2***] [DEFAULT|MATERIALIZED|ALIAS** *expr2***],**

  ...

) *ENGINE = engine_name()*

  [**PARTITION BY** *expr_list*]

  [**ORDER BY** *expr_list*]

  注意

  ClickHouse在创建表时建议携带**PARTITION BY**创建表分区。因为ClickHouse数据迁移工具是基于表的分区做数据迁移，在创建表时如果不携带**PARTITION BY**创建表分区，则在**使用ClickHouse数据迁移工具**界面无法对该表进行数据迁移。

- 方法二：创建一个与database_name2.table_name2具有相同结构的表，同时可以对其指定不同的表引擎声明。

  如果没有表引擎声明，则创建的表将与database_name2.table_name2使用相同的表引擎。

  CREATE TABLE [IF NOT EXISTS] *[database_name.]table_name* **AS** [*database_name2.*]*table_name2* [ENGINE = *engine_name*]

- 方法三：使用指定的引擎创建一个与SELECT子句的结果具有相同结构的表，并使用SELECT子句的结果填充它。

  CREATE TABLE [IF NOT EXISTS] *[database_name.]table_name ENGINE = engine_name* **AS SELECT** ...

使用示例

```
--在default数据库和default_cluster集群下创建名为test表
CREATE TABLE default.test ON CLUSTER default_cluster
(
   `EventDate` DateTime,
   `id` UInt64
)
ENGINE = ReplicatedMergeTree('/clickhouse/tables/{shard}/default/test', '{replica}')
PARTITION BY toYYYYMM(EventDate)
ORDER BY id
```

12.4.4.3 INSERT INTO 插入表数据

本章节主要介绍ClickHouse插入表数据的SQL基本语法和使用说明。

基本语法

- 方法一：标准格式插入数据。

  INSERT INTO *[database_name.]table* [(*c1, c2, c3*)] **VALUES** (*v11, v12, v13*), (*v21, v22, v23*), ...

- 方法二：使用SELECT的结果写入。

  INSERT INTO *[database_name.]table* [(c1, c2, c3)] **SELECT** ...

使用示例

```
--给test2表插入数据
insert into test2 (id, name) values (1, 'abc'), (2, 'bbbb');
--查询test2表数据
select * from test2;
┌─id─┬─name─┐
│  1 │ abc  │
│  2 │ bbbb │
└────┴──────┘
```

12.4.4.4 SELECT 查询表数据

本章节主要介绍ClickHouse查询表数据的SQL基本语法和使用说明。

基本语法

SELECT [**DISTINCT**] expr_list

[**FROM** [*database_name*.]*table* | (subquery) | table_function] [**FINAL**]

[SAMPLE sample_coeff]

[ARRAY **JOIN** ...]

[**GLOBAL**] [**ANY**|**ALL**|**ASOF**] [**INNER**|**LEFT**|**RIGHT**|**FULL**|**CROSS**] [**OUTER**|SEMI|ANTI] **JOIN** (subquery)|**table** (**ON** <expr_list>)|(**USING** <column_list>)

[PREWHERE expr]

[**WHERE** expr]

[**GROUP BY** expr_list] [**WITH** TOTALS]

[**HAVING** expr]

[**ORDER BY** expr_list] [**WITH** FILL] [**FROM** expr] [**TO** expr] [STEP expr]

[**LIMIT** [offset_value,]n **BY** columns]

[**LIMIT** [n,]m] [**WITH** TIES]

[**UNION ALL** ...]

[**INTO** OUTFILE filename]

[FORMAT format]

使用示例

```
--查看ClickHouse集群信息
select * from system.clusters;
--显示当前节点设置的宏
```

```
select * from system.macros;
--查看数据库容量
select
sum(rows) as "总行数",
formatReadableSize(sum(data_uncompressed_bytes)) as "原始大小",
formatReadableSize(sum(data_compressed_bytes)) as "压缩大小",
round(sum(data_compressed_bytes) / sum(data_uncompressed_bytes) * 100,
0) "压缩率"
from system.parts;
--查询test表容量。where条件根据实际情况添加修改
select
sum(rows) as "总行数",
formatReadableSize(sum(data_uncompressed_bytes)) as "原始大小",
formatReadableSize(sum(data_compressed_bytes)) as "压缩大小",
round(sum(data_compressed_bytes) / sum(data_uncompressed_bytes) * 100,
0) "压缩率"
from system.parts
where table in ('test')
and partition like '2020-11-%'
group by table;
```

12.4.4.5 ALTER TABLE 修改表结构

本章节主要介绍ClickHouse修改表结构的SQL基本语法和使用说明。

基本语法

ALTER TABLE [*database_name*].*name* [**ON CLUSTER** *cluster*] **ADD**|**DROP**|**CLEAR**|**COMMENT**|**MODIFY COLUMN** ...

说明

ALTER仅支持 *MergeTree ， Merge以及Distributed等引擎表。

使用示例

```
--给表t1增加列test01
ALTER TABLE t1 ADD COLUMN test01 String DEFAULT 'defaultvalue';
--查询修改后的表t1
desc t1
┌─name────┬─type───┬─default_type─┬─default_expression─┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ id      │ UInt8  │              │                    │         │                  │                │
│ name    │ String │              │                    │         │                  │                │
│ address │ String │              │                    │         │                  │                │
│ test01  │ String │ DEFAULT      │ 'defaultvalue'     │         │                  │                │
└─────────┴────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘
--修改表t1列name类型为UInt8
ALTER TABLE t1 MODIFY COLUMN name UInt8;
--查询修改后的表t1
desc t1
┌─name────┬─type───┬─default_type─┬─default_expression─┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ id      │ UInt8  │              │                    │         │                  │                │
│ name    │ UInt8  │              │                    │         │                  │                │
│ address │ String │              │                    │         │                  │                │
│ test01  │ String │ DEFAULT      │ 'defaultvalue'     │         │                  │                │
└─────────┴────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘
--删除表t1的列test01
ALTER TABLE t1 DROP COLUMN test01;
--查询修改后的表t1
desc t1
┌─name────┬─type───┬─default_type─┬─default_expression─┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ id      │ UInt8  │              │                    │         │                  │                │
│ name    │ UInt8  │              │                    │         │                  │                │
```

```
│ address │ String │              │                    │         │                  │                │
└─────────┴────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘
```

12.4.4.6 DESC 查询表结构

本章节主要介绍ClickHouse查询表结构的SQL基本语法和使用说明。

基本语法

DESC|**DESCRIBE TABLE** [*database_name*.]*table* [**INTO** OUTFILE filename] [FORMAT format]

使用示例

```
--查询表t1的表结构
desc t1;
┌─name────┬─type─┬─default_type─┬─default_expression ┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ id      │ UInt8 │              │                    │         │                  │                │
│ name    │ UInt8 │              │                    │         │                  │                │
│ address │ String │             │                    │         │                  │                │
└─────────┴────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘
```

12.4.4.7 DROP 删除表

本章节主要介绍ClickHouse删除表的SQL基本语法和使用说明。

基本语法

DROP [**TEMPORARY**] **TABLE** [**IF EXISTS**] [*database_name*.]*name* [**ON CLUSTER** *cluster*]

使用示例

```
--删除表t1
drop table t1;
```

12.4.4.8 SHOW 显示数据库和表信息

本章节主要介绍ClickHouse显示数据库和表信息的SQL基本语法和使用说明。

基本语法

show databases

show tables

使用示例

```
--查询数据库
show databases;
┌─name────┐
│ default │
│ system  │
│ test    │
└─────────┘
--查询表信息
show tables;
┌─name─┐
```

```
┌─────┐
│ t1  │
│ test│
│ test2│
│ test5│
└─────┘
```

12.4.5 ClickHouse 数据迁移

12.4.5.1 ClickHouse 数据导入导出

使用 ClickHouse 客户端导入导出数据

本章节主要介绍使用ClickHouse客户端导入导出文件数据的基本语法和使用说明。

- CSV格式数据导入

  clickhouse client --host *主机名/ClickHouse实例IP地址* **--database** *数据库名* **--port** *端口号* **--secure --format_csv_delimiter="***csv文件数据分隔符***" --query="INSERT INTO** *数据表名* **FORMAT CSV" <** *csv文件所在主机路径*

  使用示例：

  ```
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 --secure --format_csv_delimiter="," --query="INSERT INTO testdb.csv_table FORMAT CSV" < /opt/data.csv
  ```

  数据表需提前创建好。

- CSV格式数据导出

  ⚠ 注意

  导出数据为CSV格式的文件，可能存在CSV注入的安全风险，请谨慎使用。

  ---

  clickhouse client --host *主机名/ClickHouse实例IP地址* **--database** *数据库名* **--port** *端口号* **-m --secure --query="SELECT * FROM** *表名*" > *csv文件导出路径*

  使用示例：

  ```
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="SELECT * FROM test_table" > /opt/test.csv
  ```

- parquet格式数据导入

  cat *parquet格式文件* **| clickhouse client --host** *主机名/ClickHouse实例IP* **--database** *数据库名* **--port** *端口号* **-m --secure --query="INSERT INTO** *表名* **FORMAT Parquet"**

  使用示例：

  ```
  cat /opt/student.parquet | clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="INSERT INTO parquet_tab001 FORMAT Parquet"
  ```

- parquet格式数据导出

  clickhouse client --host *主机名/ClickHouse实例IP* **--database** *数据库名* **--port** *端口号* **-m --secure --query="select * from** *表名* **FORMAT Parquet" >** *parquet格式文件输出路径*

  使用示例：

  ```
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="select * from test_table FORMAT Parquet" > /opt/student.parquet
  ```

- ORC格式数据导入

cat *orc格式文件路径* **| clickhouse client --host** *主机名/ClickHouse实例IP* **--database** *数据库名* **--port** *端口号* **-m --secure --query="INSERT INTO** *表名* **FORMAT ORC"**

使用示例：

```
cat /opt/student.orc | clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="INSERT INTO orc_tab001 FORMAT ORC"
#orc格式文件格式文件数据可以从HDFS中导出，例如：
hdfs dfs -cat /user/hive/warehouse/hivedb.db/emp_orc/000000_0_copy_1 | clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="INSERT INTO orc_tab001 FORMAT ORC"
```

- ORC格式数据导出

  clickhouse client --host *主机名/ClickHouse实例IP* **--database** *数据库名* **--port** *端口* -m **--secure --query="select** * **from** *表名* **FORMAT ORC"** > *输出的ORC格式文件路径*

  使用示例：

  ```
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="select * from csv_tab001 FORMAT ORC" > /opt/student.orc
  ```

- JSON格式数据导入

  INSERT INTO *表名* **FORMAT JSONEachRow** *JSON格式字符串1 JSON格式字符串2*

  使用示例：

  ```
  INSERT INTO test_table001 FORMAT JSONEachRow {"PageViews":5, "UserID":"4324182021466249494", "Duration":146,"Sign":-1}
  {"UserID":"4324182021466249494","PageViews":6,"Duration":185,"Sign":1}
  ```

- JSON格式数据导出

  clickhouse client --host *主机名/ClickHouse实例IP* **--database** *数据库名* **--port** *端口号* **-m --secure --query="SELECT** * **FROM** *表名* **FORMAT JSON|JSONEachRow|JSONCompact|..."** > *json文件输出路径*

  使用示例

  ```
  #导出json
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="SELECT * FROM test_table FORMAT JSON" > /opt/test.json

  #导出json(JSONEachRow)
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="SELECT * FROM test_table FORMAT JSONEachRow" > /opt/test_jsoneachrow.json

  #导出json(JSONCompact)
  clickhouse client --host 10.5.208.5 --database testdb --port 21427 -m --secure --query="SELECT * FROM test_table FORMAT JSONCompact" > /opt/test_jsoncompact.json
  ```

12.4.5.2 将 Kafka 数据同步至 ClickHouse

您可以通过创建Kafka引擎表将Kafka数据自动同步至ClickHouse集群，具体操作详见本章节描述。

前提条件

- 已创建Kafka集群。已安装Kafka客户端。
- 已创建ClickHouse集群，并且ClickHouse集群和Kafka集群在同一VPC下，网络可以互通。已安装ClickHouse客户端。

Kafka 引擎表使用语法说明

- **语法**

```
CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
(
    name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1],
    name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2],
    ...
) ENGINE = Kafka()
SETTINGS
    kafka_broker_list = 'host1:port1,host2:port2',
    kafka_topic_list = 'topic1,topic2,...',
    kafka_group_name = 'group_name',
    kafka_format = 'data_format';
    [kafka_row_delimiter = 'delimiter_symbol',]
    [kafka_schema = '',]
    [kafka_num_consumers = N]
```

- **参数说明**

表 12-55 Kafka 引擎表参数说明

参数名	是否必选	参数说明
kafka_broker_list	是	Kafka集群broker实例的IP和端口列表。例如：*kafka集群broker实例IP1*:**9092**,*kafka集群broker实例IP2*:**9092**,*kafka集群broker实例IP3*:**9092**。 Kafka集群broker实例IP获取方法如下： • MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Kafka”。单击“实例”，查看Kafka角色实例的IP地址。 **说明** 若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。 • MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 > Kafka”。单击“实例”，查看Kafka角色实例的IP地址。
kafka_topic_list	是	Kafka的topic列表。
kafka_group_name	是	Kafka的Consumer Group名称，可以自己指定。
kafka_format	是	Kafka消息体格式。例如JSONEachRow、CSV、XML等。
kafka_row_delimiter	否	每个消息体（记录）之间的分隔符。
kafka_schema	否	如果解析格式需要一个schema时，此参数必填。
kafka_num_consumers	否	单个表的消费者数量。默认值是：1，如果一个消费者的吞吐量不足，则指定更多的消费者。消费者的总数不应该超过topic中分区的数量，因为每个分区只能分配一个消费者。

Kafka 数据同步至 ClickHouse 操作示例

步骤1 参考**使用Kafka客户端**，切换到Kafka客户端安装目录。

1. 以Kafka客户端安装用户，登录Kafka安装客户端的节点。
2. 执行以下命令，切换到客户端安装目录。

   cd /opt/client
3. 执行以下命令配置环境变量。

   source bigdata_env
4. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   a. 如果是MRS 3.1.0版本集群，则需要先执行：**export CLICKHOUSE_SECURITY_ENABLED=true**

   b. **kinit** *组件业务用户*

步骤2 执行以下命令，创建Kafka的Topic。详细的命令使用可以参考**管理Kafka主题**。

kafka-topics.sh --topic *kafkacktest2* **--create --zookeeper** *ZooKeeper角色实例IP:2181*/**kafka --partitions** *2* **--replication-factor** *1*

说明

- **--topic**参数值为要创建的Topic名称，本示例创建的名称为kafkacktest2 。
- **--zookeeper**：ZooKeeper角色实例所在节点IP地址，填写三个角色实例其中任意一个的IP地址即可。ZooKeeper角色实例所在节点IP获取参考如下。
  - MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > ZooKeeper > 实例”。查看ZooKeeper角色实例的IP地址。
  - MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 实例”。查看ZooKeeper角色实例的IP地址。
- **--partitions**主题分区数和**--replication-factor**主题备份个数不能大于Kafka角色实例数量。

步骤3 参考**从零开始使用ClickHouse**登录ClickHouse客户端。

1. 执行以下命令，切换到客户端安装目录。

   cd /opt/Bigdata/client
2. 执行以下命令配置环境变量。

   source bigdata_env
3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限，具体请参见**ClickHouse用户及权限管理**章节，为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行本步骤。

   kinit *组件业务用户*

   例如，**kinit** clickhouseuser。
4. 执行以下命令连接到要导入数据的ClickHouse实例节点。

   clickhouse client --host *ClickHouse的实例IP* **--user** *登录名* **--password** *密码* **--port** *ClickHouse的端口号* **--database** *数据库名* **--multiline**

步骤4 参考**Kafka引擎表使用语法说明**，在ClickHouse中创建Kafka引擎表。例如，如下建表语句在default数据库下，创建表名为kafka_src_tbl3，Topic名为kafkacktest2、消息格式为JSONEachRow的Kafka引擎表。

```
create table kafka_src_tbl3 on cluster default_cluster
(id UInt32, age UInt32, msg String)
ENGINE=Kafka()
SETTINGS
 kafka_broker_list='kafka集群broker实例IP1:9092,kafka集群broker实例IP2:9092,kafka集群broker实例IP3:9092',
 kafka_topic_list='kafkacktest2',
 kafka_group_name='cg12',
 kafka_format='JSONEachRow';
```

步骤5 创建ClickHouse本地复制表。例如，如下创建表名为kafka_dest_tbl3的ReplicatedMergeTree表。

```
create table kafka_dest_tbl3 on cluster default_cluster
( id UInt32, age UInt32, msg String )
engine = ReplicatedMergeTree('/clickhouse/tables/{shard}/default/kafka_dest_tbl3', '{replica}')
partition by age
order by id;
```

步骤6 创建MATERIALIZED VIEW，该视图会在后台转换Kafka引擎中的数据并将其放入创建的ClickHouse表中。

```
create materialized view consumer3 on cluster default_cluster to kafka_dest_tbl3 as select * from
kafka_src_tbl3;
```

步骤7 再次执行**步骤1**，进入Kafka客户端安装目录。

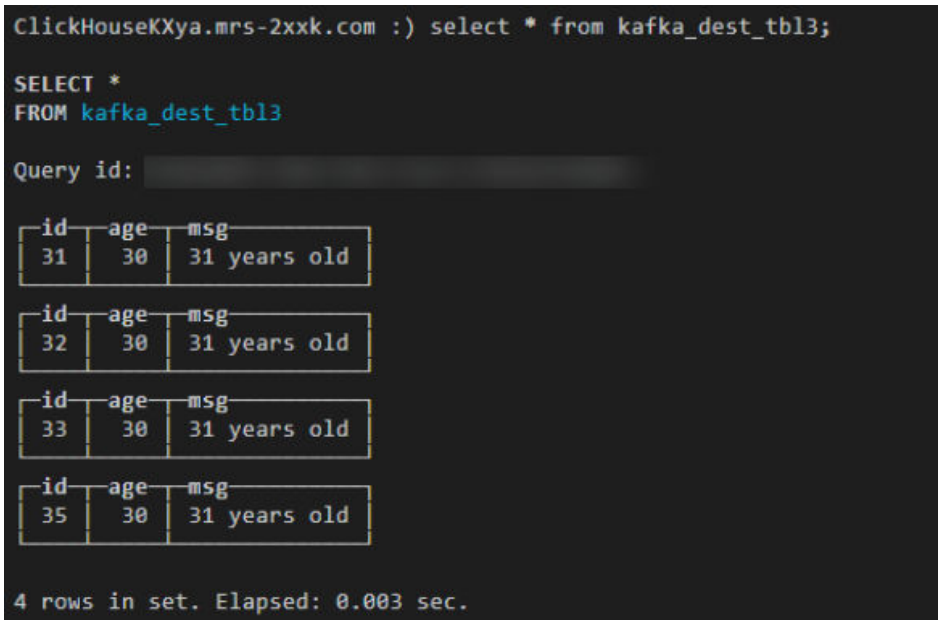
步骤8 执行以下命令，在Kafka的Topic中产生消息。例如，如下命令向**步骤2**中创建的Topic发送消息。

kafka-console-producer.sh --broker-list *kafka集群broker实例IP1*:9092,*kafka集群broker实例IP2*:9092,*kafka集群broker实例IP3*:9092 **--topic** *kafkacktest2*

```
>{"id":31, "age":30, "msg":"31 years old"}
>{"id":32, "age":30, "msg":"31 years old"}
>{"id":33, "age":30, "msg":"31 years old"}
>{"id":35, "age":30, "msg":"31 years old"}
```

步骤9 使用ClickHouse客户端登录**步骤3**中ClickHouse实例节点，查询ClickHouse表数据。例如，查询kafka_dest_tbl3本地复制表，Kafka消息中的数据已经同步到该表。

```
select * from kafka_dest_tbl3;
```



----结束

12.4.5.3 使用 ClickHouse 数据迁移工具

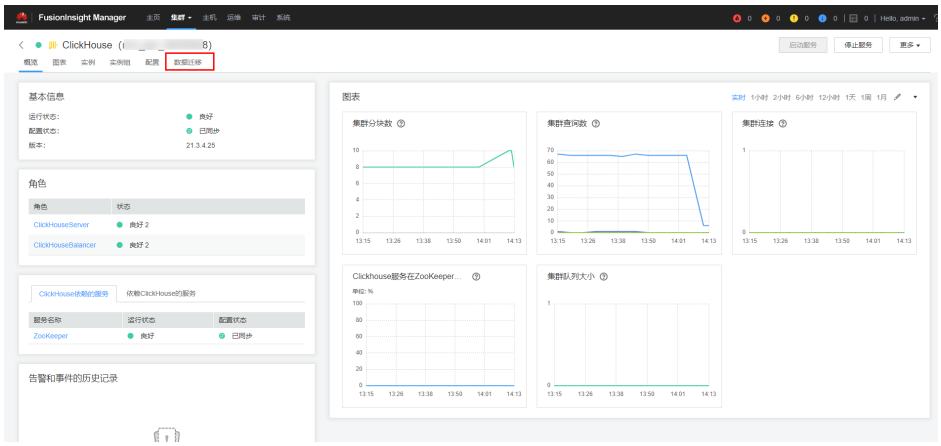
ClickHouse数据迁移工具可以将某几个ClickHouseServer实例节点上的一个或多个MergeTree引擎分区表的部分分区迁移至其他ClickHouseServer节点上相同的表中。在扩容场景中，可以使用该工具将原节点上的部分数据迁移至新增节点上，从而达到扩容后的数据均衡。

前提条件

- ClickHouse服务运行正常，Zookeeper服务运行正常，迁入、迁出节点的ClickHouseServer实例状态正常。
- 请确保迁入节点已有待迁移数据表，且确保该表是MergeTree系列引擎的分区表。
- 创建迁移任务前请确保所有对待迁移数据表的写入任务已停止，且任务启动后，只允许对待迁移数据表进行查询操作，禁止对该表进行写入、删除等操作，否则可能会造成迁移前后数据不一致。
- 迁入节点的ClickHouse数据目录有足够的空间。

操作步骤

步骤1 登录Manager，选择“集群 > 服务 > ClickHouse”， 在ClickHouse服务界面单击“数据迁移”页签，进入数据迁移界面。



步骤2 单击“创建迁移任务”。

步骤3 在创建迁移任务界面，填写迁移任务的相关参数，具体参考如下表12-56。

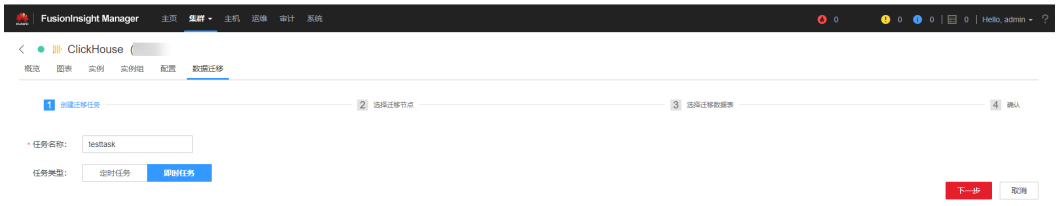
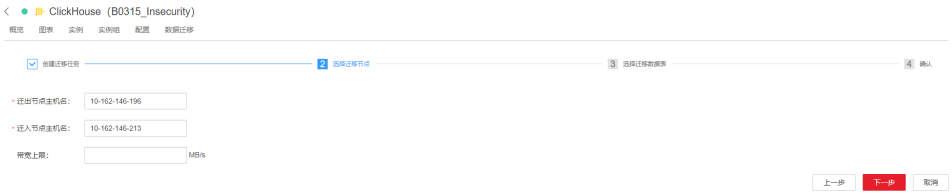


表 12-56 迁移任务参数说明

参数名	参数取值说明
任务名称	填写具体的任务名称。可由字母、数组及下划线组成，长度为1~50位，且不能与已有的迁移任务相同。
任务类型	• 定时任务：选择定时任务时，可以设置“开始时间”参数，设定任务在当前时间以后的某个时间点执行。 • 即时任务：任务启动后立即开始执行。
开始时间	在“任务类型”参数选择“定时任务”时填写，有效值为当前时间以后的某个时间（最长为90天以后）。

步骤4 在选择迁移节点界面，填写“迁入节点主机名”、“迁出节点主机名”，单击“下一步”。



说明

- “迁入节点主机名”与“迁出节点主机名”只能各填写一个主机名，不支持多节点迁移。

  具体的参数值可以在ClickHouse服务界面单击“实例”页签，查看当前ClickHouseServer实例所在“主机名称”列获取。
- “带宽上限”为可选参数，若不填写则为无上限，最大可设置为10000MB/s。

步骤5 在选择迁移数据表界面，单击“数据库”后的 ，选择待迁出节点上存在的数据库，在“数据表”处选择待迁移的数据表，数据表下拉列表中展示的是所选数据库中的MergeTree系列引擎的分区表。“节点信息”中展示的为当前迁入节点、迁出节点上ClickHouse服务数据目录的空间使用情况，单击“下一步”。

步骤6 确认任务信息，确认无误后可以单击“提交”提交任务。

数据迁移工具将根据待迁移数据表的大小自动计算需要迁移的分区，数据迁移量则是计算出的需要迁移的分区总大小。

步骤7 提交迁移任务成功后，单击操作列的“启动”。如果任务类型是即时任务则开始执行任务，如果是定时任务则开始倒计时。

步骤8 迁移任务执行过程中，可单击“取消”取消正在执行的迁移任务，若取消任务，则会回退掉迁入节点上已迁移的数据。

可以单击“更多 > 详情”查看迁移过程中的日志信息。

步骤9 迁移完成后，选择“更多 > 结果”查看迁移结果；选择“更多 > 删除”清理ZooKeeper以及迁出节点上该迁移任务相关的目录。

----结束

12.4.6 用户管理及认证

12.4.6.1 ClickHouse 用户及权限管理

用户权限模型

ClickHouse用户权限管理实现了对集群中各个ClickHouse实例上用户、角色、权限的统一管理。通过Manager UI的权限管理模块进行创建用户、创建角色、绑定ClickHouse访问权限配置等操作，通过用户绑定角色的方式，实现用户权限控制。

管理资源：Clickhouse权限管理支持的资源如表12-57所示。

资源权限：ClickHouse支持的资源权限如表12-58所示。

表 12-57 ClickHouse 支持的权限管理对象

资源列表	是否集成	备注
数据库	是（一级）	-
表	是（二级）	-
视图	是（二级）	与表一致

表 12-58 资源权限列表

资源对象	可选权限	备注
数据库（DATABASE）	CREATE	CREATE DATABASE/TABLE/VIEW/DICTIONARY权限
表/视图（TABLE/VIEW）	SELECT/INSERT	-

前提条件

- ClickHouse服务运行正常，Zookeeper服务运行正常。
- 用户在集群中创建数据库或者表时使用ON CLUSTER语句，保证各个ClickHouse节点上数据库、表的元信息相同。

说明

ClickHouse赋权成功后，权限生效时间大约为1分钟。

添加 ClickHouse 角色

步骤1 登录Manager，选择“系统 > 权限 > 角色”，在“角色”界面单击“添加角色”按钮，进入添加角色页面。

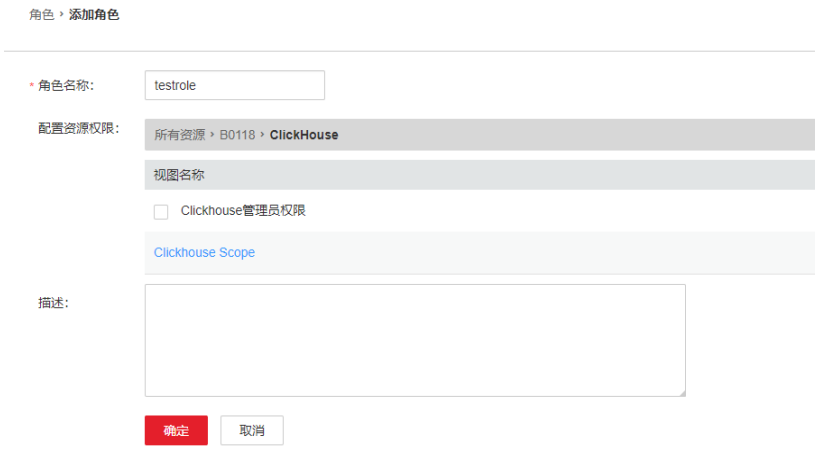
步骤2 在添加角色界面输入“角色名称”，在配置资源权限处单击集群名称，进入服务列表页面，单击ClickHouse服务，进入ClickHouse权限资源页面。

根据业务需求确定是否要创建具有管理员权限的角色。

说明

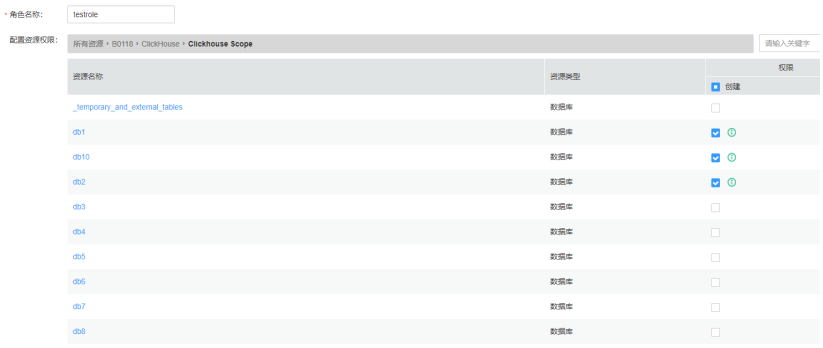
- ClickHouse管理员权限为：除去对user/role的创建、删除和修改之外的所有数据库操作权限。
- 对于用户和角色的管理，仅有ClickHouse的内置用户**clickhouse**具有权限。

- 是，执行**步骤3**。
- 否，执行**步骤4**。



步骤3 勾选“ClickHouse管理员权限”，单击“确定”操作结束。

步骤4 单击“ClickHouse Scope”，进入ClickHouse数据库资源列表。勾选“创建”权限，则该角色将拥有该数据库下的创建（CREATE）权限。



根据业务需求确定是否赋权。

- 是，单击“确定”操作结束。
- 否，执行**步骤5**。

步骤5 单击“资源名称 > *待操作的数据库资源名称*”，进入表、视图页面，根据业务需要，勾选“读”（SELECT权限）或者“写”（INSERT权限）权限，单击“确定”。

----结束

添加用户并将 ClickHouse 对应角色绑定到该用户

步骤1 登录Manager，选择“系统 > 权限 > 用户”，单击“添加用户”，进入添加用户页面。

步骤2 “用户类型”选择“人机”，在“密码”和“确认密码”参数设置该用户对应的密码。

说明

- 用户名：添加的用户名不能包含字符“-”，否则会导致认证失败。
- 密码：设置的密码不能携带“$”、“.”、“#”特殊字符，否则会导致认证失败。

步骤3 在“角色”处单击“添加”，在弹框中选择具有ClickHouse权限的角色，单击“确定”添加到角色，单击“确定”完成操作。

步骤4 登录ClickHouse客户端安装节点，使用新添加的用户及设置的密码连接ClickHouse服务。

- 执行以下命令，切换到客户端安装目录。

  cd /opt/*客户端安装目录*

- 执行以下命令配置环境变量。

  source bigdata_env

- 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限，具体请参见**添加ClickHouse角色**，为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行本步骤。

  a. 如果是MRS 3.1.0版本集群，则需要先执行：**export CLICKHOUSE_SECURITY_ENABLED=true**

  b. **kinit** *组件业务用户*

- 使用新添加的用户登录验证。

  clickhouse client --host *ClickHouse的实例IP* **--multiline --user** ***步骤1****中添加的用户* **--password** ***步骤2****中设置的用户密码* **--port** *ClickHouse的端口号* **--secure**

----结束

异常场景下登录客户端操作赋权

ClickHouse集群默认每个节点上的表元信息是相同的，因此在Manager的权限管理页面上默认采集的是任意ClickHouse节点的表信息，如果有个别节点上创建DATABASE/TABLE时未使用ON CLUSTER语句，则权限操作可能无法展示该资源，不保证可以对其赋权。对于这样单个ClickHouse节点中的本地表，如果需要赋权，则可以通过后台客户端进行操作。

说明

以下操作，需要提前获取到需要赋权的角色、数据库或表名称、对应的ClickHouseServer实例所在的节点IP。

- ClickHouseServer的实例IP地址可登录集群FusionInsight Manager，然后选择“集群 > 服务 > ClickHouse > 实例”，获取ClickHouseServer实例对应的业务IP地址。
- 系统域名：默认为hadoop.com。可登录集群FusionInsight Manager，单击“系统 > 权限 > 域和互信”，“本端域”参数值即为系统域名。在执行命令时改为小写。

步骤1 以root用户登录ClickHouseServer实例所在的节点。

步骤2 执行以下命令获取“clickhouse.keytab”文件路径。

ls ${BIGDATA_HOME}/FusionInsight_ClickHouse_*/install/FusionInsight-ClickHouse-*/clickhouse/keytab/clickhouse.keytab

步骤3 以客户端安装用户，登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤5 执行以下命令配置环境变量。

source bigdata_env

如果是MRS 3.1.0版本集群并且集群已启用Kerberos认证，则还需要执行：

export CLICKHOUSE_SECURITY_ENABLED=true

步骤6 执行如下命令使用客户端命令连接ClickHouseServer实例。

如果当前集群已启用Kerberos认证，执行以下命令：

clickhouse client --host *ClickHouseServer实例所在节点IP* **--user clickhouse/hadoop.***<系统域名>* **--password** ***步骤2*****中获取的***clickhouse.keytab路径* **--port** *ClickHouse的端口号* **--secure**

如果当前集群未启用Kerberos认证，执行以下命令：

clickhouse client --host *ClickHouseServer实例所在节点IP* **--user clickhouse --port** *ClickHouse的端口号*

步骤7 对某DATABASE进行赋权操作，执行如下命令。

授权操作语法，其中*DATABASE*为要操作的数据库名称，*role*为需要操作的角色。

GRANT [ON CLUSTER *cluster_name***]** *privilege* **ON** *{DATABASE|TABLE}* **TO** *{user | role]*

例如，给用户testuser授予数据库t2的CREATE权限：

GRANT CREATE ON *m2* **to** *testuser***;**

步骤8 对TABLE/VIEW进行赋权操作，执行如下命令，其中*TABLE*为要操作的表或视图名称，*user*为需要操作的角色。

对某数据库下的表赋予查询权限：

GRANT SELECT ON *TABLE* **TO** *user;*

对某数据库下的表赋予写入权限：

GRANT INSERT ON *TABLE* **TO** *user;*

说明

更多ClickHouse授权操作及详细权限说明可参考**https://clickhouse.tech/docs/zh/sql-reference/statements/grant/**。

步骤9 执行如下命令，退出客户端。

quit;

----结束

12.4.6.2 ClickHouse 使用 OpenLDAP 认证

ClickHouse支持和OpenLDAP进行对接，通过在ClickHouse上添加OpenLDAP服务器配置和创建用户，实现帐号和权限的统一集中管理和权限控制等操作。此方案适合从OpenLDAP服务器中批量向ClickHouse中导入用户。

本章节操作仅支持MRS 3.1.0及以上集群版本。

前提条件

- MRS集群及ClickHouse实例运行正常，已安装ClickHouse客户端。
- OpenLDAP已安装且状态正常。

对接 OpenLDAP 服务器创建 ClickHouse 用户

步骤1 登录集群Manager页面，选择“集群 > 服务 > ClickHouse > 配置 > 全部配置”。

步骤2 选择“ClickHouseServer（角色）> 自定义”，在“clickhouse-config-customize”配置项中添加如下OpenLDAP配置参数。

表 12-59 OpenLDAP 参数说明

参数名	参数值说明	参数取值参考
ldap_servers.ldap_server_name.host	OpenLDAP服务器主机名或IP，不能为空。	localhost
ldap_servers.ldap_server_name.port	OpenLDAP服务器端口。 如果enable_tls参数设置为true，则默认端口号为636，否则为389。	636
ldap_servers.ldap_server_name.auth_dn_prefix	用于构造要绑定到的DN的前缀和后缀。 生成的DN将被构造为auth_dn_prefix + escape(user_name) + auth_dn_suffix字符串。 auth_dn_suffix通常应将逗号“,”作为其第一个非空格字符。	uid=
ldap_servers.ldap_server_name.auth_dn_suffix		,ou=Group,dc=node1,dc=com

参数名	参数值说明	参数取值参考
ldap_servers.ldap_server_name.enable_tls	触发使用OpenLDAP服务器安全连接的标志。 • 纯文本（ldap://）协议指定“no”（不推荐）。 • LDAP over SSL/TLS (ldaps://)协议指定“yes”。	yes
ldap_servers.ldap_server_name.tls_require_cert	SSL/TLS对端证书校验行为。 取值范围为： 'never'、'allow'、'try'、'require'。	allow

说明

其他参数说明详细可以参考**<ldap_servers>配置参数详解**。

步骤3 添加完配置后，单击“保存”，在弹出对话框中单击“确定”，配置保存成功后，单击“完成”。

步骤4 Manager页面，单击“实例”，选择ClickHouseServer实例，单击“更多 > 重启实例 ”，弹出对话框输入密码，单击“确定”。重启实例对话框，单击“确定”，根据界面提示信息确认实例重启成功，单击“完成”重启操作完成。

步骤5 登录ClickHouseServer实例所在主机节点，进入“${BIGDATA_HOME}/FusionInsight_ClickHouse_*版本号*/*x_x*_ClickHouseServer/etc”目录。

cd ${BIGDATA_HOME}/FusionInsight_ClickHouse_*/*x_x*_ClickHouseServer/etc

步骤6 执行以下命令，查看配置文件config.xml，确认OpenLDAP参数是否配置成功。

cat config.xml

```
[root@k                3 etc]# cat config.xml
<yandex>
  <ldap_servers>
    <ldap_server_name>
      <auth_dn_prefix>uid=</auth_dn_prefix>
      <port>636</port>
      <host>localhost</host>
      <enable_tls>yes</enable_tls>
      <tls_require_cert>allow</tls_require_cert>
      <auth_dn_suffix>,ou=Group,dc=node1,dc=com</auth_dn_suffix>
    </ldap_server_name>
  </ldap_servers>
  <zookeeper>
```

步骤7 以root用户登录ClickHouseServer实例所在的节点。

步骤8 执行以下命令获取“clickhouse.keytab”文件路径。

ls ${BIGDATA_HOME}/FusionInsight_ClickHouse_*/install/FusionInsight-ClickHouse-*/clickhouse/keytab/clickhouse.keytab

步骤9 以客户端安装用户，登录安装客户端的节点。

步骤10 执行以下命令，切换到ClickHouse客户端安装目录。

cd /opt/client

步骤11 执行以下命令配置环境变量。

source bigdata_env

步骤12 执行如下命令使用客户端命令连接ClickHouseServer实例。

- 如果当前集群已启用Kerberos认证，使用clickhouse.keytab连接ClickHouseServer实例：

  clickhouse client --host *ClickHouseServer实例所在节点IP* **--user clickhouse/hadoop.***<系统域名>* **--password** ***步骤8****中获取的clickhouse.keytab路径* **--port** *ClickHouse的端口号*

  说明

  系统域名：默认为hadoop.com。具体可登录集群FusionInsight Manager，单击“系统 > 权限 > 域和互信”，“本端域”参数值即为系统域名。在执行命令时改为小写。

- 如果当前集群未启用Kerberos认证，使用clickhouse管理员用户连接ClickHouseServer实例：

  clickhouse client --host *ClickHouseServer实例所在节点IP* **--user clickhouse --port** *ClickHouse的端口号*

步骤13 创建OpenLDAP中的普通用户。

如以下语句，在集群default_cluster上创建testUser用户，设置ldap_server为**步骤6**中<ldap_servers>标签下的OpenLDAP服务名，本示例为ldap_server_name。

CREATE USER *testUser* **ON CLUSTER** *default_cluster* **IDENTIFIED WITH ldap_server BY** 'ldap_server_name'**;**

testUser用户为OpenLDAP中已有的用户名，请根据实际情况修改。

步骤14 退出客户端，使用新建的用户登录验证配置是否成功。

exit;

clickhouse client --host *ClickHouseServer实例IP* **--user** *testUser* **--password** *testUser对应的密码* **--port** *ClickHouse的端口号*

----结束

<ldap_servers>配置参数详解

- host

  OpenLDAP服务器主机名或IP，必选参数，不能为空。

- port

  OpenLDAP服务器端口，如果enable_tls参数设置为true，则默认为636，否则为389。

- auth_dn_prefix，auth_dn_suffix

  用于构造要绑定到的DN的前缀和后缀。

  实际上，生成的DN将被构造为auth_dn_prefix + escape(user_name) + auth_dn_suffix字符串。

  注意，这意味着auth_dn_suffix通常应将逗号“，”作为其第一个非空格字符。

- enable_tls

  触发使用OpenLDAP服务器安全连接的标志。

为纯文本（ldap:// ）协议指定“no”（不推荐）。
为LDAP over SSL/TLS (ldaps://)协议指定“yes”（建议为默认值）。

- tls_minimum_protocol_version
  SSL/TLS的最小协议版本。
  接受的值是： 'ssl2'、'ssl3'、'tls1.0'、'tls1.1'、'tls1.2'（默认值）。
- tls_require_cert
  SSL/TLS对端证书校验行为。
  接受的值是： 'never'、'allow'、'try'、'require'（默认值）。
- tls_cert_file
  证书文件。
- tls_key_file
  证书密钥文件。
- tls_ca_cert_file
  CA证书文件。
- tls_ca_cert_dir
  CA证书所在的目录。
- tls_cipher_suite
  允许加密套件。

12.4.7 通过数据文件备份恢复 ClickHouse 数据

操作场景

本章节主要介绍通过把ClickHouse中的表数据导出到CSV文件进行备份，后续可以通过备份的CSV文件数据再进行恢复操作。

前提条件

- 已安装ClickHouse客户端。
- 在Manager已创建具有ClickHouse相关表权限的用户。
- 已准备好备份服务器。

备份数据

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限。如果当前集群未启用Kerberos认证，则无需执行本步骤。

1. 如果是MRS 3.1.0版本集群，则需要先执行：
   export CLICKHOUSE_SECURITY_ENABLED=true

2. **kinit** *组件业务用户*

   例如，**kinit** clickhouseuser。

步骤5 执行ClickHouse组件的客户端命令，将要备份ClickHouse表数据导出到指定目录下。

clickhouse client --host *主机名/实例IP* **--secure --port 21427 --query=**"*表查询语句*" > *输出的csv格式文件路径*

例如，如下是在ClickHouse实例10.244.225.167下备份test表数据到default_test.csv文件中。

clickhouse client --host 10.244.225.167 --secure --port 21427 --query="select * from default.test FORMAT CSV" > /opt/clickhouse/default_test.csv

步骤6 将导出的csv数据文件上传至备份服务器。

----结束

恢复数据

步骤1 将备份服务器上的备份数据文件上传到ClickHouse客户端所在目录。

例如，上传default_test.csv备份文件到：/opt/clickhouse目录下。

步骤2 以客户端安装用户，登录安装客户端的节点。

步骤3 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤4 执行以下命令配置环境变量。

source bigdata_env

步骤5 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限。如果当前集群未启用Kerberos认证，则无需执行本步骤。

1. 如果是MRS 3.1.0版本集群，则需要先执行**:**

   export CLICKHOUSE_SECURITY_ENABLED=true

2. **kinit** ***组件业务用户***

   例如，**kinit** clickhouseuser。

步骤6 执行ClickHouse组件的客户端命令，登录ClickHouse集群。

clickhouse client --host *主机名/实例IP* **--secure --port 21427**

步骤7 创建与CSV备份数据文件格式对应的表。

CREATE TABLE [IF NOT EXISTS] *[database_name.]table_name* **[ON CLUSTER** *Cluster名***]**

(

name1 **[***type1***] [DEFAULT|***materialized***|ALIAS** *expr1***],**

name2 **[***type2***] [DEFAULT|***materialized***|ALIAS** *expr2***],**

...

) *ENGINE = engine*

步骤8 将备份数据文件中的内容导入到**步骤7**创建的表中进行数据恢复。

clickhouse client --host *主机名/实例IP* **--secure --port 21427 --query="insert into** *表信息* **FORMAT CSV"** < *csv文件路径*

例如，如下在ClickHouse实例10.244.225.167中，恢复default_test.csv备份文件数据到test_cpy表中。

clickhouse client --host 10.244.225.167 --secure --port 21427 --query="insert into default.test_cpy FORMAT CSV" < **/opt/clickhouse/default_test.csv**

----结束

12.4.8 ClickHouse 日志介绍

日志描述

日志路径：ClickHouse相关日志的默认存储路径为“${BIGDATA_LOG_HOME}/clickhouse”。

日志归档规则：ClickHouse日志启动了自动压缩归档功能，缺省情况下，当日志大小超过100MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>.[编号].gz”。默认最多保留最近的10个压缩文件，压缩文件保留个数可以在Manager界面中配置。

表 12-60 ClickHouse 日志列表

日志类型	日志文件名	描述
运行日志	/var/log/Bigdata/clickhouse/clickhouseServer/clickhouse-server.err.log	ClickHouseServer服务运行错误日志文件路径。
	/var/log/Bigdata/clickhouse/clickhouseServer/checkService.log	ClickHouseServer服务运行关键日志文件路径。
	/var/log/Bigdata/clickhouse/clickhouseServer/clickhouse-server.log	
	/var/log/Bigdata/clickhouse/balance/start.log	ClickHouseBalancer服务启动日志文件路径。
	/var/log/Bigdata/clickhouse/balance/error.log	ClickHouseBalancer服务运行错误日志文件路径。
	/var/log/Bigdata/clickhouse/balance/access_http.log	ClickHouseBalancer服务运行日志文件路径。
数据迁移日志	/var/log/Bigdata/clickhouse/migration/数据迁移任务名/clickhouse-copier_{timestamp}_{processId}/copier.log	参考**使用ClickHouse数据迁移工具**使用迁移工具时产生的运行日志。
	/var/log/Bigdata/clickhouse/migration/数据迁移任务名/clickhouse-copier_{timestamp}_{processId}/copier.err.log	参考**使用ClickHouse数据迁移工具**使用迁移工具时产生的错误日志。

日志级别

ClickHouse提供了如表12-61所示的日志级别。

运行日志的级别优先级从高到低分别是error、warning、trace、information、debug，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-61 日志级别

日志类型	级别	描述
运行日志	error	error表示系统运行的错误信息。
	warning	warning表示当前事件处理存在异常信息。
	trace	trace表示当前事件处理跟踪信息。
	information	information表示记录系统及各事件正常运行状态信息。
	debug	debug表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 登录FusionInsight Manager系统。

步骤2 选择“集群 > 服务 > ClickHouse > 配置”。

步骤3 单击“全部配置”。

步骤4 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤5 选择所需修改的日志级别。

步骤6 单击“保存”，然后单击“确定”，成功后配置生效。

----结束

说明

配置完成后即生效，不需要重启服务。

日志格式

ClickHouse的日志格式如下所示：

表 12-62 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level><产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2021.02.23 15:26:30.691301 [6085] {} <Error> DynamicQueryHandler: Code: 516, e.displayText() = DB::Exception: default: Authentication failed: password is incorrect or there is no user with such name, Stack trace (when copying this message, always include the lines below): 0. Poco::Exception::Exception(std::__1::basic_string<char, std::__1::char_traits<char>, std::__1::allocator<char> > const&, int) @ 0x1250e59c

12.5 使用 DBService

12.5.1 DBService 日志介绍

日志描述

日志存储路径：DBService相关日志的默认存储路径为“/var/log/Bigdata/dbservice”。

- gaussDB：“/var/log/Bigdata/dbservice/DB”（gaussDB运行日志目录），“/var/log/Bigdata/dbservice/scriptlog/gaussdbinstall.log”（gaussDB安装日志），“/var/log/gaussdbuninstall.log”（gaussDB卸载日志）。
- HA：“/var/log/Bigdata/dbservice/ha/runlog”（HA运行日志目录），“/var/log/Bigdata/dbservice/ha/scriptlog”（HA脚本日志目录）。
- DBServer：“/var/log/Bigdata/dbservice/healthCheck”（服务进程健康状态检查日志目录）。

  “/var/log/Bigdata/dbservice/scriptlog”（运行日志目录），“/var/log/Bigdata/audit/dbservice/”（审计日志目录）。

日志归档规则：DBService的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过1MB的时候，会自动压缩，压缩后的日志文件名规则为：“*<原有日志名>_<编号>*.gz”。最多保留最近的20个压缩文件。

说明

日志归档规则用户不能修改。

表 12-63 DBService 日志列表

日志类型	日志文件名	描述
DBServer运行相关日志	dbservice_serviceCheck.log	服务检查脚本运行日志
	dbservice_processCheck.log	进程检查脚本运行日志
	backup.log	备份恢复操作运行日志（需执行DBService备份恢复操作）
	checkHaStatus.log	HA检查日志
	cleanupDBService.log	卸载日志（需执行DBService卸载日志操作）
	componentUserManager.log	数据库用户添加删除操作日志 （需添加依赖DBService的服务）
	install.log	安装日志
	preStartDBService.log	预启动日志
	start_dbserver.log	DBServer启动操作日志（需执行启动DBService服务的操作）
	stop_dbserver.log	DBServer停止操作日志（需执行停止DBService服务的操作）
	status_dbserver.log	DBServer状态检查日志（需执行**$DBSERVICE_HOME/sbin/status-dbserver.sh**）
	modifyPassword.log	DBService修改密码脚本运行日志（需执行**$DBSERVICE_HOME/sbin/modifyDBPwd.sh**）

日志类型	日志文件名	描述
	modifyDBPwd_yyyy-mm-dd.log	修改密码工具运行日志（需执行 **$DBSERVICE_HOME/sbin/modifyDBPwd.sh**）
	dbserver_switchover.log	DBServer执行主备倒换脚本的日志（需执行主备倒换操作）
GAUSSDB运行日志	gaussdb.log	记录数据库运行信息
	gs_ctl-current.log	记录gs_ctl工具的操作
	gs_guc-current.log	记录gs_guc工具的操作，主要是参数修改
	gaussdbinstall.log	gaussDB安装日志
	gaussdbuninstall.log	gaussDB卸载日志
HA脚本相关运行日志	floatip_ha.log	Floatip资源脚本日志
	gaussDB_ha.log	gaussDB资源脚本日志
	ha_monitor.log	HA进程监控日志
	send_alarm.log	告警发送日志
	ha.log	HA运行日志
DBService审计日志	dbservice_audit.log	dbservice操作审计日志（例如：备份恢复操作）

日志格式

DBService的日志格式如下所示：

表 12-64 日志格式

日志类型	格式	示例
运行日志	[<yyyy-MM-dd HH:mm:ss>] <Log Level>: [<*产生该日志的脚本名称*: *行号*>]: <*log中的message*>	[2020-12-19 15:56:42] INFO [postinstall.sh:653] Is cloud flag is false. (main)

日志类型	格式	示例
审计日志	[<yyyy-MM-dd HH:mm:ss,SSS>] UserName:<*用户名称*> UserIP:<*用户IP*> Operation:<*操作内容*> Result:<*操作结果*> Detail:<*具体信息*>	[2020-05-26 22:00:23] UserName:omm UserIP:192.168.10.21 Operation:DBService data backup Result: SUCCESS Detail: DBService data backup is successful.

12.6 使用 Flink

12.6.1 从零开始使用 Flink

本节提供使用Flink运行wordcount作业的操作指导。

前提条件

- MRS集群中已安装Flink组件。
- 集群正常运行，已安装集群客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

使用 Flink 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行如下命令初始化环境变量。

source /opt/hadoopclient/bigdata_env

步骤4 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 登录Manager，下载认证凭据。

   登录集群的Manager界面，具体请参见**访问MRS Manager（MRS 3.x之前版本）**，选择“系统设置 > 用户管理”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/hadoopclient/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。
4. 配置安全认证，在“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

例如：

security.kerberos.login.keytab: /opt/hadoopclient/Flink/flink/conf/user.keytab

security.kerberos.login.principal: test

5. 参考“组件操作指南 > 使用Flink > 参考 > 签发证书样例”章节生成“generate_keystore.sh”脚本并放置在Flink的客户端bin目录下，执行如下命令进行安全加固，请参考“组件操作指南 > 使用Flink > 安全加固 > 认证和加密”，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中关于SSL的值，针对MRS2.x及之前版本，安全集群默认没有开启外部SSL，用户如果需要启用外部SSL，请参考“组件操作指南 > 使用Flink > 安全加固”进行配置后再次运行该脚本即可。

   说明

   - generate_keystore.sh脚本无需手动生成。
   - 执行认证和加密后会将生成的flink.keystore、flink.truststore、security.cookie自动填充到“flink-conf.yaml”对应配置项中。

6. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/hadoopclient/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/hadoopclient/Flink/flink/conf/”目录下新建目录，例如ssl。

        cd /opt/hadoopclient/Flink/flink/conf/

        mkdir ssl

     ii. 移动flink.keystore和flink.truststore文件到“/opt/hadoopclient/Flink/flink/conf/ssl/”中。

        mv flink.keystore ssl/

        mv flink.truststore ssl/

     iii. 修改flink-conf.yaml文件中如下两个参数为相对路径。

```
security.ssl.internal.keystore: ssl/flink.keystore
security.ssl.internal.truststore: ssl/flink.truststore
```

步骤5 运行wordcount作业。

须知

用户在Flink提交作业或者运行作业时，应具有如下权限：

- 如果启用Ranger鉴权，当前用户必须属于hadoop组或者已在Ranger中为该用户添加“/flink”的读写权限。
- 如果停用Ranger鉴权，当前用户必须属于hadoop组。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。

yarn-session.sh -nm "*session-name*"

flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

- 执行如下命令在Yarn上提交单个作业。

  flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。

      yarn-session.sh -nm "*session-name*"

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - flink.keystore和flink.truststore文件路径为相对路径时：
    - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/hadoopclient/Flink/flink/conf/”，则在“opt/hadoopclient/Flink/flink/conf/”目录下执行命令。

      yarn-session.sh -t ssl/ -nm "*session-name*"

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster -yt ssl/ /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

步骤6 作业提交成功后，客户端界面显示如下。

图 12-5 在 Yarn 上提交作业成功

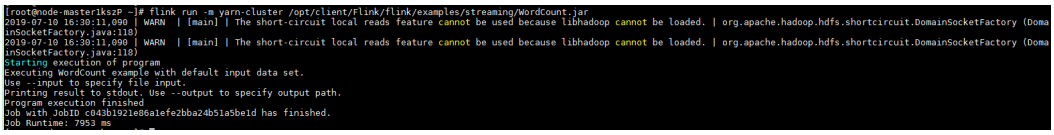


图 12-6 启动 session 成功

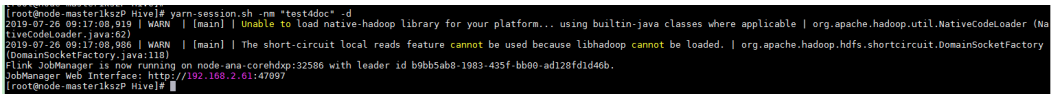
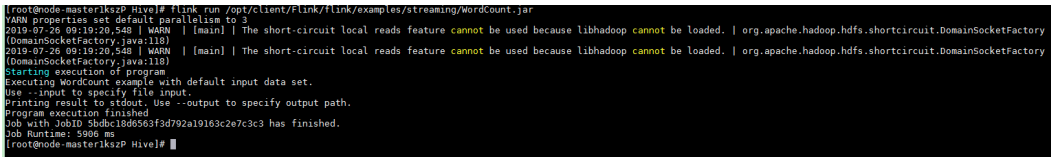
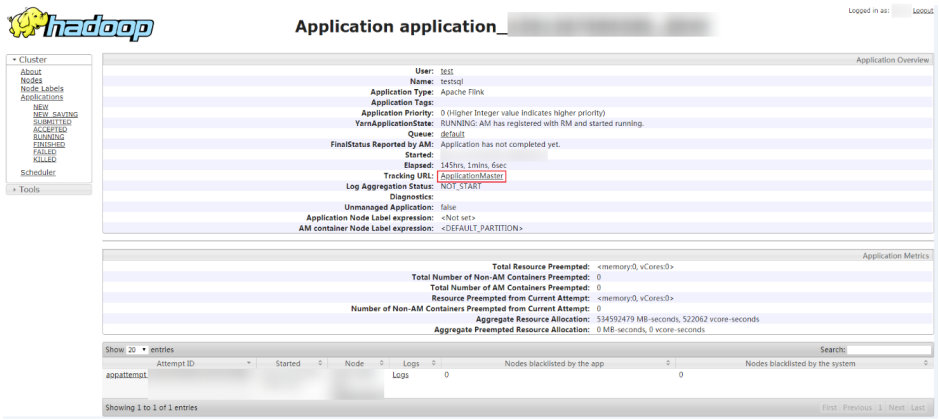


图 12-7 在 session 中提交作业成功



步骤7 使用运行用户进入Yarn服务的原生页面，具体操作参考“组件操作指南 > 使用Flink > 查看Flink作业”，找到对应作业的application，单击application名称，进入到作业详情页面。

- 若作业尚未结束，可单击“Tracking URL”链接进入到Flink的原生页面，查看作业的运行信息。
- 若作业已运行结束，对于在session中提交的作业，可以单击“Tracking URL”链接登录Flink原生页面查看作业信息。



----结束

使用 Flink 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行如下命令初始化环境变量。

source /opt/hadoopclient/bigdata_env

步骤4 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 登录Manager，下载认证凭据。

   登录集群的Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“系统 > 权限 > 用户”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/hadoopclient/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。
4. 将客户端安装节点的业务IP、Manager的浮动IP和Master节点IP添加到配置文件“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中的“jobmanager.web.access-control-allow-origin”和“jobmanager.web.allow-access-address”配置项中，IP地址之间使用英文逗号分隔。

   jobmanager.web.access-control-allow-origin: *xx.xx.xxx.xxx,xx.xx.xxx.xxx,xx.xx.xxx.xxx*
   jobmanager.web.allow-access-address: *xx.xx.xxx.xxx,xx.xx.xxx.xxx,xx.xx.xxx.xxx*

说明

- 客户端安装节点的业务IP获取方法：
  - 集群内节点：

    登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中当前的集群并单击集群名，进入集群信息页面。

    在“节点管理”中查看安装客户端所在的节点IP。
  - 集群外节点：安装客户端所在的弹性云服务器的IP。
- Manager的浮动IP获取方法：
  - 登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中当前的集群并单击集群名，进入集群信息页面。

    在“节点管理”中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。
  - 远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示MRS Manager浮动IP地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到MRS Manager的浮动IP地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。

5. 配置安全认证，在“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/hadoopclient/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test

6. 参考“组件操作指南 > 使用Flink > 参考 > 签发证书样例”章节生成“generate_keystore.sh”脚本并放置在Flink的客户端bin目录下，执行如下命令进行安全加固，请参考“组件操作指南 > 使用Flink > 安全加固 > 认证和加密”，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中关于SSL的值。

   sh generate_keystore.sh <*password*>

说明

执行认证和加密后会在Flink客户端的“conf”目录下生成“flink.keystore”和“flink.truststore”文件，并且在客户端配置文件“flink-conf.yaml”中将以下配置项进行了默认赋值：

- 将配置项“security.ssl.keystore”设置为“flink.keystore”文件所在绝对路径。
- 将配置项“security.ssl.truststore”设置为“flink.truststore”文件所在的绝对路径。
- 将配置项“security.cookie”设置为“generate_keystore.sh”脚本自动生成的一串随机规则密码。
- 默认“flink-conf.yaml”中“security.ssl.encrypt.enabled: false”，“generate_keystore.sh”脚本将配置项“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值设置为调用“generate_keystore.sh”脚本时输入的密码。
- MRS 3.1.0及之后版本，如果需要使用密文时，设置“flink-conf.yaml”中“security.ssl.encrypt.enabled: true”，“generate_keystore.sh”脚本不会配置“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值，需要使用Manager明文加密API进行获取，执行**curl -k -i -u** *user name*:*password* **-X POST -HContent-type:application/json -d '{"plainText":"***password***"}' 'https://***x.x.x.x***:28443/web/api/v2/tools/encrypt'**

  其中*user name*:password分别为当前系统登录用户名和密码；**"plainText"**的password为调用“generate_keystore.sh”脚本时的密码；x.x.x.x为集群Manager的浮动IP。

7. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/hadoopclient/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/hadoopclient/Flink/flink/conf/”目录下新建目录，例如ssl。

     cd /opt/hadoopclient/Flink/flink/conf/

     mkdir ssl

     ii. 移动flink.keystore和flink.truststore文件到“/opt/hadoopclient/Flink/flink/conf/ssl/”中。

     mv flink.keystore ssl/

     mv flink.truststore ssl/

     iii. 修改flink-conf.yaml文件中如下两个参数为相对路径。

```
security.ssl.keystore: ssl/flink.keystore
security.ssl.truststore: ssl/flink.truststore
```

步骤5 运行wordcount作业。

须知

用户在Flink提交作业或者运行作业时，应具有如下权限：

- 如果启用Ranger鉴权，当前用户必须属于hadoop组或者已在Ranger中为该用户添加“/flink”的读写权限。
- 如果停用Ranger鉴权，当前用户必须属于hadoop组。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。

    yarn-session.sh -nm "*session-name***"**

    flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - 执行如下命令在Yarn上提交单个作业。

    flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。

      yarn-session.sh -nm "*session-name***"**

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - flink.keystore和flink.truststore文件路径为相对路径时：
    - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/hadoopclient/Flink/flink/conf/”，则在“opt/hadoopclient/Flink/flink/conf/”目录下执行命令。

      yarn-session.sh -t ssl/ -nm "*session-name***"**

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster -yt ssl/ /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

步骤6 作业提交成功后，客户端界面显示如下。

图 12-8 在 Yarn 上提交作业成功

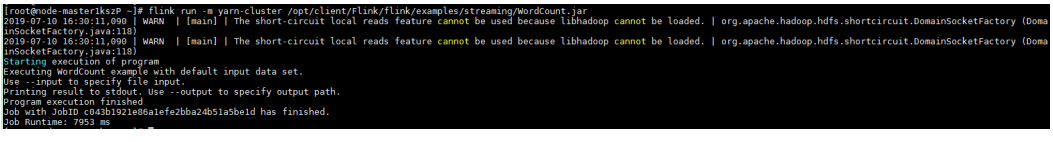
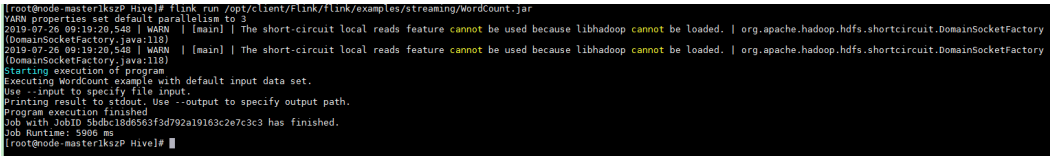


图 12-9 启动 session 成功

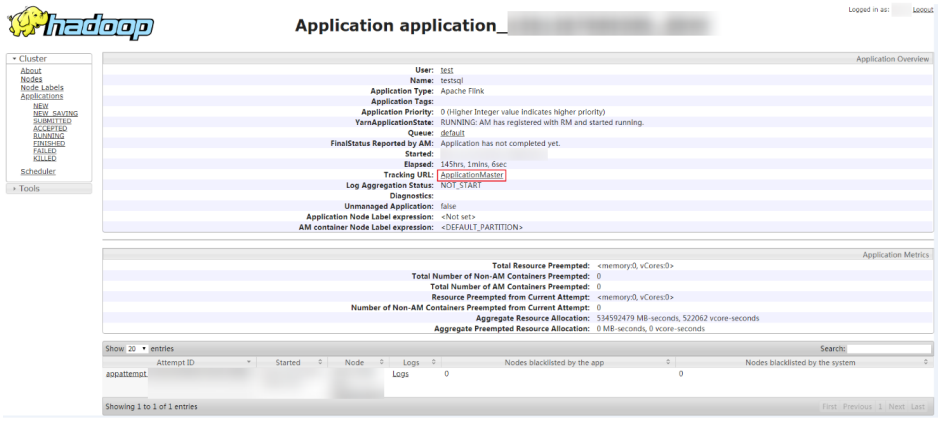
图 12-10 在 session 中提交作业成功



步骤7 使用运行用户进入Yarn服务的原生页面，具体操作参考“组件操作指南 > 使用Flink > 查看Flink作业”，找到对应作业的application，单击application名称，进入到作业详情页面

- 若作业尚未结束，可单击“Tracking URL”链接进入到Flink的原生页面，查看作业的运行信息。
- 若作业已运行结束，对于在session中提交的作业，可以单击“Tracking URL”链接登录Flink原生页面查看作业信息。

图 12-11 application



----结束

12.6.2 查看 Flink 作业信息

用户可以通过Yarn的WebUI，在图形化界面查看Flink作业的相关信息。

前提条件

集群已安装Flink服务。

访问 Yarn 的 WebUI

步骤1 进入Yarn服务页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Yarn > 概述”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > Yarn > 概览”。

步骤2 单击“ResourceManager WebUI”后面对应的链接，进入Yarn的WebUI页面。

----结束

12.6.3 配置管理 Flink

12.6.3.1 配置参数路径

Flink所有的配置参数都需要在客户端侧进行配置，配置文件路径：*客户端安装路径*/Flink/flink/conf/flink-conf.yaml。

说明

- 建议用户直接修改客户端的“flink-conf.yaml”配置文件进行配置，YAML文件的配置格式为*key: value*。

  例：**taskmanager.heap.size: 1024mb**

  注意配置项key:与value之间需有空格分隔。
- 如果在Flink服务的配置中修改了参数，配置完成之后需要重新下载安装客户端。

12.6.3.2 JobManager & TaskManager

配置场景

JobManager和TaskManager是Flink的主要组件，针对各种安全场景和性能场景，可以在客户端侧配置相关参数。

配置描述

主要配置项包括通信端口，内存管理，连接重试等。

针对MRS 3.x之前版本，参数说明见表**表12-65**。

表 12-65 参数说明

参数	是否必选	默认值	描述
taskmanager.rpc.port	否	默认值为32326-32390。	TaskManager的IPC端口范围。
taskmanager.data.port	否	默认值为32391-32455。	TaskManager数据交换端口范围。
taskmanager.data.ssl.enabled	否	false	TaskManager之间数据传输是否使用SSL加密，仅在全局开关security.ssl开启时有效。
taskmanager.numberOfTaskSlots	否	3	TaskManager占用的slot数，一般配置成物理机的核数，yarn-session模式下只能使用-s参数传递，yarn-cluster模式下只能使用-ys参数传递。

参数	是否必选	默认值	描述
parallelism.default	否	1	Job各个算子运行的并发数。
taskmanager.memory.size	否	0	TaskManager在JVM堆内存中保留空间的大小，此内存用于排序，哈希表和中间状态的缓存。如果未指定，则会使用JVM堆内存乘以比例taskmanager.memory.fraction。单位：MB。
taskmanager.memory.fraction	否	0.7	TaskManager在JVM堆内存中保留空间的比例，此内存用于排序，哈希表和中间状态的缓存。
taskmanager.memory.off-heap	是	false	TaskManager是否使用堆外内存，此内存用于排序，哈希表和中间状态的缓存。建议对于大内存，开启此配置提高内存操作的效率。
taskmanager.memory.segment-size	否	32768	TaskManager中memory segment大小，这是保留内存空间的基本单位，以及用于配置网络缓存栈。单位：bytes。
taskmanager.memory.preallocate	否	false	TaskManager是否在启动时分配保留内存空间。当开启堆外内存时，建议开启此配置项。
taskmanager.registration.initial-backoff	否	500 ms	两次连续注册的初始间隔时间。单位：ms/s/m/h/d。 **说明** 时间数值和单位之间有半角字符空格。ms/s/m/h/d表示毫秒、秒、分钟、小时、天。
taskmanager.registration.refused-backoff	否	5 min	JobManager拒绝注册后到允许再次注册的间隔时间。
task.cancellation.interval	否	30000	两次连续任务取消操作的间隔时间。

针对MRS 3.x及之后版本，参数说明见**表12-66**。

表 12-66 参数说明

参数	描述	默认值	是否必选配置
taskmanager.rpc.port	TaskManager的IPC端口范围。	默认值为32326-32390。	否
client.rpc.port	Flink client端Akka system监听端口。	默认值为32651-32720。	否
taskmanager.data.port	TaskManager数据交换端口范围。	默认值为32391-32455。	否
taskmanager.data.ssl.enabled	TaskManager之间数据传输是否使用SSL加密，仅在全局开关security.ssl开启时有效。	false	否
jobmanager.heap.size	JobManager堆内存大小，yarn-session模式下只能使用-jm参数传递，yarn-cluster模式下只能使用-yjm参数传递，如果小于YARN配置文件中yarn.scheduler.minimum-allocation-mb大小，则使用YARN配置中的值。单位：B/KB/MB/GB/TB。	1024mb	否
taskmanager.heap.size	TaskManager堆内存大小，yarn-session模式下只能使用-tm参数传递，yarn-cluster模式下只能使用-ytm参数传递，如果小于YARN配置文件中yarn.scheduler.minimum-allocation-mb大小，则使用YARN配置中的值。单位：B/KB/MB/GB/TB。	1024mb	否
taskmanager.numberOfTaskSlots	TaskManager占用的slot数，一般配置成物理机的核数，yarn-session模式下只能使用-s参数传递，yarn-cluster模式下只能使用-ys参数传递。	1	否
parallelism.default	默认并行度，用于未指定并行度的作业。	1	否
taskmanager.network.numberOfBuffers	TaskManager网络传输缓冲栈数量，如果作业运行中出错提示系统中可用缓冲不足，可以增加这个配置项的值。	2048	否
taskmanager.memory.fraction	TaskManager在JVM堆内存中保留空间的比例，此内存用于排序，哈希表和中间状态的缓存。	0.7	否
taskmanager.memory.off-heap	TaskManager是否使用堆外内存，此内存用于排序，哈希表和中间状态的缓存。建议对于大内存，开启此配置提高内存操作的效率。	false	是

参数	描述	默认值	是否必选配置
taskmanager.memory.segment-size	内存管理器和网络堆栈使用的内存缓冲区大小。单位：bytes。	32768	否
taskmanager.memory.preallocate	TaskManager是否在启动时分配保留内存空间。当开启堆外内存时，建议开启此配置项。	false	否
taskmanager.debug.memory.startLogThread	调试Flink内存和GC相关问题时可开启，TaskManager会定时采集内存和GC的统计信息，包括当前堆内，堆外，内存池的使用率和GC时间。	false	否
taskmanager.debug.memory.logIntervalMs	TaskManager定时采集内存和GC的统计信息的采集间隔。	0	否
taskmanager.maxRegistrationDuration	TaskManager向JobManager注册自己的最长时间，如果超过时间，TaskManager会关闭。	5 min	否
taskmanager.initial-registration-pause	两次连续注册的初始间隔时间。该值需带一个时间单位（ms/s/min/h/d）（比如5秒）。	500ms **说明** 时间数值和单位之间有半角字符空格。ms/s/m/h/d表示毫秒、秒、分钟、小时、天。	否
taskmanager.max-registration-pause	TaskManager注册失败最大重试间隔。单位：ms/s/m/h/d。	30s	否
taskmanager.refused-registration-pause	TaskManager注册连接被JobManager拒绝后的重试间隔。单位：ms/s/m/h/d。	10s	否
task.cancellation.interval	两次连续任务取消操作的间隔时间。单位：ms。	30000	否

参数	描述	默认值	是否必选配置
classloader.resolve-order	从用户代码加载类时定义类解析策略，这意味着是首先检查用户代码jar（“child-first”）还是应用程序类路径（“parent-first”）。默认设置指示首先从用户代码jar加载类，这意味着用户代码jar可以包含和加载不同于Flink使用的（依赖）依赖项。	child-first	否
slot.idle.timeout	Slot Pool中空闲Slot的超时时间（以ms为单位）。	50000	否
slot.request.timeout	从Slot Pool请求Slot的超时（以ms为单位）。	300000	否
task.cancellation.timeout	取消任务超时时间（以ms为单位），超时后会触发TaskManager致命错误。设置为0，取消任务卡住则不会报错。	180000	否
taskmanager.network.detailed-metrics	启用网络队列长度的详细指标监控。	false	否
taskmanager.network.memory.buffers-per-channel	每个传出/传入通道（子分区/输入通道）使用的最大网络缓冲区数.在基于信用的流量控制模式下，这表示每个输入通道中有多少信用。它应配置至少2以获得良好的性能。1个缓冲区用于接收子分区中的飞行中数据，1个缓冲区用于并行序列化。	2	否
taskmanager.network.memory.floating-buffers-per-gate	每个输出/输入门（结果分区/输入门）使用的额外网络缓冲区数。在基于信用的流量控制模式中，这表示在所有输入通道之间共享多少浮动信用。浮动缓冲区基于积压（子分区中的实时输出缓冲区）反馈来分布，并且可以帮助减轻由子分区之间的不平衡数据分布引起的背压。如果节点之间的往返时间较长和/或群集中的机器数量较多，则应增加此值。	8	否
taskmanager.network.memory.fraction	用于网络缓冲区的JVM内存的占比。这决定了TaskManager可以同时拥有多少流数据交换通道以及通道缓冲的程度。如果作业被拒绝或者收到系统没有足够缓冲区的警告，请增加此值或“taskmanager.network.memory.min”和“taskmanager.network.memory.max”。另请注意，“taskmanager.network.memory.min”和“taskmanager.network.memory.max”可能会覆盖此占比。	0.1	否

参数	描述	默认值	是否必选配置
taskmanager.network.memory.max	网络缓冲区的最大内存大小。该值需带一个大小单位（B/KB/MB/GB/TB）。	1 GB	否
taskmanager.network.memory.min	网络缓冲区的最小内存大小。该值需带一个大小单位（B/KB/MB/GB/TB）。	64 MB	否
taskmanager.network.request-backoff.initial	输入通道的分区请求的最小退避。	100	否
taskmanager.network.request-backoff.max	输入通道的分区请求的最大退避。	10000	否
taskmanager.registration.timeout	TaskManager注册的超时时间，在该时间内未成功注册，TaskManager将终止。该值需带一个时间单位（ms/s/min/h/d）。	5 min	否
resourcemanager.taskmanager-timeout	释放空闲TaskManager的超时（以ms为单位）。	30000	否

12.6.3.3 Blob

配置场景

JobManager节点上的Blob服务端是用于接收用户在客户端上传的Jar包，或将Jar包发送给TaskManager，传输log文件等。Flink提供配置Blob服务端的一些配置项，用户请在“flink-conf.yaml”配置文件中进行配置。

配置描述

用户可以配置端口，SSL，重试次数，并发等配置项。

表 12-67 参数说明

参数	描述	默认值	是否必选配置
blob.server.port	blob服务器端口。	默认值为32456-32520。	否
blob.service.ssl.enabled	blob传输通道是否加密传输，仅在全局开关security.ssl开启时有。	true	是

参数	描述	默认值	是否必选配置
blob.fetch.retries	TaskManager从JobManager下载blob文件的重试次数。	50	否
blob.fetch.num-concurrent	JobManager支持的下载blob的并发数。	50	否
blob.fetch.backlog	JobManager支持的blob下载队列大小，比如下载Jar包等。单位：个。	1000	否
library-cache-manager.cleanup.interval	当用户取消flink job后，jobmanager删除HDFS上存放用户jar包的时间，单位为s。	3600	否

说明

针对MRS 3.x之前版本，不支持配置library-cache-manager.cleanup.interval参数项。

12.6.3.4 Distributed Coordination (via Akka)

配置场景

Flink客户端与JobManager的通信，JobManager与TaskManager的通信和TaskManager与TaskManager的通信都基于Akka actor模型。Flink提供Akka连接参数的配置项，配置项请在“flink-conf.yaml”配置文件中进行配置，用户可以根据网络环境或调优策略再进行配置。

配置描述

配置项包括消息发送和等待的超时设置，akka监听机制Deathwatch的相关配置等。

针对MRS 3.x之前版本，参数说明见表12-68。

表 12-68 参数说明

参数	是否必选	默认值	描述
akka.ask.timeout	否	10 s	akka所有异步请求和阻塞请求的超时时间。如果Flink发生超时失败，可以增大这个值。当机器处理速度慢或者网络阻塞时会发生超时。单位：ms/s/m/h/d。
akka.lookup.timeout	否	10 s	查找JobManager actor对象的超时时间。单位：ms/s/m/h/d。
akka.framesize	否	10485760b	JobManager和TaskManager间最大消息传输大小。当Flink出现消息大小超过限制的错误时，可以增大这个值。单位：b/B/KB/MB。

参数	是否必选	默认值	描述
akka.watch.heartbeat.interval	否	10 s	Akka DeathWatch机制检测失联TaskManager的心跳间隔。如果TaskManager经常发生由于心跳消息丢失或延误而被错误标记为失联的情况，可以增大这个值。单位：ms/s/m/h/d。 **说明** DeathWatch的详细解释可以参考akka官网：**http://doc.akka.io/docs/akka/snapshot/scala/remoting.html#failure-detector**。
akka.watch.heartbeat.pause	否	60 s	Akka DeathWatch可接受的心跳暂停时间，较小的数值表示不允许不规律的心跳。单位：ms/s/m/h/d。 **说明** DeathWatch的详细解释可以参考akka官网：**http://doc.akka.io/docs/akka/snapshot/scala/remoting.html#failure-detector**。
akka.watch.threshold	否	12	DeathWatch失败检测阈值，较小的数值容易把正常TaskManager标记为失败，较大的值增加了失败检测的时间。 **说明** DeathWatch的详细解释可以参考akka官网：**http://doc.akka.io/docs/akka/snapshot/scala/remoting.html#failure-detector**。
akka.tcp.timeout	否	20 s	发送连接TCP超时时间，如果经常发生满网络环境下连接TaskManager超时，可以增大这个值。单位：ms/s/m/h/d。
akka.throughput	否	15	Akka批量处理消息的数量，一次操作完后把处理线程归还线程池。较小的数值代表actor消息处理的公平调度，较大的值以牺牲调度公平的代价提高整体性能。
akka.log.lifecycle.events	否	false	Akka远程时间日志开关，当需要调试时可打开此开关。
akka.startup-timeout	否	默认与akka.ask.timeout的值一致	Akka启动remote组件的超时时间。单位：ms/s/m/h/d。
akka.ssl.enabled	是	true	Akka通信SSL开关，仅在全局开关security.ssl开启时有。

针对MRS 3.x及之后版本，参数说明见**表12-69**。

表 12-69 参数说明

参数	描述	默认值	是否必选配置
akka.ask.timeout	akka所有异步请求和阻塞请求的超时时间。如果Flink发生超时失败，可以增大这个值。当机器处理速度慢或者网络阻塞时会发生超时。单位：ms/s/m/h/d。	10s	否
akka.lookup.timeout	查找JobManager actor对象的超时时间。单位：ms/s/m/h/d。	10s	否
akka.framesize	JobManager和TaskManager间最大消息传输大小。当Flink出现消息大小超过限制的错误时，可以增大这个值。单位：b/B/KB/MB。	10485760b	否
akka.watch.heartbeat.interval	Akka DeathWatch机制检测失联TaskManager的心跳间隔。如果TaskManager经常发生由于心跳消息丢失或延误而被错误标记为失联的情况，可以增大这个值。单位：ms/s/m/h/d。 **说明** DeathWatch的详细解释可以参考akka官网：**http://doc.akka.io/docs/akka/snapshot/scala/remoting.html#failure-detector**。	10s	否
akka.watch.heartbeat.pause	Akka DeathWatch可接受的心跳暂停时间，较小的数值表示不允许不规律的心跳。单位：ms/s/m/h/d。 **说明** DeathWatch的详细解释可以参考akka官网：**http://doc.akka.io/docs/akka/snapshot/scala/remoting.html#failure-detector**。	60s	否
akka.watch.threshold	DeathWath失败检测阈值，较小的数值容易把正常TaskManager标记为失败，较大的值增加了失败检测的时间。 **说明** DeathWatch的详细解释可以参考akka官网：**http://doc.akka.io/docs/akka/snapshot/scala/remoting.html#failure-detector**。	12	否
akka.tcp.timeout	发送连接TCP超时时间，如果经常发生满网络环境下连接TaskManager超时，可以增大这个值。单位：ms/s/m/h/d。	20s	否
akka.throughput	Akka批量处理消息的数量，一次操作完后把处理线程归还线程池。较小的数值代表actor消息处理的公平调度，较大的值以牺牲调度公平的代价提高整体性能。	15	否

参数	描述	默认值	是否必选配置
akka.log.lifecycle.events	Akka远程时间日志开关，当需要调试时可打开此开关。	false	否
akka.startup-timeout	远程组件启动失败前的超时时间。该值需带一个时间单位（ms/s/min/h/d）	默认与akka.ask.timeout的值一致	否
akka.ssl.enabled	Akka通信SSL开关，仅在全局开关security.ssl开启时有。	true	是
akka.client-socket-worker-pool.pool-size-factor	计算线程池大小的因子，计算公式：ceil（可用处理器*因子），计算结果限制在pool-size-min和pool-size-max之间。	1.0	否
akka.client-socket-worker-pool.pool-size-max	基于因子计算的线程数上限。	2	否
akka.client-socket-worker-pool.pool-size-min	基于因子计算的线程数下限。	1	否
akka.client.timeout	【说明】客户端超时时间。该值需带一个时间单位（ms/s/min/h/d）。	60s	否
akka.server-socket-worker-pool.pool-size-factor	【说明】计算线程池大小的因子，计算公式：ceil（可用处理器*因子），计算结果限制在pool-size-min和pool-size-max之间。	1.0	否
akka.server-socket-worker-pool.pool-size-max	基于因子计算的线程数上限。	2	否
akka.server-socket-worker-pool.pool-size-min	基于因子计算的线程数下限。	1	否

12.6.3.5 SSL

配置场景

当需要配置安全Flink集群时，需要配置SSL相关配置项。

配置描述

配置项包括SSL开关，证书，密码，加密算法等。

针对MRS 3.x之前版本，参数说明见表12-70。

表 12-70 参数说明

参数	是否必选	默认值	描述
security.ssl.internal.enabled	是	按照集群的安装模式自动配置。 • 安全模式：默认为true。 • 普通模式：默认为false。	内部通信SSL总开关。
security.ssl.internal.keystore	是	-	Java keystore文件。
security.ssl.internal.keystore-password	是	-	keystore文件解密密码。
security.ssl.internal.key-password	是	-	keystore文件中服务端key的解密密码。
security.ssl.internal.truststore	是	-	truststore文件包含公共CA证书。
security.ssl.internal.truststore-password	是	-	truststore文件解密密码。
security.ssl.protocol	是	TLSv1.2	SSL传输的协议版本。
security.ssl.algorithms	是	默认值为“TLS_RSA_WITH_AES_128_CBC_SHA256,TLS_DHE_RSA_WITH_AES_128_CBC_SHA256,TLS_DHE_DSS_WITH_AES_128_CBC_SHA256”	支持的SSL标准算法，具体可参考java官网：http://docs.oracle.com/javase/8/docs/technotes/guides/security/StandardNames.html#ciphersuites。

参数	是否必选	默认值	描述
security.ssl.rest.enabled	是	按照集群的安装模式自动配置。 • 安全模式：默认为true。 • 普通模式：默认为false。	外部通信SSL总开关。
security.ssl.rest.keystore	是	-	Java keystore文件。
security.ssl.rest.keystore-password	是	-	keystore文件解密密码。
security.ssl.rest.key-password	是	-	keystore文件中服务端key的解密密码。
security.ssl.rest.truststore	是	-	truststore文件包含公共CA证书。
security.ssl.rest.truststore-password	是	-	truststore文件解密密码。

针对MRS 3.x及之后版本，参数说明见**表12-71**。

表 12-71 参数说明

参数	描述	默认值	是否必选配置
security.ssl.enabled	内部通信SSL总开关。	按照集群的安装模式自动配置。 • 安全模式：默认为true。 • 非安全模式：默认为false。	是
security.ssl.keystore	Java keystore文件。	-	是
security.ssl.keystore-password	keystore文件解密密码。	-	是
security.ssl.key-password	keystore文件中服务端key的解密密码。	-	是
security.ssl.truststore	truststore文件包含公共CA证书。	-	是

参数	描述	默认值	是否必选配置
security.ssl.truststore-password	truststore文件解密密码。	-	是
security.ssl.protocol	SSL传输的协议版本。	TLSv1.2	是
security.ssl.algorithms	支持的SSL标准算法，具体可参考java官网：**http://docs.oracle.com/javase/8/docs/technotes/guides/security/StandardNames.html#ciphersuites**。	默认值为"TLS_DHE_RSA_WITH_AES_128_GCM_SHA256,TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256,TLS_DHE_RSA_WITH_AES_256_GCM_SHA384,TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384"	是

12.6.3.6 Network communication (via Netty)

配置场景

Flink运行Job时，Task之间的数据传输和反压检测都依赖Netty，某些环境下可能需要对Netty参数进行配置。

配置描述

对于高级调优，可调整以下Netty配置项，默认配置已可满足大规模集群并发高吞吐量的任务，参数详情可参考Netty官网：**http://netty.io/**。

表 12-72 参数说明

参数	描述	默认值	是否必选配置
taskmanager.network.netty.num-arenas	Netty内存块数。	1	否
taskmanager.network.netty.server.numThreads	Netty服务器线程的数量。	1	否
taskmanager.network.netty.client.numThreads	Netty客户端线程数。	1	否
taskmanager.network.netty.client.connectTimeoutSec	Netty客户端连接超时。单位：s。	120	否

参数	描述	默认值	是否必选配置
taskmanager.network.netty.sendReceiveBufferSize	Netty发送和接收缓冲区大小。 默认为系统缓冲区大小（cat / proc / sys / net / ipv4 / tcp_ [rw] mem ），在现代Linux中为4MB。单位：bytes。	4096	否
taskmanager.network.netty.transport	Netty传输类型，“nio”或“epoll”。	nio	否

12.6.3.7 JobManager Web Frontend

配置场景

JobManager启动时，会在同一进程内启动web服务器。

- 用户可以访问web服务器获取当前Flink集群的信息，包括JobManager，TaskManager及集群内运行的Job。
- 用户可以对web服务器参数进行配置。

配置描述

配置包括端口，临时目录，显示项目，错误重定向，安全相关等。

针对MRS 3.x之前版本，参数说明见**表12-73**。

表 12-73 参数说明

参数	是否必选	默认值	描述
jobmanager.web.port	否	32261-32325	web端口，支持范围：32261-32325。
jobmanager.web.allow-access-address	是	*	web访问白名单，ip以逗号隔开。只有在白名单中的ip才能访问web。

针对MRS 3.x及之后版本，参数说明见**表12-74**。

表 12-74 参数说明

参数	描述	默认值	是否必选配置
flink.security.enable	用户安装Flink集群时，需要选择“安全模式”或“普通模式”。 ● 当选择“安全模式”，配置项“flink.security.enable”被自动配置为“true”。 ● 当选择“普通模式”，配置项“flink.security.enable”被自动配置为“false”。 对于已经安装好的Flink集群，用户可以通过查看配置项“flink.security.enable”的值来区分当前安装的是安全模式还是普通模式。	按照集群的安装模式自动配置	否
rest.bind-port	web端口，支持范围：32261-32325。	32261-32325	否
jobmanager.web.history	显示“flink.security.enable”最近的job数目。	5	否
jobmanager.web.checkpoints.disable	禁用checkpoint统计。	false	否
jobmanager.web.checkpoints.history	Checkpoint统计记录数。	10	否
jobmanager.web.backpressure.cleanup-interval	未访问反压记录清理周期。单位：ms。	600000	否
jobmanager.web.backpressure.refresh-interval	反压记录刷新周期。单位：ms。	60000	否
jobmanager.web.backpressure.num-samples	计算反压使用的堆栈跟踪记录数。	100	否
jobmanager.web.backpressure.delay-between-samples	计算反压的采样间隔。单位：ms	50	否
jobmanager.web.ssl.enabled	web是否使用SSL加密传输，仅在全局开关security.ssl开启时有。	false	是
jobmanager.web.accesslog.enable	web操作日志使能开关，日志会存放在webaccess.log中。	true	是

参数	描述	默认值	是否必选配置
jobmanager.web.x-frame-options	http安全头X-Frame-Options的值，可选范围为：SAMEORIGIN、DENY、ALLOW-FROM uri。	DENY	是
jobmanager.web.cache-directive	web页面是否支持缓存。	no-store	是
jobmanager.web.expires-time	web页面缓存过期时长。单位：ms。	0	是
jobmanager.web.allow-access-address	web访问白名单，ip以逗号隔开。只有在白名单中的ip才能访问web。	*	是
jobmanager.web.access-control-allow-origin	网页同源策略，防止跨域攻击。	*	是
jobmanager.web.refresh-interval	web网页刷新时间。单位：ms。	3000	是
jobmanager.web.logout-timer	配置无操作情况下自动登出时间间隔。单位：ms。	600000	是
jobmanager.web.403-redirect-url	web403页面，访问若遇到403错误，则会重定向到配置的页面。	自动配置	是
jobmanager.web.404-redirect-url	web404页面，访问若遇到404错误，则会重定向到配置的页面。	自动配置	是
jobmanager.web.415-redirect-url	web415页面，访问若遇到415错误，则会重定向到配置的页面。	自动配置	是
jobmanager.web.500-redirect-url	web500页面，访问若遇到500错误，则会重定向到配置的页面。	自动配置	是
rest.await-leader-timeout	客户端等待Leader地址的时间（以ms为单位）。	30000	否
rest.client.max-content-length	客户端处理的最大内容长度（以字节为单位）。	104857600	否
rest.connection-timeout	客户端建立TCP连接的最长时间（以ms为单位）。	15000	否
rest.idleness-timeout	连接保持空闲状态的最长时间（以ms为单位）。	300000	否
rest.retry.delay	客户端在连续重试之间等待的时间（以ms为单位）。	3000	否
rest.retry.max-attempts	如果可重试算子操作失败，客户端将尝试重试的次数。	20	否

参数	描述	默认值	是否必选配置
rest.server.max-content-length	服务端处理的最大内容长度（以字节为单位）。	104857600	否
rest.server.numThreads	异步处理请求的最大线程数。	4	否
web.timeout	web监控超时时间（以ms为单位）。	10000	否

12.6.3.8 File Systems

配置场景

task运行中会创建结果文件，Flink支持对文件创建行为进行配置。

配置描述

配置项包括文件覆盖策略，目录创建。

表 12-75 参数说明

参数	描述	默认值	是否必选配置
fs.overwrite-files	文件输出写操作是否默认覆盖已有文件。	false	否
fs.output.always-create-directory	当文件写入程序的并行度大于1时，输出文件的路径下会创建一个目录，并将不同的结果文件（每个并行写程序任务一个）放入该目录。 • 如果此选项设置为true，那么并行度为1的写入程序也将创建一个目录并将一个结果文件放入其中。 • 如果该选项设置为false，则并行度为1的写入程序将直接在输出路径中创建文件，而不再创建目录。	false	否

12.6.3.9 State Backend

配置场景

Flink提供了HA和作业的异常恢复，并且提供版本升级时作业的暂停恢复。对于作业状态的存储，Flink依赖于state backend，作业的重启依赖于重启策略，用户可以对这两部分进行配置。

配置描述

配置项包括state backend类型，存储路径，重启策略等。

表 12-76 参数说明

参数	描述	默认值	是否必选配置
state.backend.fs.checkpointdir	当backend为filesystem时的路径，路径必须能够被JobManager访问到，本地路径只支持local模式，集群模式下请使用HDFS路径。	hdfs:///flink/checkpoints	否
state.savepoints.dir	Flink用于恢复和更新作业的保存点存储目录。当触发保存点的时候，保存点元数据信息将会保存到该目录中。	hdfs:///flink/savepoint	安全模式下必配
restart-strategy	默认重启策略，用于未指定重启策略的作业。三个值可选： • fixed-delay • failure-rate • none	none	否
restart-strategy.fixed-delay.attempts	fixed-delay策略重试次数，具体策略的介绍请参见：**https://ci.apache.org/projects/flink/flink-docs-release-1.12/dev/task_failure_recovery.html**。	• 作业中开启了checkpoint，则默认值为Integer.MAX_VALUE。 • 作业中未开启checkpoint，默认值为3。	否
restart-strategy.fixed-delay.delay	fixed-delay策略重试间隔时间。单位：ms/s/m/h/d。	• 作业中开启了checkpoint，默认值是10 s。 • 作业中不开启checkpoint，默认值和配置项akka.ask.timeout的值一致。	否

参数	描述	默认值	是否必选配置
restart-strategy.failure-rate.max-failures-per-interval	故障率策略下作业失败前给定时间段内的最大重启次数。具体策略的介绍请参见： **https://ci.apache.org/projects/flink/flink-docs-release-1.12/dev/task_failure_recovery.html**。	1	否
restart-strategy.failure-rate.failure-rate-interval	failure-rate策略重试时间。单位：ms/s/m/h/d。	60 s	否
restart-strategy.failure-rate.delay	failure-rate策略重试间隔时间。单位：ms/s/m/h/d。	默认值和akka.ask.timeout配置值一样，请参见**Distributed Coordination (via Akka)**	否

12.6.3.10 Kerberos-based Security

配置场景

Flink安全模式下必须配置Kerberos相关配置项。

配置描述

配置项包括kerberos的keytab、principal、cookie等。

说明

针对MRS 3.x之前版本，配置项不包括cookie。

表 12-77 参数说明

参数	描述	默认值	是否必选配置
security.kerberos.login.keytab	该参数为客户端参数，keytab路径。	根据实际业务配置	是
security.kerberos.login.principal	该参数为客户端参数，如果keytab和principal都设置，默认会使用keytab认证。	根据实际业务配置	否

参数	描述	默认值	是否必选配置
security.kerberos.login.contexts	该参数为服务器端参数，flink生成jass文件的contexts。	Client、KafkaClient	是
security.enable	该参数为客户端参数，flink内部模块认证使能开关。	按照集群的安装模式自动配置： ● 安全模式：true ● 非安全模式：false	是
security.cookie	该参数为客户端参数，模块认证token，在security.enable打开时必须配置，不能是空串。	根据实际业务配置	是

说明

针对MRS 3.x之前版本，参数说明不包括security.enable、security.cookie。

12.6.3.11 HA

配置场景

Flink的HA模式依赖于ZooKeeper，所以必须配置ZooKeeper相关配置。

配置描述

配置项包括ZooKeeper地址，路径，安全认证等。

表 12-78 参数说明

参数	描述	默认值	是否必选配置
high-availability	HA模式，是启用HA还是非HA模式。当前支持两种模式： 1. none，只运行单个jobManager，jobManager的状态不进行Checkpoint。 2. ZooKeeper。 ● 非YARN模式下，支持多个jobManager，通过选举产生leader。 ● YARN模式下只存在一个jobManager。	zookeeper	否

参数	描述	默认值	是否必选配置
high-availability.zookeeper.quorum	ZooKeeper quorum地址。	自动配置	否
high-availability.zookeeper.path.root	Flink在ZooKeeper上创建的根目录，存放HA模式必须的元数据。	/flink	否
high-availability.storageDir	存放state backend中JobManager元数据，ZooKeeper只保存实际数据的指针。	hdfs:///flink/recovery	否
high-availability.zookeeper.client.session-timeout	ZooKeeper客户端会话超时时间。单位：ms。	60000	否
high-availability.zookeeper.client.connection-timeout	ZooKeeper客户端连接超时时间。单位：ms。	15000	否
high-availability.zookeeper.client.retry-wait	ZooKeeper客户端重试等待时间。单位：ms。	5000	否
high-availability.zookeeper.client.max-retry-attempts	ZooKeeper客户端最大重试次数。	3	否
high-availability.job.delay	当jobManager恢复后重启job的延迟时间。	默认值和akka.ask.timeout配置值保持一致。	否
high-availability.zookeeper.client.acl	设置ZooKeeper节点的ACL（open creator）。设置ACL选项请参考：**https://zookeeper.apache.org/doc/r3.5.1-alpha/zookeeperProgrammers.html#sc_BuiltinACLSchemes**。	按照集群的安装模式自动配置： • 安全模式：creator • 非安全模式：open	是

参数	描述	默认值	是否必选配置
zookeeper.sasl.disable	基于SASL认证的使能开关。	按照集群的安装模式自动配置： • 安全模式：false • 非安全模式：true	是
zookeeper.sasl.service-name	• 如果ZooKeeper服务端配置了不同于“ZooKeeper”的服务名，可以设置此配置项。 • 如果客户端和服务端的服务名不一致，认证会失败。	zookeeper	是

说明

针对MRS 3.x之前版本，不支持high-availability.job.delay配置参数。

12.6.3.12 Environment

配置场景

对于JVM配置有特定要求的场景，可以通过配置项传递JVM参数到客户端，JobMananger，TaskManager等。

配置描述

可配置JVM参数。

表 12-79 参数说明

参数	描述	默认值	是否必选配置
env.java.opts	JVM参数，会传递到启动脚本，JobManager，TaskManager，Yarn客户端。比如传递远程调试的参数等。	-Xloggc:<LOG_DIR>/gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=20M -Djdk.tls.ephemeralDHKeySize=2048 -Djava.library.path=${HADOOP_COMMON_HOME}/lib/native -Djava.net.preferIPv4Stack=true -Djava.net.preferIPv6Addresses=false -Dbeetle.application.home.path=/opt/*xxx*/Bigdata/common/runtime/security/config	否

12.6.3.13 Yarn

配置场景

Flink运行在Yarn集群上时，JobManager运行在Application Master上。JobManager的一些配置参数依赖于Yarn，通过配置YARN相关的配置，使Flink更好的运行在Yarn上。

配置描述

配置项包括yarn container的内存，虚拟内核，端口等。

表 12-80 参数说明

参数	描述	默认值	是否必选配置
yarn.maximum-failed-containers	当TaskManager所属容器出错后，重新申请container次数。默认值为Flink集群启动时TaskManager的数量。	5	否

参数	描述	默认值	是否必选配置
yarn.application-attempts	Application master重启次数，次数是算在一个validity interval的最大次数，validity interval在flink中设置为akka的timeout。重启后AM的地址和端口会变化，client需要手动连接。	2	否
yarn.heartbeat-delay	Application Master和YARN Resource Manager心跳的时间间隔。单位：seconds	5	否
yarn.containers.vcores	每个Yarn容器的虚拟核数。	默认值是TaskManager的slot数	否
yarn.application-master.port	Application Master端口号设置，支持端口范围。	32586-32650	否

12.6.3.14 Pipeline

配置场景

为适应某些场景对降低时延的需求，设计多个Job间采用Netty直接相连的方式传递数据，即分别使用NettySink用于Server端、NettySource用于Client端进行数据传输。

本章节适用于MRS 3.x及之后版本。

配置描述

配置项包括NettySink的信息存放路径、NettySink的端口监听范围、连接是否通过SSL加密以及NettySink监听所使用的网络所在域等。

表 12-81 参数说明

参数	描述	默认值	是否必选配置
nettyconnector.registerserver.topic.storage	设置NettySink的IP、端口及并发度信息在第三方注册服务器上的路径。建议用户使用ZooKeeper进行存储。	/flink/nettyconnector	否，当使用pipeline特性为必选
nettyconnector.sinkserver.port.range	设置NettySink的端口范围。	MRS集群下默认设置为28444-28843	否，当使用pipeline特性为必选

参数	描述	默认值	是否必选配置
nettyconnector.ssl.enabled	设置NettySink与NettySource之间通信是否配置SSL加密。其中加密密钥以及加密协议等请参见SSL。	false	否，当使用pipeline特性为必选
nettyconnector.message.delimiter	用来配置nettysink发送给nettysource消息的分隔符，长度为2-4个字节，不可包含“\n”，“ ”，“#”。	默认使用“$_”。	否，当使用pipeline特性为必选

12.6.4 安全配置

12.6.4.1 安全特性描述

Flink 主要完成如下安全特性：

- Flink集群中，各部件支持认证。
  - Flink集群内部各部件和外部部件之间，支持和外部部件如YARN、HDFS、ZooKeeprer进行kerberors认证。
  - Flink集群内部各部件之间，如Flink client和JobManager、JobManager和TaskManager、TaskManager和TaskManager之间支持security cookie认证。
- Flink集群中，各部件支持SSL加密传输。
- Flink集群内部各部件之间，如Flink client和JobManager、JobManager和TaskManager、TaskManager和TaskManager之间支持SSL加密传输。
- Flink web安全加固。
  - 支持白名单过滤，Flink web只能通过YARN代理访问。
  - 安全头域增强。
- Flink集群中，各部件的监听端口支持范围可配置。
- 在HA模式下，支持ACL控制。

12.6.4.2 配置对接 Kafka

Flink样例工程的数据存储在Kafka组件中。向Kafka组件发送数据（需要有Kafka权限用户），并从Kafka组件接收数据。

步骤1 确保集群安装完成，包括HDFS、Yarn、Flink和Kafka。

步骤2 创建Topic。

- 用户使用Linux命令行创建topic，执行命令前需要使用kinit命令进行人机认证，如**kinit flinkuser**。

  说明

  flinkuser需要用户自己创建，并拥有创建Kafka的topic权限。

创建topic的命令格式：{zkQuorum}表示ZooKeeper集群信息，格式为IP:port。{Topic}表示Topic名称。

bin/kafka-topics.sh --create --zookeeper {zkQuorum}/kafka --replication-factor 1 --partitions 5 --topic {Topic}

例如此处以topic1的数据为例：

```
/opt/client/Kafka/kafka/bin/kafka-topics.sh --create --zookeeper 10.96.101.32:2181,10.96.101.251:2181,10.96.101.177:2181,10.91.8.160:2181/kafka --replication-factor 1 --partitions 5 --topic topic1
```

- 服务端topic权限配置。

  将Kafka的Broker配置参数“allow.everyone.if.no.acl.found”的值修改为“true”。

步骤3 安全认证。

安全认证的方式有三种：Kerberos认证、SSL加密认证和Kerberos+SSL模式认证，用户在使用的时候可任选其中一种方式进行认证。

说明

针对MRS 3.x之前版本，安全认证的方式只支持Kerberos认证。

- **Kerberos认证配置**
  - 客户端配置。

    在Flink配置文件“flink-conf.yaml”中，增加kerberos认证相关配置（主要在“contexts”项中增加“KafkaClient”），示例如下：

    ```
    security.kerberos.login.keytab: /home/demo//keytab/flinkuser.keytab
    security.kerberos.login.principal: flinkuser
    security.kerberos.login.contexts: Client,KafkaClient
    security.kerberos.login.use-ticket-cache: false
    ```

    说明

    针对MRS 3.x之前版本，配置security.kerberos.login.keytab示例为：/home/demo/flink/release/keytab/flinkuser.keytab。

  - 运行参数。

    关于“SASL_PLAINTEXT”协议的运行参数示例如下：

    ```
    --topic topic1 --bootstrap.servers 10.96.101.32:21007 --security.protocol SASL_PLAINTEXT --sasl.kerberos.service.name kafka //10.96.101.32:21007表示kafka服务器的IP:port
    ```

- **SSL加密配置**
  - 服务端配置。

    登录FusionInsight Manager页面，选择“集群 > 服务 > Kafka > 配置”，参数类别设置为“全部配置”，搜索“ssl.mode.enable”并配置为“true”。

  - 客户端配置。

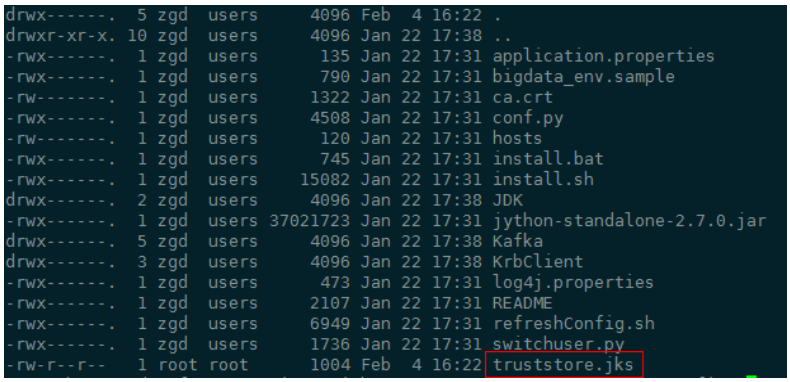
    i. 登录集群的FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > Kafka > 更多 > 下载客户端”，下载客户端压缩文件到本地机器。

    ii. 使用客户端根目录中的“ca.crt”证书文件生成客户端的“truststore”。

    执行命令如下：

    ```
    keytool -noprompt -import -alias myservercert -file ca.crt -keystore truststore.jks
    ```

    命令执行结果查看：



iii. 运行参数。

“ssl.truststore.password”参数内容需要跟创建“truststore”时输入的密码保持一致，执行以下命令运行参数。

```
--topic topic1 --bootstrap.servers 10.96.101.32:9093 --security.protocol SSL --
ssl.truststore.location /home/zgd/software/FusionInsight_Kafka_ClientConfig/truststore.jks
--ssl.truststore.password XXX
```

- **Kerberos+SSL模式配置**

  完成上文中Kerberos和SSL各自的服务端和客户端配置后，只需要修改运行参数中的端口号和协议类型即可启动Kerberos+SSL模式。

  ```
  --topic topic1 --bootstrap.servers 10.96.101.32:21009 --security.protocol SASL_SSL --
  sasl.kerberos.service.name kafka --ssl.truststore.location /home/zgd/software/
  FusionInsight_Kafka_ClientConfig/truststore.jks --ssl.truststore.password XXX
  ```

----结束

12.6.4.3 配置 Pipeline

本章节适用于MRS 3.x及之后版本。

1. 配置文件。
   - nettyconnector.registerserver.topic.storage：设置NettySink的IP、端口及并发度信息在第三方注册服务器上的路径（必填），例如：
     ```
     nettyconnector.registerserver.topic.storage: /flink/nettyconnector
     ```
   - nettyconnector.sinkserver.port.range：设置NettySink的端口范围（必填），例如：
     ```
     nettyconnector.sinkserver.port.range: 28444-28843
     ```
   - nettyconnector.ssl.enabled：设置NettySink与NettySource之间通信是否SSL加密（默认为false），例如：
     ```
     nettyconnector.ssl.enabled: true
     ```
2. 安全认证配置。
   - Zookeeper的SASL认证，依赖“flink-conf.yaml”中有关HA的相关配置。
   - SSL的keystore、truststore、keystore password、truststore password以及password等也使用“flink-conf.yaml”的相关配置，具体配置请参见**加密传输**。

12.6.5 安全加固

12.6.5.1 认证和加密

安全认证

Flink整个系统有三种认证方式：

- 使用kerberos认证：Flink yarn client与Yarn Resource Manager、JobManager与Zookeeper、JobManager与HDFS、TaskManager与HDFS、Kafka与TaskManager、TaskManager和Zookeeper。
- 使用security cookie进行认证：Flink yarn client与Job Manager、JobManager与TaskManager、TaskManager与TaskManager。
- 使用YARN内部的认证机制：Yarn Resource Manager与Application Master（简称AM）。

  说明

  - Flink的JobManager与YARN的AM是在同一个进程下。
  - 如果用户集群开启Kerberos认证需要使用kerberos认证。
  - 针对MRS 3.x之前版本，Flink不支持使用security cookie方式进行认证。

表 12-82 安全认证方式

安全认证方式	说明	配置方法
Kerberos认证	当前只支持keytab认证方式。	1. 从KDC服务器上下载用户keytab，并将keytab放到Flink客户端所在主机的某个文件夹下。 2. 在“flink-conf.yaml”上配置： a. keytab路径。 security.kerberos.login.keytab: /home/flinkuser/keytab/abc222.keytab 说明： “/home/flinkuser/keytab/abc222.keytab”表示的是用户目录。 b. principal名。 security.kerberos.login.principal: abc222 c. 对于HA模式，如果配置了ZooKeeper，还需要设置ZK kerberos认证相关的配置。配置如下： zookeeper.sasl.disable: false security.kerberos.login.contexts: Client d. 如果用户对于Kafka client和Kafka broker之间也需要做kerberos认证，配置如下： security.kerberos.login.contexts: Client,KafkaClient

安全认证方式	说明	配置方法
Security Cookie认证	-	1. 参考**签发证书样例**章节生成“generate_keystore.sh”脚本并放置在Flink客户端的“bin”目录下，调用“generate_keystore.sh”脚本，生成“Security Cookie”、“flink.keystore”文件和“flink.truststore”文件。 执行**sh generate_keystore.sh**，输入用户自定义密码。密码不允许包含#。 **说明** 执行脚本后，在Flink客户端的“conf”目录下生成“flink.keystore”和“flink.truststore”文件，并且在客户端配置文件“flink-conf.yaml”中将以下配置项进行了默认赋值。 ● 将配置项“security.ssl.keystore”设置为“flink.keystore”文件所在绝对路径。 ● 将配置项“security.ssl.truststore”设置为“flink.truststore”文件所在的绝对路径。 ● 将配置项“security.cookie”设置为“generate_keystore.sh”脚本自动生成的一串随机规则密码。 ● 默认“flink-conf.yaml”中“security.ssl.encrypt.enabled: false”，“generate_keystore.sh”脚本将配置项“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值设置为调用“generate_keystore.sh”脚本时输入的密码。 ● MRS 3.1.0及之后版本，如果需要使用密文时，设置“flink-conf.yaml”中“security.ssl.encrypt.enabled: true”，“generate_keystore.sh”脚本不会配置“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值，需要使用Manager明文加密API进行获取，执行**curl -k -i -u** *user name*:*password* **-X POST -HContent-type:application/json -d '{"plainText":"***password***"}' 'https://***x.x.x.x***:28443/web/api/v2/tools/encrypt'** 其中*user name*:password分别为当前系统登录用户名和密码；**"plainText"**的password为调用“generate_keystore.sh”脚本时的密码；x.x.x.x为集群Manager的浮动IP。 2. 打开“Security Cookie”开关，配置flink-conf.yaml文件中的“security.enable: true”，查看“security cookie”是否已配置成功，例如： `security.cookie: ae70acc9-9795-4c48-ad35-8b5adc8071744f605d1d-2726-432e-88ae-dd39bfec40a9`
YARN内部认证方式	该方式是YARN内部的认证方式，不需要用户配置。	-

说明

当前一个Flink集群只支持一个用户，一个用户可以创建多个Flink集群。

加密传输

Flink整个系统有三种加密传输方式：

- 使用Yarn内部的加密传输方式：Flink yarn client与Yarn Resource Manager、Yarn Resource Manager与Job Manager。
- SSL：Flink yarn client与JobManager、JobManager与TaskManager、TaskManager与TaskManager。
- 使用Hadoop内部的加密传输方式：JobManager和HDFS、TaskManager和HDFS、JobManager与ZooKeeper、TaskManager与ZooKeeper。

说明

Yarn内部和Hadoop内部都不需要用户配置加密，用户只需要配置SSL加密传输方式。

配置SSL传输，用户主要在客户端的“flink-conf.yaml”文件中做如下配置：

1. 打开SSL开关和设置SSL加密算法，针对MRS 3.x及之后版本，配置参数如表12-83所示，请根据实际情况修改对应参数值。

表 12-83 参数描述

参数	参数值示例	描述
security.ssl.enabled	true	打开SSL总开关。
akka.ssl.enabled	true	打开akka SSL开关。
blob.service.ssl.enabled	true	打开blob通道SSL开关。
taskmanager.data.ssl.enabled	true	打开taskmanager之间通信的SSL开关。
security.ssl.algorithms	TLS_DHE_RSA_WITH_AES_128_GCM_SHA256,TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256,TLS_DHE_RSA_WITH_AES_256_GCM_SHA384,TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384	设置SSL加密的算法。

针对MRS 3.x之前版本，配置参数如表12-84所示。

表 12-84 参数描述

参数	参数值示例	描述
security.ssl.internal.enabled	true	打开内部SSL开关。
akka.ssl.enabled	true	打开akka SSL开关。
blob.service.ssl.enabled	true	打开blob通道SSL开关。
taskmanager.data.ssl.enabled	true	打开taskmanager之间通信的SSL开关。
security.ssl.algorithms	TLS_RSA_WITH_AES128CBC_SHA256	设置SSL加密的算法。

针对MRS 3.x之前版本，如下参数见表12-85，在MRS的Flink默认配置中不存在，用户如果开启外部连接SSL，则需要添加以下参数。开启外部连接SSL后，因为YARN目前的开源版本无法代理HTTPS请求，所以无法通过YARN代理访问Flink的原生页面，用户可以在集群的同一个VPC下，创建windows虚拟机，在该虚拟机中访问Flink 原生页面。

表 12-85 参数描述

参数	参数值示例	描述
security.ssl.rest.enabled	true	打开外部SSL开关，若该参数配置为“true”，请参考表12-85配置相关参数。
security.ssl.rest.keystore	${path}/flink.keystore	keystore的存放路径。
security.ssl.rest.keystore-password	-	keystore的password，-表示需要用户输入自定义设置的密码值。
security.ssl.rest.key-password	-	ssl key的password，-表示需要用户输入自定义设置的密码值。
security.ssl.rest.truststore	${path}/flink.truststore	truststore存放路径。
security.ssl.rest.truststore-password	-	truststore的password，-表示需要用户输入自定义设置的密码值。

说明

如果打开Task Manager之间data传输通道的SSL，对性能会有较大影响，需要用户从安全和性能综合考虑。

2. 在Flink客户端的bin目录下，执行命令**sh generate_keystore.sh** *<password>*，请参考**认证和加密**，针对MRS 3.x及之后版本，**表12-86**中的配置项会被默认赋值，用户也可以手动配置。

表 12-86 参数描述

参数	参数值示例	描述
security.ssl.keystore	${path}/flink.keystore	keystore的存放路径，“flink.keystore”表示用户通过generate_keystore.sh*工具生成的keystore文件名称。
security.ssl.keystore-password	-	keystore的password，-表示需要用户输入自定义设置的密码值。
security.ssl.key-password	-	ssl key的password，-表示需要用户输入自定义设置的密码值。
security.ssl.truststore	${path}/flink.truststore	truststore存放路径，“flink.truststore”表示用户通过generate_keystore.sh*工具生成的truststore文件名称。
security.ssl.truststore-password	-	truststore的password，-表示需要用户输入自定义设置的密码值。

针对MRS 3.x之前版本，***generate_keystore.sh***不需手动生成，**表12-87**中的配置项会被默认赋值，用户也可以手动配置。

表 12-87 参数描述

参数	参数值示例	描述
security.ssl.internal.keystore	${path}/flink.keystore	keystore的存放路径，“flink.keystore”表示用户通过generate_keystore.sh*工具生成的keystore文件名称。
security.ssl.internal.keystore-password	-	keystore的password，表示需要用户输入自定义设置的密码值。
security.ssl.internal.key-password	-	ssl key的password，表示需要用户输入自定义设置的密码值。
security.ssl.internal.truststore	${path}/flink.truststore	truststore存放路径，“flink.truststore”表示用户通过generate_keystore.sh*工具生成的truststore文件名称。
security.ssl.internal.truststore-password	-	truststore的password，表示需要用户输入自定义设置的密码值。

针对MRS 3.x之前版本，如果开启外部连接SSL，即 security.ssl.rest.enabled 配置为 true，则如下参数见表12-88，用户需要配置。

表 12-88 参数说明

参数	参数值示例	描述
security.ssl.rest.enabled	true	打开外部SSL开关，若该参数配置为“true”，请参考表12-88配置相关参数。
security.ssl.rest.keystore	${path}/flink.keystore	keystore的存放路径
security.ssl.rest.keystore-password	-	keystore的password，表示需要用户输入自定义设置的密码值。
security.ssl.rest.key-password	-	ssl key的password，表示需要用户输入自定义设置的密码值。
security.ssl.rest.truststore	${path}/flink.truststore	truststore存放路径
security.ssl.rest.truststore-password	-	truststore的password，表示需要用户输入自定义设置的密码值。

说明

path”目录是用来存放SSL keystore、truststore相关配置文件，该目录是由用户自定义创建。相对路径和绝对路径的不同导致执行命令存在差异，详细说明在3和4中说明。

3. 配置keystore或truststore文件路径为相对路径时，Flink Client执行命令的目录需要可以直接访问该相对路径。Flink有两种执行方式来传输keystore和truststore文件。
   - 在Flink的CLI yarn-session.sh命令中增加“-t”选项来传输keystore和truststore文件到各个执行节点。如：
     ```
     ./bin/yarn-session.sh -t ssl/
     ```
   - 在Flink run命令中增加“-yt”选项来传输keystore和truststore文件到各个执行节点。如：
     ```
     ./bin/flink run -yt ssl/ -ys 3  -m yarn-cluster -c
     org.apache.flink.examples.java.wordcount.WordCount /opt/client/Flink/flink/examples/batch/
     WordCount.jar
     ```

     说明

     - 在举例当中的“ssl/”是Flink Client端目录下的子目录，该目录是用来存放SSL keystore、truststore相关配置文件。
     - Flink Client执行命令的当前路径需要能访问到“ssl/”相对路径。
4. 配置keystore或truststore文件路径为绝对路径时，需要在Flink Client以及各个节点的该绝对路径上放置keystore和truststore文件。

说明

针对MRS 3.x之前版本，提交作业的用户需要具有读取keystore和truststore文件的权限。

Flink有两种方式执行应用程序，且执行命令中不需要使用“-t”或“-yt”来传输keystore和truststore文件。

- 使用Flink的CLI yarn-session.sh命令执行应用程序。如：
  ```
  ./bin/yarn-session.sh
  ```
- 使用Flink run命令执行应用程序。如：
  ```
  ./bin/flink run  -ys 3 -m yarn-cluster -c
  org.apache.flink.examples.java.wordcount.WordCount /opt/client/Flink/flink/examples/batch/
  WordCount.jar
  ```

12.6.5.2 ACL 控制

Flink在HA模式下，支持用ZooKeeper来管理集群和发现服务。ZooKeeper支持SASL ACL控制，即只有通过SASL（kerberos）认证的用户，才有往ZK上操作文件的权限。如果要在Flink上使用SASL ACL控制，需要在Flink配置文件中设置如下配置：

```
high-availability.zookeeper.client.acl: creator
zookeeper.sasl.disable: false
```

具体配置项介绍请参考**表12-78**。

12.6.5.3 web 安全

编码规范

说明：Web Service客户端和服务器间使用相同的编码方式，是为了防止出现乱码现象，也是实施输入校验的基础。

安全加固：web server响应消息统一采用UTF-8字符编码。

支持 IP 白名单过滤

说明：防止非法用户登录，需在web server侧添加IP Filter过滤源IP非法的请求。

安全：支持IP Filter实现Web白名单配置，配置项是“jobmanager.web.allow-access-address”，默认情况下只支持YARN用户接入。

说明

安装客户端之后需要将客户端节点IP追加到jobmanager.web.allow-access-address配置项中。

禁止将文件绝对路径发送到客户端

说明：文件绝对路径发送到客户端会暴露服务端的目录结构信息，有助于攻击者遍历了解系统，为攻击者攻击提供帮助。

安全加固：Flink配置文件中所有配置项中如果包含以/开头的，则删掉第一级目录。

同源策略

同源策略适用于MRS 3.x及之后版本。

如果两个URL的协议，主机和端口均相同，则它们同源；如果不同源，默认不能相互访问；除非被访问者在其服务端显示指定访问者的来源。

安全加固：响应头“Access-Control-Allow-Origin”头域默认配置为YARN集群ResourceManager的IP地址，如果源不是来自YARN的，则不能互相访问。

防范 XSS

防范XSS适用于MRS 3.x及之后版本。

不启用XSS Filter会有以下几种威胁：

- 敏感信息泄露（如sessionid）。
- 页面被篡改。
- 重定向到恶意网站。
- 其他网站对Flink web发起DoS攻击。

安全加固：用户添加“X-XSS-Protection”安全头域，默认配置为“X-XSS-Protection: 1; mode=block”，则可以启用XSS保护，过滤XSS攻击；并在检查到XSS攻击时，停止渲染页面（例如IE8中，检查到攻击时，整个页面会被一个#替换）。

防范敏感信息泄露

防范敏感信息泄露适用于MRS 3.x及之后版本。

带有敏感数据的Web页面都应该禁止缓存，以防止敏感信息泄漏或通过代理服务器上网的用户数据互窜现象。

安全加固：添加“Cache-control”、“Pragma”、“Expires”安全头域，默认值为：“Cache-Control：no-store”，“Pragma ： no-cache”，“Expires ： 0”。

实现了安全加固，Flink和web server交互的内容将不会被缓存。

防止劫持

防止劫持适用于MRS 3.x及之后版本。

由于单击劫持（ClickJacking）和框架盗链都利用到框架技术，所以需要采用安全措施。

安全加固：添加“X-Frame-Options”安全头域，给浏览器提供允许一个页面可否在“iframe”、“frame”或“object”网站中的展现页面的指示，如果默认配置为“X-Frame-Options: DENY”，则确保任何页面都不能被嵌入到别的“iframe”、“frame”或“object”网站中，从而避免了单击劫持 (clickjacking) 的攻击。

对 Web Service 接口调用记录日志

对Web Service接口调用记录日志适用于MRS 3.x及之后版本。

对“Flink webmonitor restful”接口调用进行日志记录。

安全加固：“access log”支持配置：“jobmanager.web.accesslog.enable”，默认为“true”。且日志保存在单独的“webaccess.log”文件中。

跨站请求（CSRF）伪造防范

跨站请求（CSRF）伪造防范适用于MRS 3.x及之后版本。

在B/S应用中，对于涉及服务器端数据改动（如增加、修改、删除）的操作必须进行跨站请求伪造的防范。跨站请求伪造是一种挟制终端用户在当前已登录的Web应用程序上执行非本意的操作的攻击方法。

安全加固：现有请求修改的接口有2个post，1个delete，其余均是get请求，非get请求的接口均已删除。

异常处理

异常处理适用于MRS 3.x及之后版本。

应用程序出现异常时，捕获异常，过滤返回给客户端的信息，并在日志中记录详细的错误信息。

安全加固：

- 默认的错误提示页面，进行信息过滤，并在日志中记录详细的错误信息。
- 新加四个配置项，默认配置为FusionInsight提供的跳转URL，错误提示页面跳转到固定配置的URL中，防止暴露不必要的信息。

表 12-89 四个配置项参数介绍

参数	描述	默认值	是否必选配置
jobmanager.web.403-redirect-url	web403页面，访问若遇到403错误，则会重定向到配置的页面。	-	是
jobmanager.web.404-redirect-url	web404页面，访问若遇到404错误，则会重定向到配置的页面。	-	是
jobmanager.web.415-redirect-url	web415页面，访问若遇到415错误，则会重定向到配置的页面。	-	是
jobmanager.web.500-redirect-url	web500页面，访问若遇到500错误，则会重定向到配置的页面。	-	是

HTML5 安全

HTML5安全适用于MRS 3.x及之后版本。

HTML5是下一代的Web开发规范，为开发者提供了许多新的功能并扩展了标签。这些新的标签及功能增加了攻击面，存在被攻击的风险（例如跨域资源共享、客户端存储、WebWorker、WebRTC、WebSocket等）。

安全加固：添加“Access-Control-Allow-Origin”配置，如运用到跨域资源共享功能，可对HTTP响应头的“Access-Control-Allow-Origin”属性进行控制。

说明

Flink不涉及如客户端存储、WebWorker、WebRTC、WebSocket等安全风险。

12.6.6 安全声明

- Flink的安全都为开源社区提供和自身研发。有些是需要用户自行配置的安全特性，如认证、SSL传输加密等，这些特性可能对性能和使用方便性造成一定影响。
- Flink作为大数据计算和分析平台，对客户输入的数据是否包含敏感信息无法感知，因此需要客户保证输入数据是脱敏的。
- 客户可以根据应用环境，权衡配置安全与否。
- 任何与安全有关的问题，请联系运维人员。

12.6.7 使用 Flink WebUI

12.6.7.1 概述

12.6.7.1.1 Flink WebUI 应用简介

Flink WebUI提供基于Web的可视化开发平台，用户只需要编写SQL即可开发作业，极大降低作业开发门槛。同时通过作业平台能力开放，支持业务人员自行编写SQL开发作业来快速应对需求，大大减少Flink作业开发工作量。

说明

Flink WebUI功能仅支持MRS 3.1.0及之后版本。

Flink WebUI 特点

Flink WebUI主要有以下特点：

- 企业级可视化运维：运维管理界面化、作业监控、作业开发Flink SQL标准化等。
- 快速建立集群连接：通过集群连接功能配置访问一个集群，需要客户端配置、用户认证密钥文件。
- 快速建立数据连接：通过数据连接功能配置访问一个组件。创建“数据连接类型”为“HDFS”类型时需创建集群连接，其他数据连接类型的“认证类型”为“KERBEROS”需创建集群连接，“认证类型”为“SIMPLE”不需创建集群连接。

  说明

  “数据连接类型”为“Kafka”时，认证类型不支持“KERBEROS”。
- 可视化开发平台：支持自定义输入/输出映射表，满足不同输入来源、不同输出目标端的需求。
- 图形化作业管理：简单易用。

Flink WebUI 关键能力

FlinkWebUI关键能力如**表12-90**：

表 12-90 Flink WebUI 关键能力

关键能力分类	描述
批流一体	● 支持一套Flink SQL定义批作业和流作业。
Flink SQL内核能力	● Flink SQL支持自定义大小窗、24小时以内流计算、超出24小时批处理。 ● Flink SQL支持Kafka、HDFS读取；支持写入Kafka、Redis和HDFS；支持Redis维表Join。 ● 支持同一个作业定义多个Flink SQL，多个指标合并在一个作业计算。当一个作业是相同主键、相同的输入和输出时，该作业支持多个窗口的计算。 ● 支持AVG、SUM、COUNT、MAX和MIN统计方法。
Flink SQL可视化定义	● 集群连接管理，配置Kafka、Redis、HDFS等服务所属的集群信息。 ● 数据连接管理，配置Kafka、Redis、HDFS等服务信息。 ● 数据表管理，定义Sql访问的数据表信息，用于生成DDL语句。 ● Flink SQL作业定义，根据用户输入的Sql，校验、解析、优化、转换成Flink作业并提交运行。
Flink作业可视化管理	● 支持可视化定义流作业和批作业。 ● 支持作业资源、故障恢复策略、Checkpoint策略可视化配置。 ● 流作业和批作业的状态监控。 ● Flink作业运维能力增强，包括原生监控页面跳转。
性能&可靠性	● 流处理支持24小时窗口聚合计算，毫秒级性能。 ● 批处理支持90天窗口聚合计算，分钟级计算完成。 ● 支持对流处理和批处理的数据进行过滤配置，过滤无效数据。 ● 读取HDFS数据时，提前根据计算周期过滤。 ● Flink作业里的数据来源于Redis，Flink作业已设置“故障恢复策略”，计算时，数据从Redis读入，作业故障无数据丢失。 ● 作业定义平台故障、服务降级，不支持再定义作业，但是不影响已有作业计算。 ● 作业故障有自动重启机制，重启策略可配置。

12.6.7.1.2 Flink WebUI 应用流程

Flink WebUI应用流程参考如下步骤：

图 12-12 Flink WebUI 应用流程

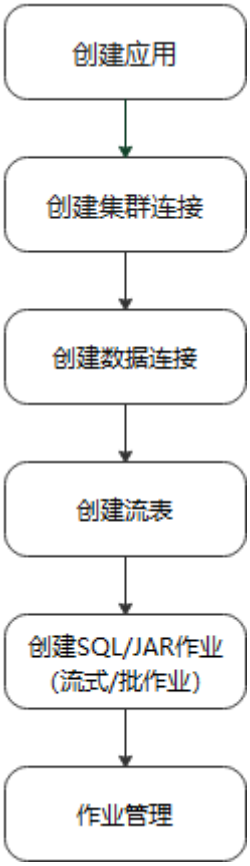


表 12-91 Flink WebUI 应用流程说明

阶段	说明	参考章节
创建应用	通过应用来隔离不同的上层业务。	**在Flink WebUI创建应用**
创建集群连接	通过集群连接配置访问不同的集群。	**在Flink WebUI创建集群连接**
创建数据连接	通过数据连接，访问不同的数据服务，包括HDFS、Kafka、Redis等。	**在Flink WebUI创建数据连接**
创建流表	通过数据表，定义源表、维表、输出表的基本属性和字段信息。	**使用Flink WebUI的流表管理**
创建SQL/JAR作业（流式/批作业）	定义Flink作业的API，包括Flink SQL和Flink Jar作业。	**使用Flink WebUI的作业管理**
作业管理	管理创建的作业，包括作业启动、开发、停止、删除和编辑等。	**使用Flink WebUI的作业管理**

12.6.7.2 FlinkServer 权限管理

12.6.7.2.1 概述

Manager的**admin**用户没有FlinkServer的业务操作权限，使用FlinkServer的业务操作需要给用户赋予相关权限。

FlinkServer中应用（租户）是最大管理范围，包含集群连接管理、数据连接管理、应用管理、流表和作业管理等。

FlinkServer中有如表12-92所示三种资源权限：

表 12-92 FlinkServer 资源权限

权限名称	权限描述	备注
管理员权限	具有所有应用的编辑、查看权限。	是FlinkServer的最高权限。如果已经具有管理员权限，则会自动具备所有应用的权限。
应用编辑权限	具有当前应用编辑权限的用户，可以执行创建、编辑和删除集群连接、数据连接，创建流表、创建作业及运行作业等操作。	同时具有当前应用查看权限。
应用查看权限	具有当前应用查看权限的用户，可以查看应用。	-

12.6.7.2.2 基于用户和角色的鉴权

该任务指导系统管理员在Manager创建并设置FlinkServer的角色。FlinkServer角色可设置管理员权限以及应用的编辑和查看权限。

用户需要在FlinkServer中对指定的用户设置权限，才能够更新数据、查询数据和删除数据等。

前提条件

管理员已根据业务需要规划权限。

操作步骤

步骤1 登录Manager。

步骤2 选择“系统 > 权限 > 角色”。

步骤3 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

步骤4 设置角色“配置资源权限”。

FlinkServer权限类型：

- FlinkServer管理员权限：是最高权限，具有FlinkServer所有应用的业务操作权限。

- FlinkServer应用权限：可设置对应用的“应用查看”、“应用编辑”权限。

表 12-93 设置角色

任务场景	角色授权操作
设置管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > Flink”，勾选“FlinkServer管理操作权限”。
设置用户对应用的指定权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Flink > FlinkServer应用”。 2. 在“权限”列，勾选“应用查看”或“应用编辑”。

步骤5 单击“确定”完成，返回角色管理。

说明

FlinkServer角色创建成功后，可参考“用户指南 > FusionInsight Manager操作指导 > 系统设置 > 权限设置> 用户管理> 创建用户”章节创建一个FlinkServer用户并绑定角色和用户组。

----结束

12.6.7.3 访问 Flink WebUI

操作场景

MRS集群安装Flink组件后，用户可以通过Flink的WebUI，在图形化界面进行集群连接、数据连接、流表管理和作业管理等。

该任务指导用户在MRS集群中访问Flink WebUI。

说明

Internet Explorer浏览器可能存在兼容性问题，建议使用Google Chrome浏览器50及以上版本访问Flink WebUI。

对系统的影响

第一次访问Manager和Flink WebUI，需要在浏览器中添加站点信任以继续访问Flink WebUI。

操作步骤

步骤1 使用具有FlinkServer管理员权限的用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Flink”。

步骤2 在“Flink WebUI”右侧，单击链接，访问Flink的WebUI。

Flink WebUI支持以下功能：

- 使用系统管理可以支持以下功能：
  - 使用集群连接管理可以创建、查看、编辑、测试和删除集群连接。
  - 使用数据连接管理可以创建、查看、编辑、测试和删除数据连接。数据连接类型包含HDFS、Kafka和Redis等。

  - 使用应用管理可以创建、查看、删除应用。
- 使用流表管理可以新建、查看、编辑和删除流表。
- 使用作业管理可以新建、查看、启动、开发、编辑、停止和删除作业等。

----结束

12.6.7.4 在 Flink WebUI 创建应用

操作场景

通过应用来隔离不同的上层业务。

创建应用

步骤1 使用具有FlinkServer管理员权限的用户访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 应用管理”，进入应用管理页面。

步骤3 单击“创建应用”，在弹出的页面中参考**表12-94**填写信息，单击“确定”，完成应用创建。

表 12-94 创建应用信息

参数名称	参数描述
应用名称	应用名称。只能包含英文字母、数字和下划线，且不能多于32个字符。
描述信息	应用描述信息。不能多于85个字符。

应用创建成功后，在Flink WebUI左上角即可切换待操作的应用，然后进行相关的作业开发。

----结束

12.6.7.5 在 Flink WebUI 创建集群连接

操作场景

通过集群连接配置访问不同的集群。

创建集群连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 集群连接管理”，进入集群连接管理页面。

步骤3 单击“创建集群连接”，在弹出的页面中参考**表12-95**填写信息，单击“确定”，完成集群连接创建。

表 12-95 创建集群连接信息

参数名称	参数描述
集群连接名称	集群连接的名称，只能包含英文字母、数字和下划线，且不能多于100个字符。
描述	集群连接名称描述信息。
版本	选择集群版本。
是否安全版本	• 是，安全集群选择是。需要输入访问用户名和上传用户凭证； • 否，非安全集群选择否。
访问用户名	访问用户需要包含访问集群中服务所需要的最小权限。只能包含英文字母、数字和下划线，且不能多于100个字符。 “是否安全版本”选择“是”时存在此参数。
客户端配置文件	集群客户端配置文件，格式为tar。
用户凭据	FusionInsight Manager中用户的认证凭据，格式为tar。 “是否安全版本”选择“是”时存在此参数。 输入访问用户名后才可上传文件。

说明

集群客户端配置文件获取方法：

1. 登录FusionInsight Manager，选择“集群 > 概览”。
2. 选择“更多 > 下载客户端 > 仅配置文件”，选择平台类型后单击“确定”。

用户凭据获取方法：

1. 登录FusionInsight Manager，单击“系统”。
2. 在对应用户的“操作”列，选择“更多 > 下载认证凭据”，选择集群后单击“确定”。

----结束

编辑集群连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 集群连接管理”，进入集群连接管理页面。

步骤3 在待修改项的“操作”列单击“编辑”，在弹出的页面中参考**表12-95**修改连接信息，单击“确定”完成修改。

----结束

测试集群连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 集群连接管理”，进入集群连接管理页面。

步骤3 在待测试项的“操作”列单击“测试”进行测试。

----结束

搜索集群连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 集群连接管理”，进入集群连接管理页面。

步骤3 在页面右上角，用户可以根据“集群连接名称”，输入查询条件后进行搜索和查看集群连接。

----结束

删除集群连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 集群连接管理”，进入集群连接管理页面。

步骤3 在待删除项的“操作”列单击“删除”，在弹出的页面中单击“确定”完成删除。

----结束

12.6.7.6 在 Flink WebUI 创建数据连接

操作场景

通过数据连接，访问不同的数据服务，当前FlinkServer支持HDFS、Kafka、Redis类型的数据连接。

创建数据连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 数据连接管理”，进入数据连接管理页面。

步骤3 单击“创建数据连接”，在弹出的页面中选择数据连接类型，参考**表12-96**填写信息，单击“确定”，完成数据连接创建。

表 12-96 创建数据连接信息

参数名称	参数描述	示例
数据连接类型	选择数据连接的类型，包含HDFS、Kafka、Redis。	-
数据连接名称	数据连接的名称。只能包含英文字母、数字和下划线，且不能多于100个字符。	-
集群连接	配置管理里的集群连接名称。 HDFS类型数据连接和认证类型为“KERBEROS”的Redis类型数据连接需配置该参数。	-

参数名称	参数描述	示例
Kafka broker	Kafka Broker实例的连接信息，格式为“IP地址:端口”，多个实例之间通过逗号分割。 Kafka类型数据连接需配置该参数。	192.168.0.1:21005,192.168.0.2:21005
Redis部署方式	Redis部署方式，当前仅支持“Cluster”。 Redis类型数据连接需配置该参数。	Cluster
Redis服务器列表	Redis实例的连接信息，格式为“IP地址:端口”，多个实例之间通过逗号分割。 Redis类型数据连接需配置该参数。	192.168.0.1:22400,192.168.0.2:22400
认证类型	• SIMPLE：表示对接的服务是非安全模式，无需认证。 • KERBEROS：表示对接的服务是安全模式，安全模式的服务统一使用Kerberos认证协议进行安全认证。 Redis类型数据连接需配置该参数。	-

----结束

编辑数据连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 数据连接管理”，进入数据连接管理页面。

步骤3 在待修改项的“操作”列单击“编辑”，在弹出的页面中参考**表12-96**修改连接信息，单击“确定”完成修改。

----结束

测试数据连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 数据连接管理”，进入数据连接管理页面。

步骤3 在待测试项的“操作”列单击“测试”进行测试。

----结束

搜索数据连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 数据连接管理”，进入数据连接管理页面。

步骤3 在页面右上角，用户可以根据“名称”进行搜索查看数据连接。

----结束

删除数据连接

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 选择“系统管理 > 数据连接管理”，进入数据连接管理页面。

步骤3 在待删除项的“操作”列单击“删除”，在弹出的页面中单击“确定”完成删除。

----结束

12.6.7.7 使用 Flink WebUI 的流表管理

操作场景

通过数据表，定义源表、维表、输出表的基本属性和字段信息。

新建流表

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“流表管理”进入流表管理页面。

步骤3 单击“新建流表”，在新建流表页面参考**表12-97**填写信息，单击“确定”，完成流表创建。

表 12-97 新建流表信息

参数名称	参数描述	备注
流/表名称	流/表的名称，只能包含英文字母、数字和下划线，且长度为1～64个字符。	例如：flink_sink
描述	流/表的描述信息，且长度为1～1024个字符。	-
映射表类型	Flink SQL本身不带有数据存储功能，所有涉及表创建的操作，实际上均是对于外部数据表、存储的引用映射。 类型包含Kafka、HDFS、Redis。	-
类型	包含数据源表Source，数据结果表Sink，数据维表Table。不同映射表类型包含的表如下所示。 • Kafka：Source、Sink • HDFS：Source、Sink • Redis：Sink、Table	-
数据连接	选择数据连接。	-
Topic	读取的Kafka的topic，支持从多个Kakfa topic中读取，topic之间使用英文分隔符进行分隔。 “映射表类型”选择“Kafka”时存在此参数。	-

参数名称	参数描述	备注
文件路径	要传输的HDFS目录或单个文件路径。 “映射表类型”选择“HDFS”时存在此参数。	例如： “/user/sqoop/ ”或“/user/sqoop/example.csv”
编码	选择不同“映射表类型”对应的编码如下： ● Kafka：CSV、JSON ● HDFS：CSV ● Redis： – “类型”为“Sink”时：String、List、Set、Zset、Hash – “类型”为“Table”时：String、Zset	-
前缀	“映射表类型”选择“Kafka”，且“类型”选择“Source”，“编码”选择“JSON”时含义为：多层嵌套json的层级前缀，使用英文逗号(,)进行分隔。	例如：data,info表示取嵌套json中data，info下的内容，作为json格式数据输入
	“映射表类型”选择“Redis”时含义为：操作数据时会自动在key上补充前缀或手动输入。	例如：key的值是key1，前缀是test，则最终写入redis的key是test:key1
分隔符	选择不同“映射表类型”对应的含义如下： ● Kafka：用于指定CSV字段分隔符。当数据“编码”为“CSV”时存在此参数。 ● Redis：字段分隔符。	例如：“,”
行分隔符	文件中的换行符，包含“\r”、“\n”、“\r\n”。 “映射表类型”选择“HDFS”时存在此参数。	-
列分隔符	文件中的字段分隔符。 “映射表类型”选择“HDFS”时存在此参数。	例如：“,”
数据有效期	数据的有效期。分为“永久有效”、“有效时长”和“截止日期”三类。 “映射表类型”选择“Redis”，“类型”选择“Sink”时存在此参数。	-
流/表结构	填写流/表结构，包含名称，类型。	-
Proctime	指系统时间，与数据本身的时间戳无关，即在Flink算子内计算完成的时间。 “类型”选择“Source”时存在此参数。	-

参数名称	参数描述	备注
Event Time	指事件产生的时间，即数据产生时自带时间戳。“类型”选择“Source”时存在此参数。	-

----结束

编辑流表

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“流表管理”进入流表管理页面。

步骤3 在待修改项的“操作”列单击“编辑”，在弹出的页面中参考**表12-97**修改流表信息，单击“确定”完成修改。

----结束

搜索流表

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“流表管理”进入流表管理页面。

步骤3 在页面右上角，用户可以输入关键字搜索查看流表信息。

----结束

删除流表

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“流表管理”进入流表管理页面。

步骤3 在待删除项的“操作”列单击“删除”，在弹出的页面单击“确定”完成删除。

----结束

12.6.7.8 使用 Flink WebUI 的作业管理

操作场景

定义Flink的作业，包括Flink SQL和Flink Jar作业。

新建流表

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 单击“新建作业”，在新建作业页面参考**表12-98**填写信息，单击“确定”，创建作业成功并进入作业开发界面。

表 12-98 新建作业信息

参数名称	参数描述
类型	作业类型，包括Flink SQL和Flink Jar。
名称	作业名称，只能包含英文字母、数字和下划线，且不能多于64个字符。
作业类型	作业数据来源类型，包括流作业和批作业。
描述	作业描述，不能超过100个字符。

步骤4 （可选）如果需要立即进行作业开发，可以在作业开发界面进行作业配置。

- 新建Flink SQL作业
  a. 在作业开发界面进行作业开发。
  b. 可以单击上方“语义校验”对输入内容校验，单击“SQL格式化”对SQL语句进行格式化。
  c. 作业SQL开发完成后，请参考**表12-99**设置基础参数，还可根据需要设置自定义参数，然后单击“保存”。

  表 12-99 基础参数

参数名称	参数描述
并行度	并行数量，只能填写正整数，且不能多于64字符。
算子最大并行度	算子最大的并行度，只能填写正整数，且不能多于64字符。
JobManager内存（MB）	JobManager的内存。输入值最小为512，且不能超过64个字符。
提交队列	作业提交队列。不填默认提交到default。只能包含英文字母，数字和下划线，且不能超过30字符。
taskManager	taskManager运行参数。该参数需配置以下内容： ■ slot数量：不填默认是1； ■ 内存（MB）：输入值最小为512。

参数名称	参数描述
开启CheckPoint	是否开启CheckPoint。开启后，需配置以下内容： ■ 时间间隔（ms）：必填； ■ 模式：必填； 可选项为：EXACTLY_ONCE、AT_LEAST_ONCE； ■ 最小间隔（ms）：输入值最小为10； ■ 超时时间：输入值最小为10； ■ 最大并发量：正整数，且不能超过64个字符； ■ 是否清理：是/否； ■ 是否开启增量Checkpoint：是/否。
故障恢复策略	作业的故障恢复策略，包含以下三种。 ■ fixed-delay：需配置“重试次数”和“失败重试间隔（s）”； ■ failure-rate：需配置“最大重试次数”、“时间间隔（min）”和“失败重试间隔（s）”； ■ none：无。

d. 单击左上角“提交”提交作业。

- 新建Flink Jar作业

a. 单击“选择”，上传本地Jar文件，并参考**表12-100**配置参数或添加自定义参数。

表 12-100 参数配置

参数名称	参数描述
本地jar文件	上传jar文件。直接上传本地文件，大小不能超过10M。
Main Class	Main-Class类型。 ■ 默认：默认根据Jar包文件的Mainfest文件指定类名。 ■ 指定：手动指定类名。
类名	类名。 “Main Class”选择“指定”时存在该参数。
类参数	类参数，为Main-Class的参数（参数间用空格分隔）。

参数名称	参数描述
并行度	并行数量，只能填写正整数，且不能多于64字符。
JobManager内存（MB）	JobManager的内存。输入值最小为512，且不能超过64个字符。
提交队列	作业提交队列。不填默认提交到default。只能包含英文字母，数字和下划线，且不能超过30字符。
taskManager	taskManager运行参数。该参数需配置以下内容： ■ slot数量：不填默认是1； ■ 内存（MB）：输入值最小为512。

b. 单击“保存”保存配置，单击“提交”提交作业。

步骤5 返回作业管理页面，可以查看到已创建的作业名称、类型、状态、作业种类和描述等信息。

----结束

启动作业

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待启动项的“操作”列单击“启动”运行作业。作业状态为“草稿”、“保存”、“提交失败”、“运行成功”、“运行失败”和“停止”的作业可以启动。

----结束

开发作业

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待开发项的“操作”列单击“开发”进入作业开发页面，参考**步骤4**进行作业开发，在左侧列表可以查看已创建的流表及字段。

----结束

编辑作业名称和描述

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待修改项的“操作”列单击“编辑”，修改“描述”，修改完成后单击“确定”保存修改。

----结束

查看作业详情

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待查看项的“操作”列选择“更多 > 作业详情”可以查看作业运行详情。

说明

只能查看状态为“运行中”的作业详情。

----结束

Checkpoint 故障恢复

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待恢复项的“操作”列选择“更多 > Checkpoint故障恢复”进行Checkpoint故障恢复。作业状态为“运行失败”、“运行成功”和“停止”的作业可以进行Checkpoint故障恢复。

----结束

筛选/搜索作业

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在页面右上角，用户可以根据作业名称进行筛选，或输入关键字搜索查看作业信息。

----结束

停止作业

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待停止项的“操作”列单击“停止”，停止作业运行。作业状态为“提交中”、“提交成功”和“运行中”的作业可以停止。

----结束

删除作业

步骤1 访问Flink WebUI，请参考**访问Flink WebUI**。

步骤2 单击“作业管理”进入作业管理页面。

步骤3 在待删除项的“操作”列单击“删除”在弹出的页面单击“确定”删除作业。作业状态为“草稿”、“保存”、“提交失败”、“运行成功”、“运行失败”和“停止”状态的作业可以删除。

----结束

12.6.8 Flink 日志介绍

日志描述

日志存储路径：

- Flink作业运行日志：“${BIGDATA_DATA_HOME}/hadoop/data${i}/nm/containerlogs/application_${appid}/container_{$contid}”。

  说明

  运行中的任务日志存储在以上路径中，运行结束后会基于Yarn的配置确定是否汇聚到HDFS目录中。

- FlinkResource运行日志：“/var/log/Bigdata/flink/flinkResource”。

日志归档规则：

1. FlinkResource运行日志：
   - 服务日志默认20MB滚动存储一次，最多保留20个文件，不压缩。

     说明

     针对MRS 3.x之前版本，Executor日志默认30MB滚动存储一次，最多保留20个文件，不压缩。
   - 日志大小和压缩文件保留个数可以在Manager界面中配置或者修改客户端“/opt/client/Flink/flink/conf/”中的log4j-cli.properties、log4j.properties、log4j-session.properties中对应的配置项。其中“/opt/client”为客户端安装目录。

表 12-101 FlinkResource 日志列表

日志类型	日志文件名	描述
FlinkResource运行日志	checkService.log	健康检查日志。
	kinit.log	初始化日志。
	postinstall.log	服务安装日志。
	prestart.log	prestart脚本日志。
	start.log	启动日志。

日志级别

Flink中提供了如表12-102所示的日志级别。日志级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-102 日志级别

级别	描述
ERROR	ERROR表示当前时间处理存在错误信息。

级别	描述
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 请参考**修改集群服务配置参数**，进入Flink的“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

----结束

说明

- 配置完成后不需要重启服务，重新下载客户端使配置生效。
- 也可以直接修改客户端“/opt/client/Flink/flink/conf/”中log4j-cli.properties、log4j.properties、log4j-session.properties文件中对应的日志级别配置项。其中“/opt/client”为客户端安装目录。
- 通过客户端提交作业时会在客户端log文件夹中生成相应日志文件，由于系统默认umask值是0022，所以日志默认权限为644；如果需要修改文件权限，需要修改umask值；例如修改omm用户umask值：
  - 在“/home/omm/.baskrc”文件末尾添加“umask 0026”；
  - 执行命令**source /home/omm/.baskrc**使文件权限生效。

日志格式

表 12-103 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2019-06-27 21:30:31,778	INFO	[flink-akka.actor.default-dispatcher-3]	TaskManager container_e10_1498290698388_0004_02_000007 has started.	org.apache.flink.yarn.YarnFlinkResourceManager (FlinkResourceManager.java:368)

12.6.9 Flink 性能调优

12.6.9.1 DataStream 调优

12.6.9.1.1 配置内存

操作场景

Flink是依赖内存计算，计算过程中内存不够对Flink的执行效率影响很大。可以通过监控GC（Garbage Collection），评估内存使用及剩余情况来判断内存是否变成性能瓶颈，并根据情况优化。

监控节点进程的YARN的Container GC日志，如果频繁出现Full GC，需要优化GC。

说明

GC的配置：在客户端的“conf/flink-conf.yaml”配置文件中，在“env.java.opts”配置项中添加参数：“-Xloggc:<LOG_DIR>/gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=20M”。 此处默认已经添加GC日志。

操作步骤

- 优化GC。

  调整老年代和新生代的比值。在客户端的“conf/flink-conf.yaml”配置文件中，在“env.java.opts”配置项中添加参数：“-XX:NewRatio”。如“-XX:NewRatio=2”，则表示老年代与新生代的比值为2:1，新生代占整个堆空间的1/3，老年代占2/3。
- 开发Flink应用程序时，优化DataStream的数据分区或分组操作。
  - 当分区导致数据倾斜时，需要考虑优化分区。
  - 避免非并行度操作，有些对DataStream的操作会导致无法并行，例如WindowAll。
  - keyBy尽量不要使用String。

12.6.9.1.2 设置并行度

操作场景

并行度控制任务的数量，影响操作后数据被切分成的块数。调整并行度让任务的数量和每个任务处理的数据与机器的处理能力达到更优。

查看CPU使用情况和内存占用情况，当任务和数据不是平均分布在各节点，而是集中在个别节点时，可以增大并行度使任务和数据更均匀的分布在各个节点。增加任务的并行度，充分利用集群机器的计算能力。

操作步骤

任务的并行度可以通过以下四种层次（按优先级从高到低排列）指定，用户可以根据实际的内存、CPU、数据以及应用程序逻辑的情况调整并行度参数。

- 算子层次

  一个算子、数据源和sink的并行度可以通过调用setParallelism()方法来指定，例如

```
final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();

DataStream<String> text = [...]
DataStream<Tuple2<String, Integer>> wordCounts = text
    .flatMap(new LineSplitter())
```

```
    .keyBy(0)
    .timeWindow(Time.seconds(5))
    .sum(1).setParallelism(5);

wordCounts.print();

env.execute("Word Count Example");
```

- 执行环境层次

  Flink程序运行在执行环境中。执行环境为所有执行的算子、数据源、data sink定义了一个默认的并行度。

  执行环境的默认并行度可以通过调用setParallelism()方法指定。例如：

  ```
  final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
  env.setParallelism(3);
  DataStream<String> text = [...]
  DataStream<Tuple2<String, Integer>> wordCounts = [...]
  wordCounts.print();
  env.execute("Word Count Example");
  ```

- 客户端层次

  并行度可以在客户端将job提交到Flink时设定。对于CLI客户端，可以通过“-p”参数指定并行度。例如：

  ```
  ./bin/flink run -p 10 ../examples/*WordCount-java*.jar
  ```

- 系统层次

  在系统级可以通过修改Flink客户端conf目录下的“flink-conf.yaml”文件中的“parallelism.default”配置选项来指定所有执行环境的默认并行度。

12.6.9.1.3 配置进程参数

操作场景

Flink on YARN模式下，有JobManager和TaskManager两种进程。在任务调度和运行的过程中，JobManager和TaskManager承担了很大的责任。

因而JobManager和TaskManager的参数配置对Flink应用的执行有着很大的影响意义。用户可通过如下操作对Flink集群性能做优化。

操作步骤

步骤1 配置JobManager内存。

JobManager负责任务的调度，以及TaskManager、RM之间的消息通信。当任务数变多，任务平行度增大时，JobManager内存都需要相应增大。

您可以根据实际任务数量的多少，为JobManager设置一个合适的内存。

- 在使用yarn-session命令时，添加“-jm MEM”参数设置内存。
- 在使用yarn-cluster命令时，添加“-yjm MEM”参数设置内存。

步骤2 配置TaskManager个数。

每个TaskManager每个核同时能跑一个task，所以增加了TaskManager的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加TaskManager的个数，以提高运行效率。

步骤3 配置TaskManager Slot数。

每个TaskManager多个核同时能跑多个task，相当于增大了任务的并发度。但是由于所有核共用TaskManager的内存，所以要在内存和核数之间做好平衡。

- 在使用yarn-session命令时，添加“-s NUM”参数设置SLOT数。
- 在使用yarn-cluster命令时，添加“-ys NUM”参数设置SLOT数。

步骤4 配置TaskManager内存。

TaskManager的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加。

- 将在使用yarn-session命令时，添加“-tm MEM”参数设置内存。
- 将在使用yarn-cluster命令时，添加“-ytm MEM”参数设置内存。

----结束

12.6.9.1.4 设计分区方法

操作场景

合理的设计分区依据，可以优化task的切分。在程序编写过程中要尽量分区均匀，这样可以实现每个task数据不倾斜，防止由于某个task的执行时间过长导致整个任务执行缓慢。

操作步骤

以下是几种分区方法。

- **随机分区：**将元素随机地进行分区。
  ```
  dataStream.shuffle();
  ```
- **Rebalancing (Round-robin partitioning)：**基于round-robin对元素进行分区，使得每个分区负责均衡。对于存在数据倾斜的性能优化是很有用的。
  ```
  dataStream.rebalance();
  ```
- **Rescaling：**以round-robin的形式将元素分区到下游操作的子集中。如果你想要将数据从一个源的每个并行实例中散发到一些mappers的子集中，用来分散负载，但是又不想要完全的rebalance 介入（引入`rebalance()`），这会非常有用。
  ```
  dataStream.rescale();
  ```
- **广播：**广播每个元素到所有分区。
  ```
  dataStream.broadcast();
  ```
- **自定义分区：**使用一个用户自定义的Partitioner对每一个元素选择目标task，由于用户对自己的数据更加熟悉，可以按照某个特征进行分区，从而优化任务执行。

  简单示例如下所示：
  ```
  // fromElements构造简单的Tuple2流
  DataStream<Tuple2<String, Integer>> dataStream = env.fromElements(Tuple2.of("hello",1),
  Tuple2.of("test",2), Tuple2.of("world",100));

  // 定义用于分区的key值，返回即属于哪个partition的，该值加1就是对应的子任务的id号
  Partitioner<Tuple2<String, Integer>> strPartitioner = new Partitioner<Tuple2<String, Integer>>() {
      @Override
      public int partition(Tuple2<String, Integer> key, int numPartitions) {
          return (key.f0.length() + key.f1) % numPartitions;
      }
  };

  // 使用Tuple2进行分区的key值
  dataStream.partitionCustom(strPartitioner, new KeySelector<Tuple2<String, Integer>, Tuple2<String,
  Integer>>() {
      @Override
      public Tuple2<String, Integer> getKey(Tuple2<String, Integer> value) throws Exception {
          return value;
  ```

```
    }
}).print();
```

12.6.9.1.5 配置 netty 网络通信

操作场景

Flink通信主要依赖netty网络，所以在Flink应用执行过程中，netty的设置尤为重要，网络通信的好坏直接决定着数据交换的速度以及任务执行的效率。

操作步骤

以下配置均可在客户端的“conf/flink-conf.yaml”配置文件中进行修改适配，默认已经是相对较优解，请谨慎修改，防止性能下降。

- “taskmanager.network.netty.num-arenas”： 默认是“taskmanager.numberOfTaskSlots”，表示netty的域的数量。
- “taskmanager.network.netty.server.numThreads”和“taskmanager.network.netty.client.numThreads”：默认是“taskmanager.numberOfTaskSlots”，表示netty的客户端和服务端的线程数目设置。
- “taskmanager.network.netty.client.connectTimeoutSec”：默认是120s，表示taskmanager的客户端连接超时的时间。
- “taskmanager.network.netty.sendReceiveBufferSize”：默认是系统缓冲区大小(cat /proc/sys/net/ipv4/tcp_[rw]mem)，一般为4MB，表示netty的发送和接收的缓冲区大小。
- “taskmanager.network.netty.transport”：默认为“nio”方式，表示netty的传输方式，有“nio”和“epoll”两种方式。

12.6.9.1.6 经验总结

数据倾斜

当数据发生倾斜（某一部分数据量特别大），虽然没有GC（Gabage Collection，垃圾回收），但是task执行时间严重不一致。

- 需要重新设计key，以更小粒度的key使得task大小合理化。
- 修改并行度。
- 调用rebalance操作，使数据分区均匀。

缓冲区超时设置

- 由于task在执行过程中存在数据通过网络进行交换，数据在不同服务器之间传递的缓冲区超时时间可以通过setBufferTimeout进行设置。
- 当设置“setBufferTimeout(-1)”，会等待缓冲区满之后才会刷新，使其达到最大吞吐量；当设置“setBufferTimeout(0)”时，可以最小化延迟，数据一旦接收到就会刷新；当设置“setBufferTimeout”大于0时，缓冲区会在该时间之后超时，然后进行缓冲区的刷新。

  示例可以参考如下：

```
env.setBufferTimeout(timeoutMillis);

env.generateSequence(1,10).map(new MyMapper()).setBufferTimeout(timeoutMillis);
```

12.6.10 Flink 常见 Shell 命令

本章节适用于MRS 3.x及之后版本。

在使用Flink的Shell脚本前，首先需要执行以下操作：

步骤1 安装Flink客户端，例如安装目录为“/opt/client”。

步骤2 初始化环境变量。

source /opt/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *业务用户*

步骤4 参考**表12-104**运行相关命令。

表 12-104 Flink Shell 命令参考

命令	参数说明	描述
yarn-session.sh	-at,--applicationType <arg>：为Yarn application自定义类型。 -D <property=value>：动态参数配置。 -d,--detached：关闭交互模式，启动一个分离的Flink YARN session。 -h,--help: 显示Yarn session CLI的帮助。 -id,--applicationId <arg>：绑定到一个已经运行的Yarn session。 -j,--jar <arg>：设置用户jar包路径。 -jm,--jobManagerMemory <arg>：为JobManager设置内存。 -m,--jobmanager <arg>：要连接的JobManager的地址，使用该参数可以连接特定的JobManager。 -nl,--nodeLabel <arg>: 指定YARN application的nodeLabel 。 -nm,--name <arg>：为Yarn application自定义名称。 -q,--query：查询可用的Yarn 资源。 -qu,--queue <arg>：指定YARN 队列。 -s,--slots <arg>：设置每个Taskmanager的SLOT个数。 -t,--ship <arg>：指定待发送文件的目录。 -tm,--taskManagerMemory <arg>：为TaskManager设置内存。 -yd,--yarndetached：以分离模式启动。 -z,--zookeeperNamespace <args>：指定zookeeper的namespace。 -h：获取帮助。	启动一个常驻的Flink集群，接受来自Flink客户端的任务。

命令	参数说明	描述
flink run	-c,--class <classname>：指定一个类作为程序运行的入口点。 -C,--classpath <url>：指定classpath。 -d,--detached：以分离方式运行job。 -files,--dependencyFiles <arg>：Flink程序依赖的文件。 -n,--allowNonRestoredState： 从快照点恢复时允许跳过不能恢复的状态。比如删除了程序中某个操作符，那么在恢复快照点时需要增加该参数。 -m,--jobmanager <host:port>：指定JobManager。 -p,--parallelism <parallelism>：指定job并行度，会覆盖配置文件中配置的并行度参数。 -q,--sysoutLogging：禁止flink日志输出至控制台。 -s,--fromSavepoint <savepointPath>：指定用于恢复job的savepoint路径。 -z,--zookeeperNamespace <zookeeperNamespace>：指定zookeeper的namespace。 -yat,--yarnapplicationType <arg>：为Yarn application自定义类型。 -yD <arg>：动态参数配置。 -yd,--yarndetached：以分离模式启动。 -yh,--yarnhelp：获取yarn帮助。 -yid,--yarnapplicationId <arg>：绑定到yarn session运行job。 -yj,--yarnjar <arg>：设置Flink jar文件路径。 -yjm,--yarnjobManagerMemory <arg>：为JobManager设置内存（MB）。 -ynm,--yarnname <arg>：为Yarn application自定义名称。 -yq,--yarnquery：查询可用的YARN资源（内存、CPU）。 -yqu,--yarnqueue <arg>：指定YARN队列。 -ys,--yarnslots：设置每个TaskManager的SLOT个数。 -yt,--yarnship <arg>：指定待发送文件的路径。 -ytm,--yarntaskManagerMemory <arg>：为TaskManager设置内存（MB）。	Flink提交作业。 1."-y*"参数是指yarn-cluster模式下使用。 2.非"-y*"参数用户在用该命令提交任务前需要先用yarn-session启动Flink集群。

命令	参数说明	描述
	-yz,--yarnzookeeperNamespace <arg>：指定zookeeper的namespace，需与yarn-session.sh -z 保持一致。 -h：获取帮助。	
flink info	-c,--class <classname>：指定一个类作为程序运行的入口点。 -p,--parallelism <parallelism>：指定程序运行的并行度。 -h：获取帮助。	显示所运行程序的执行计划（JSON）
flink list	-a,--all：显示所有的Job。 -m,--jobmanager <host:port>：指定JobManager。 -r,--running：仅显示running状态的Job。 -s,--scheduled：仅显示scheduled状态的Job。 -z,--zookeeperNamespace <zookeeperNamespace>：指定zookeeper的namespace。 -yid,--yarnapplicationId <arg>：绑定YARN session。 -h：获取帮助。	查询集群中运行的程序。
flink stop	-d,--drain：在触发savepoint和停止作业之前，发送MAX_WATERMARK。 -p,--savepointPath <savepointPath>：savepoint的储存路径，默认目录state.savepoints.dir。 -m,--jobmanager <host:port>：指定JobManager。 -z,--zookeeperNamespace <zookeeperNamespace>：指定zookeeper的namespace。 -yid,--yarnapplicationId <arg>：绑定YARN session。 -h：获取帮助。	强制停止一个运行中的Job（仅支持 streaming jobs、业务代码 source 端需要 implements StoppableFunction）

命令	参数说明	描述
flink cancel	-m,--jobmanager <host:port>：指定JobManager。 -s,--withSavepoint <targetDirectory>：取消Job时触发savepoint，默认目录state.savepoints.dir -z,--zookeeperNamespace <zookeeperNamespace>：指定zookeeper的namespace。 -yid,--yarnapplicationId <arg>：绑定YARN session。 -h：获取帮助。	取消一个运行中Job
flink savepoint	-d,--dispose <arg>：指定savepoint的保存目录。 -m,--jobmanager <host:port>：指定JobManager。 -z,--zookeeperNamespace <zookeeperNamespace>：指定zookeeper的namespace。 -yid,--yarnapplicationId <arg>：绑定YARN session。 -h：获取帮助。	触发一个savepoint
source *客户端安装目录*/**bigdata_env**	无	导入客户端环境变量。 使用限制：如果用户使用自定义脚本（例如A.sh）并在脚本中调用该命令，则脚本A.sh不能传入参数。如果确实需要给A.sh传入参数，则需采用二次调用方式。 例如A.sh中调用B.sh，在B.sh中调用该命令。A.sh可以传入参数，B.sh不能传入参数。
start-scala-shell.sh	local | remote <host> <port> | yarn：运行模式	scala shell启动脚本

命令	参数说明	描述
sh generate_keystore.sh	-	用户调用“generate_keystore.sh”脚本工具生成“Security Cookie”、“flink.keystore”和“flink.truststore”。需要输入自定义密码（不能包含#）。

----结束

12.6.11 参考

12.6.11.1 签发证书样例

将该样例代码生成generate_keystore.sh脚本，放置在Flink客户端的bin目录下。

```
#!/bin/bash

KEYTOOL=${JAVA_HOME}/bin/keytool
KEYSTOREPATH="$FLINK_HOME/conf/"
CA_ALIAS="ca"
CA_KEYSTORE_NAME="ca.keystore"
CA_DNAME="CN=Flink_CA"
CA_KEYALG="RSA"
CLIENT_CONF_YAML="$FLINK_HOME/conf/flink-conf.yaml"
KEYTABPRINCEPAL=""

function getConf()
{
    if [ $# -ne 2 ]; then
        echo "invalid parameters for getConf"
        exit 1
    fi

    confName="$1"
    if [ -z "$confName" ]; then
        echo "conf name is empty."
        exit 2
    fi

    configFile=$FLINK_HOME/conf/client.properties
    if [ ! -f $configFile ]; then
        echo $configFile" is not exist."
        exit 3
    fi

    defaultValue="$2"
    cnt=$(grep $1 $configFile | wc -l)
    if [ $cnt -gt 1 ]; then
        echo $confName" has multi values in "$configFile
        exit 4
    elif [ $cnt -lt 1 ]; then
        echo $defaultValue
    else
        line=$(grep $1 $configFile)
        confValue=$(echo "${line#*=}")
        echo "$confValue"
```

```
    fi
}

function createSelfSignedCA()
{
    #varible from user input
    keystorePath=$1
    storepassValue=$2
    keypassValue=$3

    #generate ca keystore
    rm -rf $keystorePath/$CA_KEYSTORE_NAME
    $KEYTOOL -genkeypair -alias $CA_ALIAS -keystore $keystorePath/$CA_KEYSTORE_NAME -dname
$CA_DNAME -storepass $storepassValue -keypass $keypassValue -validity 3650 -keyalg $CA_KEYALG -
keysize 3072 -ext bc=ca:true
    if [ $? -ne 0 ]; then
        echo "generate ca.keystore failed."
        exit 1
    fi

    #generate ca.cer
    rm -rf "$keystorePath/ca.cer"
    $KEYTOOL -keystore "$keystorePath/$CA_KEYSTORE_NAME" -storepass "$storepassValue" -alias
$CA_ALIAS -validity 3650 -exportcert > "$keystorePath/ca.cer"
    if [ $? -ne 0 ]; then
        echo "generate ca.cer failed."
        exit 1
    fi

    #generate ca.truststore
    rm -rf "$keystorePath/flink.truststore"
    $KEYTOOL -importcert -keystore "$keystorePath/flink.truststore" -alias $CA_ALIAS -storepass
"$storepassValue" -noprompt -file "$keystorePath/ca.cer"
    if [ $? -ne 0 ]; then
        echo "generate ca.truststore failed."
        exit 1
    fi
}

function generateKeystore()
{
    #get path/pass from input
    keystorePath=$1
    storepassValue=$2
    keypassValue=$3

    #get value from conf
    aliasValue=$(getConf "flink.keystore.rsa.alias" "flink")
    validityValue=$(getConf "flink.keystore.rsa.validity" "3650")
    keyalgValue=$(getConf "flink.keystore.rsa.keyalg" "RSA")
    dnameValue=$(getConf "flink.keystore.rsa.dname" "CN=flink.com")
    SANValue=$(getConf "flink.keystore.rsa.ext" "ip:127.0.0.1")
    SANValue=$(echo "$SANValue" | xargs)
    SANValue="ip:$(echo "$SANValue"| sed 's/,/,ip:/g')"

    #generate keystore
    rm -rf $keystorePath/flink.keystore
    $KEYTOOL -genkeypair -alias $aliasValue -keystore $keystorePath/flink.keystore -dname $dnameValue -
ext SAN=$SANValue -storepass $storepassValue -keypass $keypassValue -keyalg $keyalgValue -keysize
3072 -validity 3650
    if [ $? -ne 0 ]; then
        echo "generate flink.keystore failed."
        exit 1
    fi

    #generate cer
    rm -rf $keystorePath/flink.csr
    $KEYTOOL -certreq -keystore $keystorePath/flink.keystore -storepass $storepassValue -alias $aliasValue -
file $keystorePath/flink.csr
```

```
    if [ $? -ne 0 ]; then
        echo "generate flink.csr failed."
        exit 1
    fi

    #generate flink.cer
    rm -rf $keystorePath/flink.cer
    $KEYTOOL -gencert -keystore $keystorePath/ca.keystore -storepass $storepassValue -alias $CA_ALIAS -
ext SAN=$SANValue -infile $keystorePath/flink.csr -outfile $keystorePath/flink.cer -validity 3650
    if [ $? -ne 0 ]; then
        echo "generate flink.cer failed."
        exit 1
    fi

    #import cer into keystore
    $KEYTOOL -importcert -keystore $keystorePath/flink.keystore -storepass $storepassValue -file
$keystorePath/ca.cer -alias $CA_ALIAS -noprompt
    if [ $? -ne 0 ]; then
        echo "importcert ca."
        exit 1
    fi

    $KEYTOOL -importcert -keystore $keystorePath/flink.keystore -storepass $storepassValue -file
$keystorePath/flink.cer -alias $aliasValue -noprompt;
    if [ $? -ne 0 ]; then
        echo "generate flink.truststore failed."
        exit 1
    fi
}

function configureFlinkConf()
{
    # set config
    if [ -f "$CLIENT_CONF_YAML" ]; then
        SSL_ENCRYPT_ENABLED=$(grep "security.ssl.encrypt.enabled" "$CLIENT_CONF_YAML" | awk '{print
$2}')
        if [ "$SSL_ENCRYPT_ENABLED" = "false" ];then

            sed -i s/"security.ssl.key-password:".*/"security.ssl.key-password:"\ "${keyPass}"/g
"$CLIENT_CONF_YAML"
            if [ $? -ne 0 ]; then
                echo "set security.ssl.key-password failed."
                return 1
            fi

            sed -i s/"security.ssl.keystore-password:".*/"security.ssl.keystore-password:"\ "${storePass}"/g
"$CLIENT_CONF_YAML"
            if [ $? -ne 0 ]; then
                echo "set security.ssl.keystore-password failed."
                return 1
            fi

            sed -i s/"security.ssl.truststore-password:".*/"security.ssl.truststore-password:"\ "${storePass}"/g
"$CLIENT_CONF_YAML"
            if [ $? -ne 0 ]; then
                echo "set security.ssl.keystore-password failed."
                return 1
            fi

            echo "security.ssl.encrypt.enabled is false, set security.ssl.key-password security.ssl.keystore-
password security.ssl.truststore-password success."
        else
            echo "security.ssl.encrypt.enabled is true, please enter security.ssl.key-password security.ssl.keystore-
password security.ssl.truststore-password encrypted value in flink-conf.yaml."
        fi

        keystoreFilePath="${keystorePath}"/flink.keystore
        sed -i 's#'"security.ssl.keystore:".*'#'"security.ssl.keystore:"\ "$keystoreFilePath"'#g'
"$CLIENT_CONF_YAML"
```

```
        if [ $? -ne 0 ]; then
            echo "set security.ssl.keystore failed."
            return 1
        fi


        truststoreFilePath="${keystorePath}/flink.truststore"
        sed -i 's#'"security.ssl.truststore:".*'#'"security.ssl.truststore:"\ "$truststoreFilePath"'#g'
"$CLIENT_CONF_YAML"
        if [ $? -ne 0 ]; then
            echo "set security.ssl.truststore failed."
            return 1
        fi

        command -v sha256sum >/dev/null
        if [ $? -ne 0 ];then
            echo "sha256sum is not exist, it will produce security.cookie with date +%F-%H-%M-%s-%N."
            cookie=$(date +%F-%H-%M-%s-%N)
        else
            cookie="$(echo "${KEYTABPRINCEPAL}"| sha256sum | awk '{print $1}')"
        fi

        sed -i s/"security.cookie:".*/"security.cookie:"\ "${cookie}"/g "$CLIENT_CONF_YAML"
        if [ $? -ne 0 ]; then
            echo "set security.cookie failed."
            return 1
        fi
    fi
    return 0;
}

main()
{
    #check environment variable is set or not
    if [ -z ${FLINK_HOME+x} ]; then
        echo "errro: environment variables are not set."
        exit 1
    fi
    stty -echo
    read -rp "Enter password:" password
    stty echo
    echo


    KEYTABPRINCEPAL=$(grep "security.kerberos.login.principal" "$CLIENT_CONF_YAML" | awk '{print $2}')
    if [ -z "$KEYTABPRINCEPAL" ];then
        echo "please config security.kerberos.login.principal info first."
        exit 1
    fi


    #get input
    keystorePath="$KEYSTOREPATH"
    storePass="$password"
    keyPass="$password"

    #generate self signed CA
    createSelfSignedCA "$keystorePath" "$storePass" "$keyPass"
    if [ $? -ne 0 ]; then
        echo "create self signed ca failed."
        exit 1
    fi

    #generate keystore
    generateKeystore "$keystorePath" "$storePass" "$keyPass"
    if [ $? -ne 0 ]; then
        echo "create keystore failed."
        exit 1
    fi
```

```
    echo "generate keystore/truststore success."

    # set flink config
    configureFlinkConf "$keystorePath" "$storePass" "$keyPass"
    if [ $? -ne 0 ]; then
        echo "configure Flink failed."
        exit 1
    fi

    return 0;
}

#the start main
main "$@"

exit 0
```

说明

执行命令“sh generate_keystore.sh <password>”即可，<password>由用户自定义输入

- 若<password>中包含特殊字符"$"，应使用如下方式，以防止被转义，“sh generate_keystore.sh 'Bigdata_2013'”
- 密码不允许包含“#”。
- 使用该generate_keystore.sh脚本前需要在客户端目录下执行source bigdata_env。
- 使用该generate_keystore.sh脚本会自动将security.ssl.keystore、security.ssl.truststore的绝对路径填写到flink-conf.yaml中，所以需要用户根据实际情况手动修改为相对路径。例如：
  - 将security.ssl.keystore: /opt/client/Flink/flink/conf//flink.keystore修改为security.ssl.keystore: ssl/flink.keystore；
  - 将security.ssl.truststore: /opt/client/Flink/flink/conf//flink.truststore修改为security.ssl.truststore: ssl/flink.truststore；
  - 需要在Flink客户端环境中任意目录下创建ssl文件夹，如在“/opt/client/Flink/flink/conf/”目录下新建目录ssl，将flink.keystore、flink.truststore文件放入ssl文件夹中；
  - 执行yarn-session或者flink run -m yarn-cluster命令时需要在ssl文件夹同级目录下执行：yarn-session -t ssl -d 或者 flink run -m yarn-cluster -yt ssl -d WordCount.jar 。

12.7 使用 Flume

12.7.1 从零开始使用 Flume

操作场景

Flume支持将采集的日志信息导入到Kafka。

前提条件

- 已创建启用Kerberos认证的流集群。
- 已在日志生成节点安装Flume客户端，例如安装目录为“/opt/Flumeclient”，客户端安装请参见“组件操作指南 > 使用Flume > 安装Flume客户端”。以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 已配置网络，使日志生成节点与流集群互通。

使用 Flume 客户端（MRS 3.x 之前版本）

> **说明**
>
> 普通集群不需要执行**步骤2**-**步骤6**。

步骤1 客户端安装。

步骤2 将Master1节点上的认证服务器配置文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/MRS_Current/1_X_KerberosClient/etc/kdc.conf”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤3 查看任一部署Flume角色节点的“业务IP”。

登录集群详情页面，选择“集群 > 组件管理 > Flume > 实例”，查看任一部署Flume角色节点的“业务IP”。

> **说明**
>
> 若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤4 将此节点上的用户认证文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/MRS_XXX/install/FusionInsight-Flume-Flume组件版本号/flume/conf/flume.keytab”。

其中“XXX”为产品版本号，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤5 将此节点上的配置文件“jaas.conf”，复制到安装Flume客户端的节点，保存到Flume客户端中“conf”目录。

文件完整路径为“${BIGDATA_HOME}/MRS_Current/1_X_Flume/etc/jaas.conf”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤6 登录安装Flume客户端节点，切换到客户端安装目录，执行以下命令修改文件：

vi conf/jaas.conf

修改参数“keyTab”定义的用户认证文件完整路径即**步骤4**中保存用户认证文件的目录：“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”，然后保存并退出。

步骤7 执行以下命令，修改Flume客户端配置文件“flume-env.sh”：

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/flume-env.sh**

在“-XX:+UseCMSCompactAtFullCollection”后面，增加以下内容：

```
-Djava.security.krb5.conf=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/kdc.conf -
Djava.security.auth.login.config=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/jaas.conf -
Dzookeeper.request.timeout=120000
```

例如：**"-XX:+UseCMSCompactAtFullCollection -Djava.security.krb5.conf=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号*/**conf/kdc.conf -Djava.security.auth.login.config=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号*/**conf/jaas.conf -Dzookeeper.request.timeout=120000"**

请根据实际情况，修改“Flume客户端安装目录”，然后保存并退出。

步骤8 假设Flume客户端安装路径为“/opt/FlumeClient”，执行以下命令，重启Flume客户端：

cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/**bin**

./flume-manage.sh restart

步骤9 执行以下命令，修改Flume客户端配置文件“properties.properties”。

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号*/**conf/properties.properties**

将以下内容保存到文件中：

```
################################################################################################
client.sources = static_log_source
client.channels = static_log_channel
client.sinks = kafka_sink
################################################################################################
#LOG_TO_HDFS_ONLINE_1

client.sources.static_log_source.type = spooldir
client.sources.static_log_source.spoolDir = PATH
client.sources.static_log_source.fileSuffix = .COMPLETED
client.sources.static_log_source.ignorePattern = ^$
client.sources.static_log_source.trackerDir = PATH
client.sources.static_log_source.maxBlobLength = 16384
client.sources.static_log_source.batchSize = 51200
client.sources.static_log_source.inputCharset = UTF-8
client.sources.static_log_source.deserializer = LINE
client.sources.static_log_source.selector.type = replicating
client.sources.static_log_source.fileHeaderKey = file
client.sources.static_log_source.fileHeader = false
client.sources.static_log_source.basenameHeader = true
client.sources.static_log_source.basenameHeaderKey = basename
client.sources.static_log_source.deletePolicy = never

client.channels.static_log_channel.type = file
client.channels.static_log_channel.dataDirs = PATH
client.channels.static_log_channel.checkpointDir = PATH
client.channels.static_log_channel.maxFileSize = 2146435071
client.channels.static_log_channel.capacity = 1000000
client.channels.static_log_channel.transactionCapacity = 612000
client.channels.static_log_channel.minimumRequiredSpace = 524288000

client.sinks.kafka_sink.type = org.apache.flume.sink.kafka.KafkaSink
client.sinks.kafka_sink.kafka.topic = flume_test
client.sinks.kafka_sink.kafka.bootstrap.servers = XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口号
client.sinks.kafka_sink.flumeBatchSize = 1000
client.sinks.kafka_sink.kafka.producer.type = sync
client.sinks.kafka_sink.kafka.security.protocol  = SASL_PLAINTEXT
client.sinks.kafka_sink.kafka.kerberos.domain.name = hadoop.XXX.com
client.sinks.kafka_sink.requiredAcks = 0
```

```
client.sources.static_log_source.channels = static_log_channel
client.sinks.kafka_sink.channel = static_log_channel
```

请根据实际情况，修改以下参数，然后保存并退出。

- spoolDir
- trackerDir
- dataDirs
- checkpointDir
- topic

  如果kafka中该topic不存在，默认情况下会自动创建该topic。
- kafka.bootstrap.servers

  默认情况下，安全集群对应端口21007，普通集群对应端口9092。
- kafka.security.protocol

  安全集群请配置为SASL_PLAINTEXT，普通集群请配置为PLAINTEXT。
- “kafka.kerberos.domain.name”

  普通集群无需配置此参数。安全集群对应此参数的值为Kafka集群中“kerberos.domain.name”对应的值。

  具体可到Broker实例所在节点上查看“${BIGDATA_HOME}/MRS_Current/1_X_Broker/etc/server.properties”。

  其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤10 Flume客户端将自动加载“properties.properties”的内容。

当“spoolDir”生成新的日志文件，文件内容将发送到Kafka生产者，并支持Kafka消费者消费。

----结束

使用 Flume 客户端（MRS 3.x 及之后版本）

说明

普通集群不需要执行**步骤2**-**步骤6**。

步骤1 客户端安装。

步骤2 将Master1节点上的认证服务器配置文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_Current/1_X_KerberosClient/etc/kdc.conf”。其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤3 查看任一部署Flume角色节点的“业务IP”。

登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Flume > 实例”。查看任一部署Flume角色节点的“业务IP”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤4 将此节点上的用户认证文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_Porter_XXX/install/FusionInsight-Flume-Flume组件版本号/flume/conf/flume.keytab”。

其中“XXX”为产品版本号，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤5 将此节点上的配置文件“jaas.conf”，复制到安装Flume客户端的节点，保存到Flume客户端中“conf”目录。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_Current/1_X_Flume/etc/jaas.conf”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤6 登录安装Flume客户端节点，切换到客户端安装目录，执行以下命令修改文件：

vi conf/jaas.conf

修改参数“keyTab”定义的用户认证文件完整路径即**步骤4**中保存用户认证文件的目录：“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”，然后保存并退出。

步骤7 执行以下命令，修改Flume客户端配置文件“flume-env.sh”：

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/flume-env.sh**

在“-XX:+UseCMSCompactAtFullCollection”后面，增加以下内容：

```
-Djava.security.krb5.conf=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/kdc.conf -Djava.security.auth.login.config=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/jaas.conf -Dzookeeper.request.timeout=120000
```

例如：**"-XX:+UseCMSCompactAtFullCollection -Djava.security.krb5.conf=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/kdc.conf -Djava.security.auth.login.config=***Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/jaas.conf -Dzookeeper.request.timeout=120000"**

请根据实际情况，修改“Flume客户端安装目录”，然后保存并退出。

步骤8 假设Flume客户端安装路径为“/opt/FlumeClient”，执行以下命令，重启Flume客户端：

cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号***/bin**

./flume-manage.sh restart

步骤9 执行以下命令，修改Flume客户端配置文件“properties.properties”。

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/properties.properties**

将以下内容保存到文件中：

```
################################################################################
################
client.sources = static_log_source
client.channels = static_log_channel
client.sinks = kafka_sink
################################################################################
################
#LOG_TO_HDFS_ONLINE_1

client.sources.static_log_source.type = spooldir
client.sources.static_log_source.spoolDir = PATH
client.sources.static_log_source.fileSuffix = .COMPLETED
client.sources.static_log_source.ignorePattern = ^$
client.sources.static_log_source.trackerDir = PATH
client.sources.static_log_source.maxBlobLength = 16384
client.sources.static_log_source.batchSize = 51200
client.sources.static_log_source.inputCharset = UTF-8
client.sources.static_log_source.deserializer = LINE
client.sources.static_log_source.selector.type = replicating
client.sources.static_log_source.fileHeaderKey = file
client.sources.static_log_source.fileHeader = false
client.sources.static_log_source.basenameHeader = true
client.sources.static_log_source.basenameHeaderKey = basename
client.sources.static_log_source.deletePolicy = never

client.channels.static_log_channel.type = file
client.channels.static_log_channel.dataDirs = PATH
client.channels.static_log_channel.checkpointDir = PATH
client.channels.static_log_channel.maxFileSize = 2146435071
client.channels.static_log_channel.capacity = 1000000
client.channels.static_log_channel.transactionCapacity = 612000
client.channels.static_log_channel.minimumRequiredSpace = 524288000

client.sinks.kafka_sink.type = org.apache.flume.sink.kafka.KafkaSink
client.sinks.kafka_sink.kafka.topic = flume_test
client.sinks.kafka_sink.kafka.bootstrap.servers = XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口
号,XXX.XXX.XXX.XXX:kafka端口号
client.sinks.kafka_sink.flumeBatchSize = 1000
client.sinks.kafka_sink.kafka.producer.type = sync
client.sinks.kafka_sink.kafka.security.protocol  = SASL_PLAINTEXT
client.sinks.kafka_sink.kafka.kerberos.domain.name = hadoop.XXX.com
client.sinks.kafka_sink.requiredAcks = 0

client.sources.static_log_source.channels = static_log_channel
client.sinks.kafka_sink.channel = static_log_channel
```

请根据实际情况，修改以下参数，然后保存并退出。

- spoolDir
- trackerDir
- dataDirs
- checkpointDir
- topic
  如果kafka中该topic不存在，默认情况下会自动创建该topic。
- kafka.bootstrap.servers
  默认情况下，安全集群对应端口21007，普通集群对应端口9092。
- kafka.security.protocol
  安全集群请配置为SASL_PLAINTEXT，普通集群请配置为PLAINTEXT。
- “kafka.kerberos.domain.name”
  普通集群无需配置此参数。安全集群对应此参数的值为Kafka集群中“kerberos.domain.name”对应的值。

具体可到Broker实例所在节点上查看“${BIGDATA_HOME}/FusionInsight_Current/1_X_Broker/etc/server.properties”。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤10 Flume客户端将自动加载“properties.properties”的内容。

当“spoolDir”生成新的日志文件，文件内容将发送到Kafka生产者，并支持Kafka消费者消费。

----结束

12.7.2 使用简介

Flume是一个分布式、可靠和高可用的海量日志聚合的系统。它能够将不同数据源的海量日志数据进行高效收集、聚合、移动，最后存储到一个中心化数据存储系统中。支持在系统中定制各类数据发送方，用于收集数据。同时，提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

Flume分为客户端和服务端，两者都是FlumeAgent。服务端对应着FlumeServer实例，直接部署在集群内部。而客户端部署更灵活，可以部署在集群内部，也可以部署在集群外。它们之间没有必然联系，都可以独立工作，并且提供的功能是一样的。

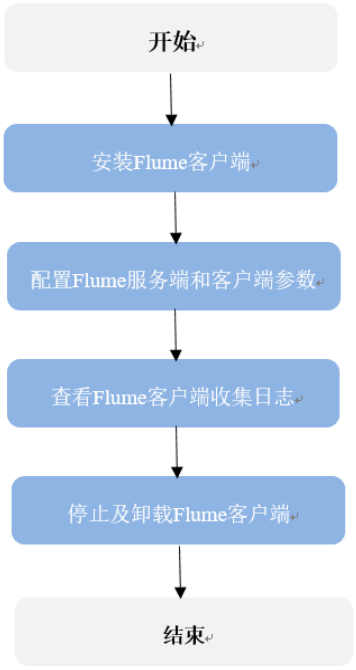
Flume客户端需要单独安装，支持将数据直接导到集群中的HDFS和Kafka等组件上，也可以结合Flume服务端一起使用。

使用流程

通过Flume采集日志的流程如下所示。

1. 安装Flume客户端。
2. 配置Flume服务端和客户端参数。
3. 查看Flume客户端收集日志。
4. 停止及卸载Flume客户端。

图 12-13 Flume 使用流程



Flume 客户端介绍

Flume客户端由Source、Channel、Sink组成，数据先进入Source然后传递到Channel，最后由Sink发送到客户端外部。各模块说明见表12-105。

表 12-105 模块说明

名称	说明
Source	Source负责接收数据或产生数据，并将数据批量放到一个或多个Channel。Source有两种类型：数据驱动和轮询。 典型的Source样例如下： • 和系统集成并接收数据的Sources：Syslog、Netcat。 • 自动生成事件数据的Sources：Exec、SEQ。 • 用于Agent和Agent之间通信的IPC Sources：Avro。 Source必须至少和一个Channel关联。
Channel	Channel位于Source和Sink之间，用于缓存Source传递的数据，当Sink成功将数据发送到下一跳的Channel或最终数据处理端，缓存数据将自动从Channel移除。 不同类型的Channel提供的持久化水平也是不一样的： • Memory Channel：非持久化 • File Channel：基于预写式日志（Write-Ahead Logging，简称WAL）的持久化实现 • JDBC Channel：基于嵌入Database的持久化实现 Channel支持事务特性，可保证简易的顺序操作，同时可以配合任意数量的Source和Sink共同工作。

名称	说明
Sink	Sink负责将数据传输到下一跳或最终目的，成功完成后将数据从Channel移除。 典型的Sink样例如下： ● 存储数据到最终目的终端Sink，比如：HDFS、Kafka ● 自动消耗的Sinks，比如：Null Sink ● 用于Agent和Agent之间通信的IPC sink：Avro Sink必须关联到一个Channel。

Flume客户端可以配置成多个Source、Channel、Sink，即一个Source将数据发送给多个Channel，再由多个Sink发送到客户端外部。

Flume还支持多个Flume客户端配置级联，即Sink将数据再发送给Source。

补充说明

1. Flume可靠性保障措施。
   - Source与Channel、Channel与Sink之间支持事务机制。
   - Sink Processor支持配置failover、load_balance机制。

     例如load_balance示例如下：

     ```
     server.sinkgroups=g1
     server.sinkgroups.g1.sinks=k1 k2
     server.sinkgroups.g1.processor.type=load_balance
     server.sinkgroups.g1.processor.backoff=true
     server.sinkgroups.g1.processor.selector=random
     ```
2. Flume多客户端聚合级联时的注意事项。
   - 级联时需要走Avro或者Thrift协议进行级联。
   - 聚合端存在多个节点时，连接配置尽量配置均衡，不要聚合到单节点上。
3. Flume客户端可以包含多个独立的数据流，即在一个配置文件properties.properties中配置多个Source、Channel、Sink。这些组件可以链接以形成多个流。

   例如在一个配置中配置两个数据流，示例如下：

   ```
   server.sources = source1 source2
   server.sinks = sink1 sink2
   server.channels = channel1 channel2

   #dataflow1
   server.sources.source1.channels = channel1
   server.sinks.sink1.channel = channel1

   #dataflow2
   server.sources.source2.channels = channel2
   server.sinks.sink2.channel = channel2
   ```

12.7.3 安装 Flume 客户端

12.7.3.1 安装 MRS 3.x 之前版本 Flume 客户端

操作场景

使用Flume搜集日志时，需要在日志主机上安装Flume客户端。用户可以创建一个新的ECS并安装Flume客户端。

本章节适用于MRS 3.x之前版本。

前提条件

- 已创建包含Flume组件的流集群。
- 日志主机需要与MRS集群在相同的VPC和子网。
- 已获取日志主机的登录方式。

操作步骤

步骤1 根据前提条件，创建一个满足要求的弹性云服务器。

步骤2 登录集群详情页面，选择“组件管理”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤3 单击“下载客户端”。

1. 在“客户端类型”选择“完整客户端”。
2. 在“下载路径”选择“远端主机”。
3. 将“主机IP”设置为ECS的IP地址，设置“主机端口”为“22”，并将“保存路径”设置为“/tmp”。
   - 如果使用SSH登录ECS的默认端口“22”被修改，请将“主机端口”设置为新端口。
   - “保存路径”最多可以包含256个字符。
4. “登录用户”设置为“root”。

   如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。
5. 在“登录方式”选择“密码”或“SSH私钥”。
   - 密码：输入创建集群时设置的root用户密码。
   - SSH私钥：选择并上传创建集群时使用的密钥文件。
6. 单击“确定”开始生成客户端文件。

   若界面显示以下提示信息表示客户端包已经成功保存。

   ```
   下载客户端文件到远端主机成功。
   ```

   若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从**步骤3**执行重新开始下载客户端。

   ```
   连接到服务器失败，请检查网络连接或参数设置。
   ```

步骤4 选择“Flume”服务，单击“实例”，查看任意一个Flume实例和两个MonitorServer实例的“业务IP”。

步骤5 使用VNC方式，登录弹性云服务器。参见弹性云服务器《用户指南》的**远程登录（VNC方式）**章节（“实例 > 登录Linux弹性云服务器 > 远程登录（VNC方式）”）。

所有镜像均支持Cloud-init特性。Cloud-init预配置的用户名“root”，密码为创建集群时设置的密码。首次登录建议修改。

步骤6 在弹性云服务器，切换到**root**用户，并将安装包复制到目录“/opt”。

sudo su - root

cp /tmp/MRS_Flume_Client.tar /opt

步骤7 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Flume_Client.tar

步骤8 执行以下命令，校验文件包。

sha256sum -c MRS_Flume_ClientConfig.tar.sha256

界面显示如下信息，表明文件包校验成功：

```
MRS_Flume_ClientConfig.tar: OK
```

步骤9 执行以下命令，解压“MRS_Flume_ClientConfig.tar”。

tar -xvf MRS_Flume_ClientConfig.tar

步骤10 执行以下命令，安装客户端运行环境到新的目录，例如“/opt/Flumeenv”。安装时自动生成目录。

sh /opt/MRS_Flume_ClientConfig/install.sh /opt/Flumeenv

查看安装输出信息，如有以下结果表示客户端运行环境安装成功：

```
Components client installation is complete.
```

步骤11 执行以下命令，配置环境变量。

source /opt/Flumeenv/bigdata_env

步骤12 执行以下命令，解压Flume客户端。

cd /opt/MRS_Flume_ClientConfig/Flume

tar -xvf FusionInsight-Flume-1.6.0.tar.gz

步骤13 执行以下命令，查看当前用户密码是否过期。

chage -l root

“Password expires”时间早于当前则表示过期。此时需要修改密码，或执行**chage -M -1 root**设置密码为未过期状态。

步骤14 执行以下命令，安装Flume客户端到新目录，例如“/opt/FlumeClient”。安装时自动生成目录。

sh /opt/MRS_Flume_ClientConfig/Flume/install.sh -d /opt/FlumeClient -f *MonitorServer实例的业务IP地址* -c Flume配置文件路径 -l /var/log/ -e *Flume的业务IP地址* -n *Flume客户端名称*

各参数说明如下：

- “-d”：表示Flume客户端安装路径。
- “-f”：可选参数，表示两个MonitorServer角色的业务IP地址，中间用英文逗号分隔，若不设置则Flume客户端将不向MonitorServer发送告警信息，同时在MRS Manager界面上看不到该客户端的相关信息。
- “-c”：可选参数，表示Flume客户端在安装后默认加载的配置文件“properties.properties”。如不添加参数，默认使用客户端安装目录的“fusioninsight-flume-1.6.0/conf/properties.properties”。客户端中配置文件为空白模板，根据业务需要修改后Flume客户端将自动加载。
- “-l”：可选参数，表示日志目录，默认值为“/var/log/Bigdata”。
- “-e”：可选参数，表示Flume实例的业务IP地址，主要用于接收客户端上报的监控指标信息。
- “-n”：可选参数，表示自定义的Flume客户端的名称。
- IBM的JDK不支持“-Xloggc”，需要修改“flume/conf/flume-env.sh”，将“-Xloggc”修改为“-Xverbosegclog”，若JDK为32位，“-Xmx”不能大于3.25GB。
- “flume/conf/flume-env.sh”中，“-Xmx”默认为4GB。若客户端机器内存过小，可调整为512M甚至1GB。

例如执行：**sh install.sh -d /opt/FlumeClient**

系统显示以下结果表示客户端运行环境安装成功：

```
install flume client successfully.
```

----结束

12.7.3.2 安装 MRS 3.x 及之后版本 Flume 客户端

操作场景

使用Flume搜集日志时，需要在日志主机上安装Flume客户端。用户可以创建一个新的ECS并安装Flume客户端。

本章节适用于MRS 3.x及之后版本。

前提条件

- 已创建包含Flume组件的集群。
- 日志主机需要与MRS集群在相同的VPC和子网。
- 已获取日志主机的登录方式。
- 安装目录可以不存在，会自动创建。但如果存在，则必须为空。目录路径不能包含空格。

操作步骤

步骤1 获取软件包。

登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”进入Flume服务界面，在右上角选择“更多 > 下载客户端”，选择“选择客户端类型”为“完整客户端”，下载Flume服务客户端文件。

客户端文件名称为“FusionInsight_Cluster_*<集群ID>*_Flume_Client.tar”，本章节以“FusionInsight_Cluster_1_Flume_Client.tar”为例进行描述。

步骤2 上传软件包。

以**user**用户将软件包上传到将要安装Flume服务客户端的节点目录上，例如“/opt/client”。

说明

user用户为安装和运行Flume客户端的用户。

步骤3 解压软件包。

以**user**用户登录将要安装Flume服务客户端的节点。进入安装包所在目录，例如“/opt/client”，执行如下命令解压安装包到当前目录。

cd /opt/client

tar -xvf FusionInsight_Cluster_1_Flume_Client.tar

步骤4 校验软件包。

执行**sha256sum -c**命令校验解压得到的文件，返回“OK”表示校验通过。例如：

sha256sum -c FusionInsight_Cluster_1_Flume_ClientConfig.tar.sha256

```
FusionInsight_Cluster_1_Flume_ClientConfig.tar: OK
```

步骤5 解压文件。

tar -xvf FusionInsight_Cluster_1_Flume_ClientConfig.tar

步骤6 在Flume客户端安装目录下执行以下命令，安装客户端到指定目录（绝对路径），例如安装到“/opt/FlumeClient”目录。客户端安装成功后安装结束。

cd /opt/client/FusionInsight_Cluster_1_Flume_ClientConfig/Flume/FlumeClient

./install.sh -d */opt/FlumeClient* **-f** *MonitorServer角色的业务IP或主机名* **-c** *用户业务配置文件properties.properties放置路径* **-s** *cpu阈值* **-l /var/log/Bigdata -e** *FlumeServer的业务IP或主机名* **-n** *Flume*

说明

- “-d”：Flume客户端安装路径。
- “-f”（可选）：两个MonitorServer角色的业务IP或主机名，中间用逗号分隔，若不设置则Flume客户端将不向MonitorServer发送告警信息，同时在FusionInsight Manager界面上看不到该客户端的相关信息。
- “-c”（可选）：指定业务配置文件，该文件需要用户根据自己业务生成，具体操作可在Flume服务端中“配置工具”页面参考**Flume业务配置指南**章节生成，并上传到待安装客户端节点上的任一目录下。若安装时未指定（即不配置该参数），可在安装后上传已经生成的业务配置文件properties.properties到“/opt/FlumeClient/fusioninsight-flume-1.9.0/conf”目录下。
- “-s”（可选）：Cgroup阈值，阈值取值范围为1~100*N之间的整数，N表示机器cpu核数。默认阈值为“-1”，表示加入到Cgroup的进程不受cpu使用率限制。
- “-l”（可选）：日志路径，默认值为“/var/log/Bigdata”（“user”用户需要对此目录有写权限）。首次安装客户端会生成名为flume-client的子目录，之后安装会依次生成名为“flume-client-*n*”的子目录，n代表一个序号，从1依次递增。在Flume客户端安装目录下的conf目录中，编辑ENV_VARS文件，搜索FLUME_LOG_DIR属性，可查看客户端日志路径。
- “-e”（可选）：FlumeServer的业务IP地址或主机名，主要用于接收客户端上报的监控指标信息。
- “-n”（可选）：Flume客户端的名称，可以通过在FusionInsight Manager上选择“集群 > *待操作集群名称* > 服务 > Flume > Flume管理”查看对应节点上客户端的名称。
- 若产生以下错误提示，可执行命令**export JAVA_HOME=***JDK路径*进行处理。
  JAVA_HOME is null in current user,please install the JDK and set the JAVA_HOME
- IBM的JDK不支持“-Xloggc”，需要修改“flume/conf/flume-env.sh”，将“-Xloggc”修改为“-Xverbosegclog”，若JDK为32位，“-Xmx”不能大于3.25GB。
- 集群混搭时，安装跨平台客户端时，请进入/opt/client/FusionInsight_Cluster_1_Flume_ClientConfig/Flume/FusionInsight-Flume-1.9.0.tar.gz路径下进行Flume客户端安装。

----结束

12.7.4 查看 Flume 客户端日志

操作场景

查看日志以便定位问题。

前提条件

Flume客户端已经正确安装。

操作步骤

步骤1 进入Flume客户端日志目录，默认为“/var/log/Bigdata”。

步骤2 执行如下命令查看日志文件列表。

ls -lR flume-client-*

日志文件示例如下：

```
flume-client-1/flume:
total 7672
-rw-------. 1 root root       0 Sep  8 19:43 Flume-audit.log
-rw-------. 1 root root 1562037 Sep 11 06:05 FlumeClient.2017-09-11_04-05-09.[1].log.zip
-rw-------. 1 root root 6127274 Sep 11 14:47 FlumeClient.log
```

```
-rw-------. 1 root root    2935 Sep  8 22:20 flume-root-20170908202009-pid72456-gc.log.0.current
-rw-------. 1 root root    2935 Sep  8 22:27 flume-root-20170908202634-pid78789-gc.log.0.current
-rw-------. 1 root root    4382 Sep  8 22:47 flume-root-20170908203137-pid84925-gc.log.0.current
-rw-------. 1 root root    4390 Sep  8 23:46 flume-root-20170908204918-pid103920-gc.log.0.current
-rw-------. 1 root root    3196 Sep  9 10:12 flume-root-20170908215351-pid44372-gc.log.0.current
-rw-------. 1 root root    2935 Sep  9 10:13 flume-root-20170909101233-pid55119-gc.log.0.current
-rw-------. 1 root root    6441 Sep  9 11:10 flume-root-20170909101631-pid59301-gc.log.0.current
-rw-------. 1 root root       0 Sep  9 11:10 flume-root-20170909111009-pid119477-gc.log.0.current
-rw-------. 1 root root   92896 Sep 11 13:24 flume-root-20170909111126-pid120689-gc.log.0.current
-rw-------. 1 root root    5588 Sep 11 14:46 flume-root-20170911132445-pid42259-gc.log.0.current
-rw-------. 1 root root    2576 Sep 11 13:24 prestartDetail.log
-rw-------. 1 root root    3303 Sep 11 13:24 startDetail.log
-rw-------. 1 root root    1253 Sep 11 13:24 stopDetail.log

flume-client-1/monitor:
total 8
-rw-------. 1 root root  141 Sep  8 19:43 flumeMonitorChecker.log
-rw-------. 1 root root 2946 Sep 11 13:24 flumeMonitor.log
```

其中**FlumeClient.log**即为Flume客户端的运行日志。

----结束

12.7.5 停止或卸载 Flume 客户端

操作场景

指导运维工程师停止、启动Flume客户端，以及在不需要Flume数据采集通道时，卸载Flume客户端。

操作步骤

- 停止Flume角色的客户端。

  假设Flume客户端安装路径为“/opt/FlumeClient”，执行以下命令，停止Flume客户端：

  cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/bin

  ./flume-manage.sh stop

  执行脚本后，显示如下信息，说明成功的停止了Flume客户端：

  ```
  Stop Flume PID=120689 successful..
  ```

  说明

  Flume客户端停止后会自动重启，如果不需自动重启，请执行以下命令：

  ./flume-manage.sh stop force

  需要启动时，可执行以下命令：

  ./flume-manage.sh start force

- 卸载Flume角色的客户端。

  假设Flume客户端安装路径为“/opt/FlumeClient”，执行以下命令，卸载Flume客户端：

  cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/inst

  ./uninstall.sh

12.7.6 使用 Flume 客户端加密工具

操作场景

安装Flume客户端后，配置文件的部分参数可能需要填写加密的字符，Flume客户端中提供了加密工具。

前提条件

已完成客户端安装。

操作步骤

步骤1 登录安装Flume客户端的节点，并切换到客户端安装目录。例如“/opt/FlumeClient”。

步骤2 切换到以下目录

cd fusioninsight-flume-*Flume组件版本号*/bin

步骤3 执行以下命令，加密原始信息：

./genPwFile.sh

输入两次待加密信息。

步骤4 执行以下命令，查看加密后的信息：

cat password.property

说明

如果加密参数是用于Flume Server，那么需要到相应的Flume Server所在节点执行加密。需要使用omm用户执行加密脚本进行加密。

- 针对MRS 3.x之前版本加密路径为“/opt/Bigdata/MRS_XXX/install/FusionInsight-Flume-*Flume组件版本号*/flume/bin/genPwFile.sh”。
- 针对MRS 3.x及之后版本加密路径为“/opt/Bigdata/FusionInsight_Porter_XXX/install/FusionInsight-Flume-*Flume组件版本号*/flume/bin/genPwFile.sh”。其中XXX为产品的版本号。

----结束

12.7.7 Flume 业务配置指南

本章节适用于MRS 3.x及之后版本。

该操作指导用户完成Flume常用业务的配置。其他一些不太常用的Source、Channel、Sink的配置请参考Flume社区提供的用户手册（**http://flume.apache.org/releases/1.9.0.html**）。

说明

- 各个表格中所示参数，黑体加粗的参数为必选参数。
- Sink的BatchSize参数必须小于Channel的transactionCapacity。
- 集群Flume配置工具界面篇幅有限，Source、Channel、Sink只展示部分参数，详细请参考如下常用配置。
- 集群Flume配置工具界面上所展示Customer Source、Customer Channel及Customer Sink需要用户根据自己开发的代码来进行配置，下述常用配置不再展示。

常用 Source 配置

- **Avro Source**

  Avro Source监听Avro端口，接收外部Avro客户端数据并放入配置的Channel中。常用配置如下表所示：

  表 12-106 Avro Source 常用配置

参数	默认值	描述
channels	-	与之相连的channel，可以配置多个。
type	avro	avro source的类型，必须为avro。
bind	-	监听主机名/IP。
port	-	绑定监听端口，该端口需未被占用。
threads	-	source工作的最大线程数。
compression-type	none	消息压缩格式：“none”或“deflate”。“none”表示不压缩，“deflate”表示压缩。
compression-level	6	数据压缩级别（1-9），数值越高，压缩率越高。
ssl	false	是否使用SSL加密。设置为true时还必须指定“密钥(keystore)”和“密钥存储密码(keystore-password)”。

参数	默认值	描述
truststore-type	JKS	Java信任库类型，“JKS”或“PKCS12”。 **说明** JKS的密钥库和私钥采用不同的密码进行保护，而PKCS12的密钥库和私钥采用相同密码进行保护。
truststore	-	Java信任库文件。
truststore-password	-	Java信任库密码。
keystore-type	JKS	ssl启用后密钥存储类型，“JKS”或“PKCS12”。 **说明** JKS的密钥库和私钥用不同的密码进行保护，而PKCS12的密钥库和私钥用相同密码进行保护。
keystore	-	ssl启用后密钥存储文件路径，开启ssl后，该参数必填。
keystore-password	-	ssl启用后密钥存储密码，开启ssl后，该参数必填。
trust-all-certs	false	是否关闭SSL server证书检查。设置为“true”时将不会检查远端source的SSL server证书，不建议在生产中使用。
exclude-protocols	SSLv3	排除的协议列表，用空格分开。默认排除SSLv3协议。
ipFilter	false	是否开启ip过滤。
ipFilter.rules	-	定义N网络的ipFilters，多个主机或IP地址用逗号分割。ipFilter设置为“true”时，配置规则有允许和禁止两种，配置格式如下： ipFilterRules=allow:ip:127.*, allow:name:localhost, deny:ip:*

- **SpoolDir Source**

  Spool Dir Source监控并传输目录下新增的文件，可实现实时数据传输。常用配置如下表所示：

  表 12-107 Spooling Directory Source 常用配置

参数	默认值	描述
channels	-	与之相连的channel，可以配置多个。
type	spooldir	spooling source的类型，必须设置为spooldir。
spoolDir	-	Spooldir source的监控目录，flume运行用户需要对该目录具有可读可写可执行权限。
monTime	0（不开启）	线程监控阈值，更新时间超过阈值后，重新启动该Source，单位：秒。
fileSuffix	.COMPLETED	文件传输完成后添加的后缀。
deletePolicy	never	文件传输完成后源文件删除策略，never或immediate。“never”表示不删除已完成传输的源文件，“immediate”表示传输完成后立刻删除源文件。
ignorePattern	^$	忽略文件的正则表达式表示。默认为“^$”，表示忽略空格。
includePattern	^.*$	包含文件的正则表达式表示。可以与ignorePattern同时使用，如果一个文件既满足ignorePattern也满足includePattern,则该文件会被忽略。另外，以“.”开头的文件不会被过滤。
trackerDir	.flumespool	传输过程中元数据存储路径。
batchSize	1000	批次写入Channel的Event数量。
decodeErrorPolicy	FAIL	编码错误策略。 **说明** 如果文件中有编码错误，请配置“decodeErrorPolicy”为“REPLACE”或“IGNORE”，Flume遇到编码错误将跳过编码错误，继续采集后续日志。
deserializer	LINE	文件解析器，值为“LINE”或“BufferedLine”。 ● 配置为“LINE”时，对从文件读取的字符逐个转码。 ● 配置为“BufferedLine”时，对文件读取的一行或多行的字符进行批量转码，性能更优。

参数	默认值	描述
deserializer.maxLineLength	2048	按行解析最大长度。
deserializer.maxBatchLine	1	按行解析最多行数，如果行数设置为多行，maxLineLength也应该设置为相应的倍数。 **说明** 用户设置Interceptor时，需要考虑多行合并后的场景，否则会造成数据丢失。如果Interceptor无法处理多行合并场景，请将该配置设置为1。
selector.type	replicating	选择器类型，“replicating”或“multiplexing”。“replicating”表示将数据复制多份，分别传递给每一个channel，每个channel接收到的数据都是相同的，而“multiplexing”表示根据event中header的value来选择特定的channel，每个channel中的数据是不同的。
interceptors	-	拦截器。多个拦截器用空格分开。
inputCharset	UTF-8	读取文件的编码格式。须与读取数据源文件编码格式相同，否则字符解析可能会出错。
fileHeader	false	是否把文件名（包含路径）添加到event的header中。
fileHeaderKey	-	设置header中数据存储结构为<key,value>模式，需要fileHeaderKey与fileHeader配合使用。若fileHeader设置为true，可参考如下示例。 示例：将fileHeaderKey定义为file，当读取到文件名为/root/a.txt的内容时，header中以file=/root/a.txt的形式存在。
basenameHeader	false	是否把文件名（不包含路径）添加到event的header中。
basenameHeaderKey	-	设置header中数据存储结构为<key,value>模式，需要basenameHeaderKey与basenameHeader配合使用。若basenameHeader设置为true，可参考如下示例。 示例：将basenameHeaderKey定义为file，当读取到文件名为a.txt的内容时，header中以file=a.txt的形式存在。
pollDelay	500	轮询监控目录下新文件时的时延。单位：毫秒。
recursiveDirectorySearch	false	是否监控配置的目录下子目录中的新文件。

参数	默认值	描述
consumeOrder	oldest	监控目录下文件的消耗次序。如果配置为oldest或者youngest，会根据监控目录下文件的最后修改时间来决定，当目录下有大量文件时，会消耗较长时间去寻找oldest或者youngest的文件。需要注意的是，如果配置为random，创建比较早的文件有可能长时间未被读取。如果配置为oldest或者youngest，那么进程会需要较多时间来查找最新的或最旧的文件。可选值：random，youngest，oldest。
maxBackoff	4000	当Channel满了以后，尝试再次去写Channel所等待的最大时间。超过这个时间，则会抛出异常。对应的Source会以一个较小的时间开始，然后每尝试一次，该时间数字指数增长直到达到当前指定的值，如果还不能成功写入，则认为失败。时间单位：秒。
emptyFileEvent	true	是否采集空文件信息发送到Sink端，默认值为true，表示将空文件信息发送到Sink端。该参数只对HDFS Sink有效，其他Sink该参数无效。以HDFS Sink为例，当参数为true时，如果spoolDir路径下存在空文件，那么HDFS的hdfs.path路径下就会创建一个同名的空文件。

说明

SpoolDir Source在按行读取过程中会忽略掉每一个event的最后一个换行符，该换行符所占用的数据量指标不会被Flume统计。

- **Kafka Source**

Kafka Source从Kafka的topic中消费数据，可以设置多个Source消费同一个topic的数据，每个Source会消费topic的不同partitions。常用配置如下表所示：

表 12-108 Kafka Source 常用配置

参数	默认值	描述
channels	-	与之相连的channel，可以配置多个。
type	org.apache.flume.source.kafka.KafkaSource	kafka source的类型，必须设置为org.apache.flume.source.kafka.KafkaSource。

参数	默认值	描述
kafka.bootstrap.servers	-	Kafka的bootstrap地址端口列表。如果集群已安装Kafka并且配置已经同步，服务端可以不配置此项，默认值为Kafka集群中所有的broker列表。客户端必须配置该项，多个值用逗号分隔。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。
kafka.topics	-	订阅的Kafka topic列表，用逗号分隔。
kafka.topics.regex	-	符合正则表达式的topic会被订阅，优先级高于“kafka.topics”，如果存在将覆盖“kafka.topics”。
monTime	0（不开启）	线程监控阈值，更新时间超过阈值后，重新启动该Source，单位：秒。
nodatatime	0（不开启）	告警阈值，从Kafka中订阅不到数据的时长超过阈值时发送告警，单位：秒。该参数可在配置文件properties.properties进行设置。
batchSize	1000	批次写入Channel的Event数量。
batchDurationMillis	1000	批次消费topic数据的最大时长，单位：ms。
keepTopicInHeader	false	是否在Event Header中保存topic。设置为true，则Kafka Sink配置的topic将无效。
setTopicHeader	true	当设置为true时，会将“topicHeader”中定义的topic名称存储到Header中。
topicHeader	topic	当setTopicHeader属性设置为true，此参数用于定义存储接收的topic名称。如果与Kafka Sink的topicHeader属性结合使用，应该注意，避免将消息循环发送到同一主题。
useFlumeEventFormat	false	默认情况下，event会以字节的形式从kafka topic传递到event的body体中。设置为true，则会以Flume的Avro二进制格式来读取Event。与KafkaSink或KakfaChannel 中同名的parseAsFlumeEvent参数一起使用时，会保留从数据源产生的任何设定的Header。

参数	默认值	描述
keepPartitionInHeader	false	是否在Event Header中保存partitionID。设置为true，则Kafka Sink将写入对应的Partition。
kafka.consumer.group.id	flume	Kafka消费组ID。多个源或代理中设置相同的ID表示它们是同一个consumer group。
kafka.security.protocol	SASL_PLAINTEXT	Kafka安全协议，普通模式集群下须配置为“PLAINTEXT”。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。
Other Kafka Consumer Properties	-	其他Kafka配置，可以接受任意Kafka支持的消费配置，配置需要加前缀“kafka.”。

- **Taildir Source**

  Taildir Source监控目录下文件的变化并自动读取文件内容，可实现实时数据传输，常用配置如下表所示：

  表 12-109 Taildir Source 常用配置

参数	默认值	描述
channels	-	与之相连的channel，可以配置多个。
type	TAILDIR	taildir source的类型，必须为TAILDIR。
filegroups	-	设置采集文件目录分组名字，分组名字中间使用空格间隔。
filegroups.<filegroupName>.parentDir	-	父目录，需要配置为绝对路径。
filegroups.<filegroupName>.filePattern	-	相对父目录的文件路径，可以包含目录，支持正则表达式，须与父目录联合使用。
positionFile	-	传输过程中元数据存储路径。
headers.<filegroupName>.<headerKey>	-	设置某一个分组采集数据时event中的key-value值。
byteOffsetHeader	false	是否在每一个event头中携带该event在源文件中的位置信息。设置为true，则该信息保存在byteoffset变量中。

参数	默认值	描述
maxBatchCount	Long.MAX_VALUE	控制从一个文件中连续读取的最大批次。如果监控目录会一直读取多个文件，且其中一个文件以非常快的速率在写入，那么其他文件可能会无法处理。因为高速写入的这个文件会陷入无限读取的循环中。这种情况下，应该降低此值。
skipToEnd	false	Flume在重启后是否直接定位到文件最新的位置处读取最新的数据。设置为true，则重启后直接定位到文件最新位置读取最新数据。
idleTimeout	120000	设置读取文件的空闲时间，单位：毫秒，如果在该时间内文件内容没有变更，关闭掉该文件，关闭后如果该文件有数据写入，重新打开并读取数据。
writePosInterval	3000	设置将元数据写入到文件的周期，单位：毫秒。
batchSize	1000	批次写入Channel的Event数量。
monTime	0（不开启）	线程监控阈值，更新时间超过阈值后，重新启动该Source，单位：秒。
fileHeader	false	是否把文件名（包含路径）添加到event的header中。
fileHeaderKey	file	设置header中数据存储结构为<key,value>模式，需要fileHeaderKey与fileHeader配合使用。若fileHeader设置为true，可参考如下示例。 示例：将fileHeaderKey定义为file，当读取到文件名为/root/a.txt的内容时，header中以file=/root/a.txt的形式存在。

- **Http Source**

  Http Source接收外部HTTP客户端发送过来的数据，并放入配置的Channel中，常用配置如下表所示：

  表 12-110 Http Source 常用配置

参数	默认值	描述
channels	-	与之相连的channel，可以配置多个。
type	http	http source的类型，必须为http。
bind	-	监听主机名/IP。

参数	默认值	描述
port	-	绑定监听端口，该端口需未被占用。
handler	org.apache.flume.source.http.JSONHandler	http请求的消息解析方式，支持Json格式解析（org.apache.flume.source.http.JSONHandler）和二进制Blob块解析（org.apache.flume.sink.solr.morphline.BlobHandler）。
handler.*	-	设置handler的参数。
exclude-protocols	SSLv3	排除的协议列表，用空格分开。默认排除SSLv3协议。
include-cipher-suites	-	包含的协议列表，用空格分开。如果设置为空，则默认支持所有协议。
enableSSL	false	http协议是否启用SSL。设置为true时还必须指定“密钥(keystore)”和“密钥存储密码(keystore-password)”。
keystore-type	JKS	Keystore类型，可以为JKS或者PKCS12。
keystore	-	http启用SSL后设置keystore的路径。
keystorePassword	-	http启用SSL后设置keystore的密码。

- **Thrift Source**

  Thrift Source监听thrift端口，接收外部Thrift客户端数据并放入配置的Channel中。常用配置如下表所示：

参数	默认值	描述
channels	-	与之相连的channel，可以配置多个。
type	thrift	thrift source的类型，必须设置为thrift。
bind	-	监听主机名/IP。
port	-	绑定监听端口，该端口需未被占用。
threads	-	允许运行的最大的worker线程数目。
kerberos	false	是否启用Kerberos认证。
agent-keytab	-	服务端使用的keytab文件地址，必须使用机机帐号。建议使用Flume服务安装目录下flume/conf/flume_server.keytab。

参数	默认值	描述
agent-principal	-	服务端使用的安全用户的Principal，必须使用机机帐户。建议使用Flume服务默认用户flume_server/hadoop.*<系统域名>*@*<系统域名>* **说明** “flume_server/hadoop.*<系统域名>*”为用户名，用户的用户名所包含的系统域名所有字母为小写。例如“本端域”参数为“9427068F-6EFA-4833-B43E-60CB641E5B6C.COM”，用户名为“flume_server/hadoop.9427068f-6efa-4833-b43e-60cb641e5b6c.com”。
compression-type	none	消息压缩格式：“none”或“deflate”。“none”表示不压缩，“deflate”表示压缩。
ssl	false	是否使用SSL加密。设置为true时还必须指定“密钥(keystore)”和“密钥存储密码(keystore-password)”。
keystore-type	JKS	SSL启用后密钥存储类型。
keystore	-	SSL启用后密钥存储文件路径，开启SSL后，该参数必填。
keystore-password	-	SSL启用后密钥存储密码，开启ssl后，该参数必填。

常用 Channel 配置

- **Memory Channel**

  Memory Channel使用内存作为缓存区，Events存放在内存队列中。常用配置如下表所示：

  表 12-111 Memory Channel 常用配置

参数	默认值	描述
type	-	memory channel的类型，必须设置为memory。
capacity	10000	缓存在channel中的最大Event数。
transactionCapacity	1000	每次存取的最大Event数。 **说明** • 此参数值需要大于source和sink的batchSize。 • 事务缓存容量必须小于或等于Channel缓存容量。

参数	默认值	描述
channelfullcount	10	channel full次数，达到该次数后发送告警。
keep-alive	3	当事务缓存或Channel缓存满时，Put、Take线程等待时间。单位：秒。
byteCapacity	JVM最大内存的80%	channel中最多能容纳所有event body的总字节数，默认是 JVM最大可用内存（ -Xmx ）的80%，单位：bytes。
byteCapacityBufferPercentage	20	channel中字节容量百分比（ % ）。

- **File Channel**

  File Channel使用本地磁盘作为缓存区，Events存放在设置的dataDirs配置项文件夹中。常用配置如下表所示：

  表 12-112 File Channel 常用配置

参数	默认值	描述
type	-	file channel的类型，必须设置为file。
checkpointDir	${BIGDATA_DATA_HOME}/hadoop/data1~N/flume/checkpoint **说明** 此路径随自定义数据路径变更。	检查点存放路径。
dataDirs	${BIGDATA_DATA_HOME}/hadoop/data1~N/flume/data **说明** 此路径随自定义数据路径变更。	数据缓存路径，设置多个路径可提升性能，中间用逗号分开。
maxFileSize	2146435071	单个缓存文件的最大值，单位：bytes。
minimumRequiredSpace	524288000	缓冲区空闲空间最小值，单位：bytes。
capacity	1000000	缓存在channel中的最大Event数。

参数	默认值	描述
transactionCapacity	10000	每次存取的最大Event数。 **说明** • 此参数值需要大于source和sink的batchSize。 • 事务缓存容量必须小于或等于Channel缓存容量。
channelfullcount	10	channel full次数，达到该次数后发送告警。
useDualCheckpoints	false	是否备份检查点。设置为“true”时，必须设置backupCheckpointDir的参数值。
backupCheckpointDir	-	备份检查点路径。
checkpointInterval	30000	检查点间隔时间，单位：秒。
keep-alive	3	当事务缓存或Channel缓存满时，Put、Take线程等待时间。单位：秒。
use-log-replay-v1	false	是否启用旧的回复逻辑。
use-fast-replay	false	是否使用队列回复。
checkpointOnClose	true	channel关闭时是否创建检查点。

- **Memory File Channel**

  Memory File Channel同时使用内存和本地磁盘作为缓存区，消息可持久化，性能优于File Channel，接近Memory Channel的性能。此Channel目前处于试验阶段，可靠性不够高，不建议在生产环境使用。常用配置如下表所示：

  表 12-113 Memory File Channel 常用配置

参数	默认值	描述
type	org.apache.flume.channel.MemoryFileChannel	memory file channel的类型，必须设置为“org.apache.flume.channel.MemoryFileChannel”。
capacity	50000	Channel缓存容量：缓存在Channel中的最大Event数。

参数	默认值	描述
transactionCapacity	5000	事务缓存容量：一次事务能处理的最大Event数。 **说明** • 此参数值需要大于source和sink的batchSize。 • 事务缓存容量必须小于或等于Channel缓存容量。
subqueueByteCapacity	20971520	每个subqueue最多保存多少byte的Event，单位：byte。 Memory File Channel采用queue和subqueue两级缓存，event保存在subqueue，subqueue保存在queue。 subqueue能保存多少event，由“subqueueCapacity”和“subqueueInterval”两个参数决定，“subqueueCapacity”限制subqueue内的Event总容量，“subqueueInterval”限制subqueue保存Event的时长，只有subqueue达到“subqueueCapacity”或“subqueueInterval”上限时，subqueue内的Event才会发往目的地。 **说明** “subqueueByteCapacity”必须大于一个batchsize内的Event总容量。
subqueueInterval	2000	每个subqueue最多保存一段多长时间的Event，单位：毫秒。
keep-alive	3	当事务缓存或Channel缓存满时，Put、Take线程等待时间。 单位：秒。
dataDir	-	缓存本地文件存储目录。
byteCapacity	JVM最大内存的80%	Channel缓存容量。 单位：bytes。
compression-type	None	消息压缩格式：“none”或“deflate”。“none”表示不压缩，“deflate”表示压缩。
channelfullcount	10	channel full次数，达到该次数后发送告警。

Memory File Channel配置样例：

```
server.channels.c1.type = org.apache.flume.channel.MemoryFileChannel
server.channels.c1.dataDir = /opt/flume/mfdata
server.channels.c1.subqueueByteCapacity = 20971520
server.channels.c1.subqueueInterval=2000
```

```
server.channels.c1.capacity = 500000
server.channels.c1.transactionCapacity = 40000
```

- **Kafka Channel**

  Kafka Channel使用Kafka集群缓存数据，Kafka提供高可用、多副本，以防Flume或Kafka Broker崩溃，Channel中的数据会立即被Sink消费。

  表 12-114 Kafka channel 常用配置

Parameter	Default Value	Description
type	-	kafka channel的类型，必须设置为“org.apache.flume.channel.kafka.KafkaChannel”。
kafka.bootstrap.servers	-	Kafka的bootstrap地址端口列表。 如果集群已安装Kafka并且配置已经同步，则服务端可以不配置此项，默认值为Kafka集群中所有的broker列表。客户端必须配置该项，多个值用逗号分隔。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。
kafka.topic	flume-channel	channel用来缓存数据的topic。
kafka.consumer.group.id	flume	从kafka中获取数据的组标识，此参数不能为空。
parseAsFlumeEvent	true	是否解析为Flume event。
migrateZookeeperOffsets	true	当Kafka没有存储offset时，是否从ZooKeeper中查找，并提交到Kafka。

Parameter	Default Value	Description
kafka.consumer.auto.offset.reset	latest	当没有offset记录时从什么位置消费，可选为“earliest”、“latest”或“none”。 “earliest”表示将offset重置为初始点， “latest”表示将offset置为最新位置点， “none”表示若没有offset则抛出异常。
kafka.producer.security.protocol	SASL_PLAINTEXT	Kafka生产安全协议。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。 **说明** 若该参数没有显示，请单击弹窗左下角的"+"显示全部参数。
kafka.consumer.security.protocol	SASL_PLAINTEXT	同上，但用于消费。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。
pollTimeout	500	consumer调用poll()函数能接受的最大超时时间，单位：毫秒。
ignoreLongMessage	false	是否丢弃超大消息。
messageMaxLength	1000012	Flume写入Kafka的消息的最大长度。

常用 Sink 配置

- **HDFS Sink**

  HDFS Sink将数据写入Hadoop分布式文件系统（HDFS）。常用配置如下表所示：

  表 12-115 HDFS Sink 常用配置

参数	默认值	描述
channel	-	与之相连的channel。

参数	默认值	描述
type	hdfs	hdfs sink的类型，必须设置为hdfs。
hdfs.path	-	HDFS上数据存储路径，必须以“hdfs://hacluster/”开头。
monTime	0（不开启）	线程监控阈值，更新时间超过阈值后，重新启动该Sink，单位：秒。
hdfs.inUseSuffix	.tmp	正在写入的hdfs文件后缀。
hdfs.rollInterval	30	按时间滚动文件，单位：秒。
hdfs.rollSize	1024	按大小滚动文件，单位：bytes。
hdfs.rollCount	10	按Event个数滚动文件。 **说明** 参数“rollInterval”、“rollSize”和“rollCount”可同时配置，三个参数采取优先原则，哪个参数值先满足，优先按照哪个参数进行压缩。
hdfs.idleTimeout	0	自动关闭空闲文件超时时间，单位：秒。
hdfs.batchSize	1000	批次写入HDFS的Event个数。
hdfs.kerberosPrincipal	-	认证HDFS的Kerberos principal，普通模式集群不配置，安全模式集群必须配置。
hdfs.kerberosKeytab	-	认证HDFS的Kerberos keytab，普通模式集群不配置，安全模式集群中，用户必须对jaas.cof文件中的keyTab路径有访问权限。
hdfs.fileCloseByEndEvent	true	收到源文件的最后一个Event时是否关闭hdfs文件。
hdfs.batchCallTimeout	-	批次写入HDFS超时控制时间，单位：毫秒。 当不配置此参数时，对每个Event写入HDFS进行超时控制。当“hdfs.batchSize”大于0时，配置此参数可以提升写入HDFS性能。 **说明** “hdfs.batchCallTimeout”设置多长时间需要考虑“hdfs.batchSize”的大小，“hdfs.batchSize”越大，“hdfs.batchCallTimeout”也要调整更长时间，设置过短时间容易导致写HDFS失败。
serializer.appendNewline	true	将一个Event写入HDFS后是否追加换行符（'\n'），如果追加该换行符，该换行符所占用的数据量指标不会被HDFS Sink统计。

参数	默认值	描述
hdfs.filePrefix	over_%{basename}	数据写入hdfs后文件名的前缀。
hdfs.fileSuffix	-	数据写入hdfs后文件名的后缀。
hdfs.inUsePrefix	-	正在写入的hdfs文件前缀。
hdfs.fileType	DataStream	hdfs文件格式，包括“SequenceFile”、“DataStream”以及“CompressedStream”。 **说明** “SequenceFile”和“DataStream”不压缩输出文件，不能设置参数“codeC”，“CompressedStream”压缩输出文件，必须设置“codeC”参数值配合使用。
hdfs.codeC	-	文件压缩格式，包括gzip、bzip2、lzo、lzop、snappy。
hdfs.maxOpenFiles	5000	最大允许打开的hdfs文件数，当打开的文件数达到该值时，最早打开的文件将会被关闭。
hdfs.writeFormat	Writable	文件写入格式，“Writable”或者“Text”。
hdfs.callTimeout	10000	写入HDFS超时控制时间，单位：毫秒。
hdfs.threadsPoolSize	-	每个HDFS sink用于HDFS io操作的线程数。
hdfs.rollTimerPoolSize	-	每个HDFS sink用于调度定时文件滚动的线程数。
hdfs.round	false	时间戳是否四舍五入。若设置为true，则会影响所有基于时间的转义序列（%t除外）。
hdfs.roundUnit	second	时间戳四舍五入单位，可选为“second”、“minute”或“hour”，分别对应为秒、分钟和小时。
hdfs.useLocalTimeStamp	true	是否启用本地时间戳，建议设置为“true”。
hdfs.closeTries	0	hdfs sink尝试关闭重命名文件的最大次数。默认为0表示sink会一直尝试重命名，直至重命名成功。

参数	默认值	描述
hdfs.retryInterval	180	尝试关闭hdfs文件的时间间隔，单位：秒。 **说明** 每个关闭请求都会有多个RPC往返Namenode，因此设置的太低可能导致Namenode超负荷。如果设置0，如果第一次尝试失败的话，该Sink将不会尝试关闭文件，并且把文件打开，或者用“.tmp”作为扩展名。
hdfs.failcount	10	数据写入hdfs失败的次数。该参数作为sink写入hdfs失败次数的阈值，当超过该阈值后上报数据传输异常告警。

- **Avro Sink**

  Avro Sink把events转化为Avro events并发送到配置的主机的监听端口。常用配置如下表所示：

  表 12-116 Avro Sink 常用配置

参数	默认值	描述
channel	-	与之相连的channel。
type	-	avro sink的类型，必须设置为avro。
hostname	-	绑定的主机名/IP。
port	-	监听端口，该端口需未被占用。
batch-size	1000	批次发送的Event个数。

参数	默认值	描述
client.type	DEFAULT	客户端实例类型，根据所配置的模型实际使用到的通信协议设置。该值可选值包括： ● DEFAULT，返回AvroRPC类型的客户端实例。 ● OTHER，返回NULL。 ● THRIFT，返回Thrift RPC类型的客户端实例。 ● DEFAULT_LOADBALANCING， 返回LoadBalancing RPC客户端实例。 ● DEFAULT_FAILOVER， 返回Failover RPC客户端实例。
ssl	false	是否使用SSL加密。设置为true时还必须指定“密钥(keystore)”和“密钥存储密码(keystore-password)”。
truststore-type	JKS	Java信任库类型，“JKS”或“PKCS12”。 **说明** JKS的密钥库和私钥采用不同的密码进行保护，而PKCS12的密钥库和私钥采用相同密码进行保护。
truststore	-	Java信任库文件。
truststore-password	-	Java信任库密码。
keystore-type	JKS	ssl启用后密钥存储类型。
keystore	-	ssl启用后密钥存储文件路径，开启ssl后，该参数必填。
keystore-password	-	ssl启用后密钥存储密码，开启ssl后，该参数必填。

参数	默认值	描述
connect-timeout	20000	第一次连接的超时时间，单位：毫秒。
request-timeout	20000	第一次请求后一次请求的最大超时时间，单位：毫秒。
reset-connection-interval	0	一次断开连接后，等待多少时间后进行重新连接，单位：秒。默认为0表示不断尝试。
compression-type	none	批数据压缩类型，“none”或“deflate”，“none”表示不压缩，“deflate”表示压缩。该值必须与AvroSource的compression-type匹配。
compression-level	6	批数据压缩级别（1-9），数值越高，压缩率越高。
exclude-protocols	SSLv3	排除的协议列表，用空格分开。默认排除SSLv3协议。

- **HBase Sink**

  HBase Sink将数据写入到HBase中。常用配置如下表所示：

  表 12-117 HBase Sink 常用配置

参数	默认值	描述
channel	-	与之相连的channel。
type	-	hbase sink的类型，必须设置为hbase。
table	-	HBase表名称。
columnFamily	-	HBase列族。
monTime	0（不开启）	线程监控阈值，更新时间超过阈值后，重新启动该Sink，单位：秒。
batchSize	1000	批次写入HBase的Event个数。
kerberosPrincipal	-	认证HBase的Kerberos principal，普通模式集群不配置，安全模式集群必须配置。

参数	默认值	描述
kerberosKeytab	-	认证HBase的Kerberos keytab，普通模式集群不配置，安全模式集群中，flume运行用户必须对jaas.cof文件中的keyTab路径有访问权限。
coalesceIncrements	true	是否在同一个处理批次中，合并对同一个hbase cell多个操作。设置为true有利于提高性能。

- **Kafka Sink**

  Kafka Sink将数据写入到Kafka中。常用配置如下表所示：

  表 12-118 Kafka Sink 常用配置

参数	默认值	描述
channel	-	与之相连的channel。
type	-	kafka sink的类型，必须设置为org.apache.flume.sink.kafka.KafkaSink。
kafka.bootstrap.servers	-	Kafka 的bootstrap 地址端口列表。如果集群安装有kafka并且配置已经同步，服务端可以不配置此项，默认值为Kafka集群中所有的broker列表，客户端必须配置该项，多个用逗号分隔。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。
monTime	0（不开启）	线程监控阈值，更新时间超过阈值后，重新启动该Sink，单位：秒。
kafka.producer.acks	1	必须收到多少个replicas的确认信息才认为写入成功。0表示不需要接收确认信息，1表示只等待leader的确认信息。-1表示等待所有的relicas的确认信息。设置为-1，在某些leader失败的场景中可以避免数据丢失。
kafka.topic	-	数据写入的topic，必须填写。
flumeBatchSize	1000	批次写入Kafka的Event个数。
kafka.security.protocol	SASL_PLAINTEXT	Kafka安全协议，普通模式集群下须配置为“PLAINTEXT”。端口和安全协议的匹配规则必须为：21007匹配安全模式（SASL_PLAINTEXT），9092匹配普通模式（PLAINTEXT）。
ignoreLongMessage	false	是否丢弃超大消息的开关。

参数	默认值	描述
messageMaxLength	1000012	Flume写入Kafka的消息的最大长度。
defaultPartitionId	-	用于指定channel中的events被传输到哪一个Kafka partition ID ，此值会被partitionIdHeader覆盖。默认情况下，如果此参数不设置，会由Kafka Producer's partitioner 进行events分发(可以通过指定key或者kafka.partitioner.class自定义的partitioner)。
partitionIdHeader	-	设置时，对应的Sink 将从Event 的Header中获取使用此属性的值命名的字段的值，并将消息发送到主题的指定分区。 如果该值无对应的有效分区，则会抛出EventDeliveryException。 如果Header 值已经存在，则此设置将覆盖参数defaultPartitionId。
Other Kafka Producer Properties	-	其他Kafka配置，可以接受任意Kafka支持的生产配置，配置需要加前缀 .kafka。

- **Thrift Sink**

  Thrift Sink把events转化为Thrift events并发送到配置的主机的监听端口。常用配置如下表所示：

  表 12-119 Thrift Sink 常用配置

参数	默认值	描述
channel	-	与之相连的channel。
type	thrift	thrift sink的类型，必须设置为thrift。
hostname	-	绑定的主机名/IP。
port	-	监听端口，该端口需未被占用。
batch-size	1000	批次发送的Event个数。
connect-timeout	20000	第一次连接的超时时间，单位：毫秒。
request-timeout	20000	第一次请求后一次请求的最大超时时间，单位：毫秒。
kerberos	false	是否启用Kerberos认证。

参数	默认值	描述
client-keytab	-	客户端使用的keytab文件地址，flume运行用户必须对认证文件具有访问权限。
client-principal	-	客户端使用的安全用户的Principal。
server-principal	-	服务端使用的安全用户的Principal。
compression-type	none	Flume发送数据的压缩类型，“none”或“deflate”，“none”表示不压缩，“deflate”表示压缩。
maxConnections	5	Flume发送数据时的最大连接池大小。
ssl	false	是否使用SSL加密。
truststore-type	JKS	Java信任库类型。
truststore	-	Java信任库文件。
truststore-password	-	Java信任库密码。
reset-connection-interval	0	一次断开连接后，等待多少时间后进行重新连接，单位：秒。默认为0表示不断尝试。

注意事项

- Flume可靠性保障措施有哪些?
  - Source&Channel、Channel&Sink之间的事务机制。
  - Sink Processor支持配置failover、load_blance机制，例如负载均衡示例如下，详细参考**http://flume.apache.org/releases/1.9.0.html**。

    ```
    server.sinkgroups=g1
    server.sinkgroups.g1.sinks=k1 k2
    server.sinkgroups.g1.processor.type=load_balance
    server.sinkgroups.g1.processor.backoff=true
    server.sinkgroups.g1.processor.selector=random
    ```
- Flume多agent聚合级联时的注意事项?
  - 级联时需要使用Avro或者Thrift协议进行级联。
  - 聚合端存在多个节点时，连接配置尽量配置均衡，不要聚合到单节点上。

12.7.8 Flume 配置参数说明

MRS 3.x之前版本需在“properties.properties”文件中配置。

MRS 3.x及之后版本，部分参数可在Manager界面配置。

基本介绍

使用Flume需要配置Source、Channel和Sink，各模块配置参数说明可通过本节内容了解。

MRS 3.x及之后版本部分参数可通过Manager界面配置，选择“集群 > 服务 > Flume > 配置工具”，选择要使用的Source、Channel以及Sink，将其拖到右侧的操作界面中，双击对应的Source、Channel以及Sink，根据实际环境可配置Source、Channel和Sink参数。“channels”、“type”等参数仅在客户端配置文件“properties.properties”中进行配置，配置文件路径为“*Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号*/**conf/properties.properties**”。

说明

部分配置可能需要填写加密后的信息，请参见**使用Flume客户端加密工具**。

常用 Source 配置

- **Avro Source**

  Avro Source监听Avro端口，接收外部Avro客户端数据并放入配置的Channel中。常用配置如**表12-120**所示：

  表 12-120 Avro Source 常用配置

参数	默认值	描述
channels	-	与之相连的Channel，可以配置多个。用空格隔开。 在单个代理流程中，是通过channel连接sources和sinks。一个source实例对应多个channels，但一个sink实例只能对应一个channel。 格式如下： **<Agent >.sources.<Source>.channels = <channel1> <channel2> <channel3>...** **<Agent >.sinks.<Sink>.channels = <channel1>** 仅可在“properties.properties”文件中配置。
type	avro	类型，需设置为“avro”。每一种source的类型都为相应的固定值。 仅可在“properties.properties”文件中配置。
bind	-	绑定和source关联的主机名或IP地址。
port	-	绑定端口号。

参数	默认值	描述
ssl	false	是否使用SSL加密。 • true • false
truststore-type	JKS	Java信任库类型。填写JKS或其他java支持的truststore类型。
truststore	-	Java信任库文件。
truststore-password	-	Java信任库密码。
keystore-type	JKS	密钥存储类型。填写JKS或其他java支持的truststore类型。
keystore	-	密钥存储文件。
keystore-password	-	密钥存储密码。

- **SpoolDir Source**

  SpoolDir Source监控并传输目录下新增的文件，可实现准实时数据传输。常用配置如**表 2 Spooling Source常用配置**所示：

  表 12-121 SpoolDir Source 常用配置

参数	默认值	描述
channels	-	与之相连的Channel，可以配置多个。 仅可在“properties.properties”文件中配置。
type	spooldir	类型，需设置为“spooldir”。 仅可在“properties.properties”文件中配置。
monTime	0（不开启）	线程监控阈值，更新时间大于阈值时会重新启动该Source，单位：秒。
spoolDir	-	监控目录。
fileSuffix	.COMPLETED	文件传输完成后添加的后缀。
deletePolicy	never	文件传输完成后源文件删除策略，支持“never”或“immediate”。分别别是从不删除和立即删除。
ignorePattern	^$	忽略文件的正则表达式表示。
trackerDir	.flumespool	传输过程中元数据存储路径。
batchSize	1000	Source传输粒度。

参数	默认值	描述
decodeErrorPolicy	FAIL	编码错误策略。仅可在“properties.properties”文件中配置。 可选FAIL、REPLACE、IGNORE。 FAIL：抛出异常并让解析失败。 REPLACE：将不能识别的字符用其它字符代替，通常是字符U+FFFD。 IGNORE：直接丢弃不能解析的字符串。 **说明** 如果文件中有编码错误，请配置“decodeErrorPolicy”为“REPLACE”或“IGNORE”，Flume遇到编码错误将跳过编码错误，继续采集后续日志。
deserializer	LINE	文件解析器，值为“LINE”或“BufferedLine”。 ● 配置为“LINE”时，对从文件读取的字符逐个转码。 ● 配置为“BufferedLine”时，对文件读取的一行或多行的字符进行批量转码，性能更优。
deserializer.maxLineLength	2048	按行解析最大长度。0到2,147,483,647。
deserializer.maxBatchLine	1	按行解析最多行数，如果行数设置为多行，“maxLineLength”也应该设置为相应的倍数。例如maxBatchLine设置为2，“maxLineLength”相应的设置为2048*2为4096。
selector.type	replicating	选择器类型，支持“replicating”或“multiplexing”。 ● “replicating”表示同样的内容会发给每一个channel。 ● “multiplexing”表示根据分发规则，有选择地发给某些channel。
interceptors	-	拦截器配置。详细配置可参考**flume官方文档**。 仅可在“properties.properties”文件中配置。

说明

Spooling Source在按行读取过程中，会忽略掉每一个Event的最后一个换行符，该换行符所占用的数据量指标不会被Flume统计。

- **Kafka Source**

  Kafka Source从Kafka的topic中消费数据，可以设置多个Source消费同一个topic的数据，每个Source会消费topic的不同partitions。常用配置如**表 3 Kafka Source常用配置**所示：

  表 12-122 Kafka Source 常用配置

参数	默认值	描述
channels	-	与之相连的Channel，可以配置多个。 仅可在“properties.properties”文件中配置。
type	org.apache.flume.source.kafka.KafkaSource	类型，需设置为“org.apache.flume.source.kafka.KafkaSource”。 仅可在“properties.properties”文件中配置。
monTime	0（不开启）	线程监控阈值，更新时间大于阈值时重新启动该Source，单位：秒。
nodatatime	0（不开启）	告警阈值，从Kafka中订阅不到数据的时长大于阈值时发送告警，单位：秒。
batchSize	1000	每次写入Channel的Event数量。
batchDurationMillis	1000	每次消费topic数据的最大时长，单位：毫秒。
keepTopicInHeader	false	是否在Event Header中保存topic，如果保存，Kafka Sink配置的topic将无效。 ● true ● false 仅可在“properties.properties”文件中配置。
keepPartitionInHeader	false	是否在Event Header中保存partitionID，如果保存，Kafka Sink将写入对应的Partition。 ● true ● false 仅可在“properties.properties”文件中配置。
kafka.bootstrap.servers	-	brokers地址列表，多个地址用英文逗号分隔。
kafka.consumer.group.id	-	Kafka消费者组ID。

参数	默认值	描述
kafka.topics	-	订阅的kafka topic列表，用英文逗号分隔。
kafka.topics.regex	-	符合正则表达式的topic会被订阅，优先级高于“kafka.topics”，如果配置将覆盖“kafka.topics”。
kafka.security.protocol	SASL_PLAINTEXT	Kafka安全协议，未启用Kerberos集群中须配置为“PLAINTEXT”。
kafka.kerberos.domain.name	-	此参数的值为Kafka集群中kerberos的“default_realm”，仅安全集群需要配置。 仅可在“properties.properties”文件中配置。
Other Kafka Consumer Properties	-	其他Kafka配置，可以接受任意Kafka支持的消费参数配置，配置需要加前缀“.kafka”。 仅可在“properties.properties”文件中配置。

- **Taildir Source**

  Taildir Source监控目录下文件的变化并自动读取文件内容，可实现实时数据传输，常用配置如表12-123所示：

  表 12-123 Taildir Source 常用配置

参数	默认值	描述
channels	-	与之相连的Channel，可以配置多个。 仅可在“properties.properties”文件中配置。
type	taildir	类型，需配置为“taildir”。 仅可在“properties.properties”文件中配置。
filegroups	-	设置采集文件目录分组名字，分组名字中间使用空格间隔。
filegroups.<filegroupName>.parentDir	-	父目录，需要配置为绝对路径。 仅可在“properties.properties”文件中配置。
filegroups.<filegroupName>.filePattern	-	相对父目录的文件路径，可以包含目录，支持正则表达式，须与父目录联合使用。 仅可在“properties.properties”文件中配置。

参数	默认值	描述
positionFile	-	传输过程中元数据存储路径。
headers.<filegroupName>.<headerKey>	-	设置某一个分组采集数据时Event中的key-value值。 仅可在“properties.properties”文件中配置。
byteOffsetHeader	false	是否在每一个Event头中携带该Event在源文件中的位置信息，该信息保存在“byteoffset”变量中。
skipToEnd	false	Flume在重启后是否直接定位到文件最新的位置处，以读取最新的数据。
idleTimeout	120000	设置读取文件的空闲时间，单位：毫秒。如果在该时间内文件内容没有变更，关闭掉该文件，关闭后如果该文件有数据写入，重新打开并读取数据。
writePosInterval	3000	设置将元数据写入到文件的周期，单位：毫秒。
batchSize	1000	批次写入Channel的Event数量。
monTime	0（不开启）	线程监控阈值，更新时间大于阈值时重新启动该Source，单位：秒。

- **Http Source**

  Http Source接收外部HTTP客户端发送过来的数据，并放入配置的Channel中，常用配置如表12-124所示：

  表 12-124 Http Source 常用配置

参数	默认值	描述
channels	-	与之相连的Channel，可以配置多个。仅可在“properties.properties”文件中配置。
type	http	类型，需配置为“http”。仅可在“properties.properties”文件中配置。
bind	-	绑定关联的主机名或IP地址。
port	-	绑定端口。

参数	默认值	描述
handler	org.apache.flume.source.http.JSONHandler	http请求的消息解析方式，支持以下两种： ● “org.apache.flume.source.http.JSONHandler”：表示Json格式解析。 ● “org.apache.flume.sink.solr.morphline.BlobHandler”：表示二进制Blob块解析。
handler.*	-	设置handler的参数。
enableSSL	false	http协议是否启用SSL。
keystore	-	http启用SSL后设置keystore的路径。
keystorePassword	-	http启用SSL后设置keystore的密码。

常用 Channel 配置

- **Memory Channel**

  Memory Channel使用内存作为缓存区，Events存放在内存队列中。常用配置如表12-125所示：

  表 12-125 Memory Channel 常用配置

参数	默认值	描述
type	-	类型，需配置为“memory”。仅可在“properties.properties”文件中配置。
capacity	10000	缓存在Channel中的最大Event数。
transactionCapacity	1000	每次存取的最大Event数。
channelfullcount	10	Channel full次数，达到该次数后发送告警。

- **File Channel**

  File Channel使用本地磁盘作为缓存区，Events存放在设置的“dataDirs”配置项文件夹中。常用配置如表12-126所示：

  表 12-126 File Channel 常用配置

参数	默认值	描述
type	-	类型，需配置为“file”。仅可在“properties.properties”文件中配置。

参数	默认值	描述
checkpointDir	${BIGDATA_DATA_HOME}/flume/checkpoint	检查点存放路径。
dataDirs	${BIGDATA_DATA_HOME}/flume/data	数据缓存路径，设置多个路径可提升性能，中间用逗号分开。
maxFileSize	2146435071	单个缓存文件的最大值，单位：字节。
minimumRequiredSpace	524288000	缓冲区空闲空间最小值，单位：字节。
capacity	1000000	缓存在Channel中的最大Event数。
transactionCapacity	10000	每次存取的最大Event数。
channelfullcount	10	Channel full次数，达到该次数后发送告警。

- **Kafka Channel**

  Kafka Channel使用kafka集群缓存数据，Kafka提供高可用、多副本，以防Flume或Kafka Broker崩溃，Channel中的数据会立即被Sink消费。常用配置如**表 10 Kafka Channel 常用配置**所示：

  表 12-127 Kafka Channel 常用配置

参数	默认值	描述
type	-	类型，需配置为“org.apache.flume.channel.kafka.KafkaChannel”.。 仅可在“properties.properties”文件中配置。
kafka.bootstrap.servers	-	kafka broker列表。
kafka.topic	flume-channel	Channel用来缓存数据的topic。
kafka.consumer.group.id	flume	Kafka消费者组ID。
parseAsFlumeEvent	true	是否解析为Flume event。
migrateZookeeperOffsets	true	当Kafka没有存储offset时，是否从ZooKeeper中查找，并提交到Kafka。

参数	默认值	描述
kafka.consumer.auto.offset.reset	latest	当没有offset记录时，从指定的位置消费数据。
kafka.producer.security.protocol	SASL_PLAINTEXT	Kafka生产者安全协议。
kafka.consumer.security.protocol	SASL_PLAINTEXT	Kafka消费者安全协议。

常用 Sink 配置

- **HDFS Sink**

  HDFS Sink将数据写入HDFS。常用配置如表12-128所示：

  表 12-128 HDFS Sink 常用配置

参数	默认值	描述
channel	-	与之相连的Channel。仅可在“properties.properties”文件中配置。
type	hdfs	类型，需配置为“hdfs”。仅可在“properties.properties”文件中配置。
monTime	0（不开启）	线程监控阈值，更新时间大于阈值时重新启动该Sink，单位：秒。
hdfs.path	-	HDFS路径。
hdfs.inUseSuffix	.tmp	正在写入的HDFS文件后缀。
hdfs.rollInterval	30	按时间滚动文件，单位：秒。
hdfs.rollSize	1024	按大小滚动文件，单位：字节。
hdfs.rollCount	10	按Event个数滚动文件。
hdfs.idleTimeout	0	自动关闭空闲文件超时时间，单位：秒。
hdfs.batchSize	1000	每次写入HDFS的Event个数。
hdfs.kerberosPrincipal	-	认证HDFS的Kerberos用户名，未启用Kerberos认证集群不配置。
hdfs.kerberosKeytab	-	认证HDFS的Kerberos keytab路径，未启用Kerberos认证集群不配置
hdfs.fileCloseByEndEvent	true	收到最后一个Event时是否关闭文件。

参数	默认值	描述
hdfs.batchCallTimeout	-	每次写入HDFS超时控制时间，单位：毫秒。 当不配置此参数时，对每个Event写入HDFS进行超时控制。当“hdfs.batchSize”大于0时，配置此参数可以提升写入HDFS性能。 **说明** “hdfs.batchCallTimeout”设置多长时间需要考虑“hdfs.batchSize”的大小，“hdfs.batchSize”越大，“hdfs.batchCallTimeout”也要调整更长时间，设置过短时间容易导致数据写入HDFS失败。
serializer.appendNewline	true	将一个Event写入HDFS后是否追加换行符（'\n'），如果追加该换行符，该换行符所占用的数据量指标不会被HDFS Sink统计。

- **Avro Sink**

  Avro Sink把events转化为Avro events并发送到配置的主机的监听端口。常用配置如表12-129所示：

  表 12-129 Avro Sink 常用配置

参数	默认值	描述
channel	-	与之相连的Channel。仅可在“properties.properties”文件中配置。
type	-	类型，需配置为“avro”。仅可在“properties.properties”文件中配置。
hostname	-	绑定关联的主机名或IP地址。
port	-	监听端口。
batch-size	1000	批次发送的Event个数。
ssl	false	是否使用SSL加密。
truststore-type	JKS	Java信任库类型。
truststore	-	Java信任库文件。
truststore-password	-	Java信任库密码。
keystore-type	JKS	密钥存储类型。
keystore	-	密钥存储文件。
keystore-password	-	密钥存储密码

- **HBase Sink**

  HBase Sink将数据写入到HBase中。常用配置如表12-130所示：

  表 12-130 HBase Sink 常用配置

参数	默认值	描述
channel	-	与之相连的Channel。仅可在“properties.properties”文件中配置。
type	-	类型，需配置为“hbase”。仅可在“properties.properties”文件中配置。
table	-	HBase表名称。
monTime	0（不开启）	线程监控阈值，更新时间大于阈值时重新启动该Sink，单位：秒。
columnFamily	-	HBase列族名称。
batchSize	1000	每次写入HBase的Event个数。
kerberosPrincipal	-	认证HBase的Kerberos用户名，未启用Kerberos认证集群不配置。
kerberosKeytab	-	认证HBase的Kerberos keytab路径，未启用Kerberos认证集群不配置。

- **Kafka Sink**

  Kafka Sink将数据写入到Kafka中。常用配置如表12-131所示：

  表 12-131 Kafka Sink 常用配置

参数	默认值	描述
channel	-	与之相连的Channel。仅可在“properties.properties”文件中配置。
type	-	类型，需配置为“org.apache.flume.sink.kafka.KafkaSink”。 仅可在“properties.properties”文件中配置。
kafka.bootstrap.servers	-	Kafkabrokers列表，多个用英文逗号分隔。
monTime	0（不开启）	线程监控阈值，更新时间大于阈值时重新启动该Sink，单位：秒。

参数	默认值	描述
kafka.topic	default-flume-topic	数据写入的topic。
flumeBatchSize	1000	每次写入Kafka的Event个数。
kafka.security.protocol	SASL_PLAINTEXT	Kafka安全协议，未启用Kerberos认证集群下须配置为“PLAINTEXT”。
kafka.kerberos.domain.name	-	Kafka Domain名称。安全集群必填。仅可在“properties.properties”文件中配置。
Other Kafka Producer Properties	-	其他Kafka配置，可以接受任意Kafka支持的生产参数配置，配置需要加前缀“.kafka”。 仅可在“properties.properties”文件中配置。

12.7.9 在配置文件 properties.properties 中使用环境变量

操作场景

本章节描述如何在配置文件“properties.properties”中使用环境变量。

本章节适用于MRS 3.x及之后版本。

前提条件

已成功安装Flume服务或Flume客户端。

操作步骤

步骤1 在“*Flume客户端安装目录*/fusioninsight-flume-*Flume组件版本号*/conf/flume-env.sh”文件中添加变量。

添加变量：

export *变量名=变量值*

示例如下：

JAVA_OPTS="-Xms2G -Xmx4G -XX:CMSFullGCsBeforeCompaction=1 -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:+UseCMSCompactAtFullCollection -DpropertiesImplementation=org.apache.flume.node.EnvVarResolverProperties"

export TAILDIR_PATH=/tmp/flumetest/201907/20190703/1/.*log.*

步骤2 重启Flume实例进程。

1. 登录FusionInsight Manager。

2. 选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，勾选Flume实例，选择“更多 > 重启实例”输入密码，单击“确定”等待实例重启成功。

须知

- 服务端flume-env.sh生效后不能通过Manager界面重启整个Flume服务，否则用户自定义环境变量丢失。
- 服务端必须保证flume-env.sh生效之后，再执行步骤3配置properties.properties文件，然后再通过界面上传。若操作顺序不规范，可能造成用户自定义环境变量丢失。

步骤3 在properties.properties配置文件中使用“${变量名}”格式引用变量，以下以客户端为例。

示例如下：

```
client.sources.s1.type = TAILDIR
client.sources.s1.filegroups = f1
client.sources.s1.filegroups.f1 = ${TAILDIR_PATH}
client.sources.s1.positionFile = /tmp/flumetest/201907/20190703/1/taildir_position.json
client.sources.s1.channels = c1
```

----结束

12.7.10 非加密传输

12.7.10.1 配置非加密传输

操作场景

该操作指导安装工程师在集群及Flume服务安装完成后，分别配置Flume服务的服务端和客户端参数，使其可以正常工作。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。

前提条件

- 已成功安装集群及Flume服务。
- 确保集群网络环境安全。

操作步骤

步骤1 配置Flume角色客户端参数。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > 服务 > Flume > 配置工具”。

b. “Agent名”选择“client”，然后选择要使用的Source、Channel以及Sink，将其拖到右侧的操作界面中并将其连接。

例如采用SpoolDir Source、File Channel和Avro Sink。

c. 双击对应的Source、Channel以及Sink，根据实际环境并参考**表12-132**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > 服务 > Flume > 配置 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置Source/Channel/Sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-132 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl	是否启用SSL认证（基于安全要求，建议启用此功能） 只有“Avro”类型的Source才有此配置项 - true表示启用 - false表示不启用	false

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤2 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

   例如采用Avro Source、File Channel和HDFS Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-133**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在 FusionInsight Manager界面选择“集群 > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置 ”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“集群 > 服务 > Flume > 配置 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
- 不同的File Channel均需要配置一个不同的checkpoint目录。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-133 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl	是否启用SSL认证（基于安全要求，建议启用此功能） 只有“Avro”类型的Source才有此配置项 ■ true表示启用 ■ false表示不启用	false

2. 登录FusionInsight Manager，选择“集群 > 服务 > Flume”，在“实例”下单击“Flume”角色。
3. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置 ”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

   说明

   - 每个Flume实例均可以上传单独的服务端配置文件。
   - 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。
4. 单击“保存”，单击“确定”。
5. 单击“完成”完成操作。

----结束

12.7.10.2 典型场景：从本地采集静态日志保存到 Kafka

操作场景

该任务指导用户使用Flume从本地（业务IP:192.168.108.11）采集静态日志保存到Kafka的Topic列表（test1）。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。该配置可以只用一个Flume场景，例如Server:Spooldir Source+File Channel+Kafka Sink.

前提条件

- 已成功安装集群、Kafka及Flume服务。
- 确保集群网络环境安全。
- 系统管理员已明确业务需求，并准备一个Kafka管理员用户**flume_kafka**。

操作步骤

步骤1 配置Flume的角色客户端参数。

1. 使用Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用SpoolDir Source、File Channel和Avro Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-134**设置对应的配置参数。

      说明

      - 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录Manager，选择“集群 > 服务 > Flume > 配置 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
      - 导入配置文件时，建议配置Source/Channel/Sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
   d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-134 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一	test
spoolDir	待采集的文件所在的目录路径，此参数不能为空。该路径需存在，且对flume运行用户有读写执行权限。	/srv/BigData/hadoop/data1/zb
trackerDir	flume采集文件信息元数据保存路径。	/srv/BigData/hadoop/data1/tracker

参数名称	参数值填写规则	参数样例
batchSize	Flume一次发送的事件个数（数据条数）。增大会提升性能，降低实时性；反之降低性能，提升实时性。	61200
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数。建议和Source的batchSize设置为同样大小，不能小于batchSize	61200
hostname	要发送数据的主机名或者IP，此参数不能为空。须配置为与之相连的avro source所在的主机名或IP。	192.168.108.11
port	要发送数据的端口，此参数不能为空。须配置为与之相连的avro source监听的端口。	21154

参数名称	参数值填写规则	参数样例
ssl	是否启用SSL认证（基于安全要求，建议启用此功能） 只有“Avro”类型的Source才有此配置项 ■ true表示启用 ■ false表示不启用	false

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤2 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。
      采用Avro Source、File Channel和Kafka Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-135**设置对应的配置参数。

      说明

      - 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在Manager界面选择“集群 > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置 ”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“集群 > 服务 > Flume >配置 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
      - 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
      - 不同的File Channel均需要配置一个不同的checkpoint目录。

   d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-135 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
bind	avro source绑定的ip地址，此参数不能为空。须配置为服务端配置文件即将要上传的主机IP。	192.168.108.11
port	avro source监听的端口，此参数不能为空。须配置为未被使用的端口。	21154

参数名称	参数值填写规则	参数样例
ssl	是否启用SSL认证（基于安全要求，用户启用此功能）。 只有“avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/data
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数。建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
kafka.topics	订阅的Kafka topic列表，用逗号分隔，此参数不能为空。	test1
kafka.bootstrap.servers	Kafka 的bootstrap 地址端口列表,默认值为Kafka集群中所有的Kafka列表。如果集群安装有Kafka并且配置已经同步，可以不配置此项。	192.168.101.10:21007

2. 登录FusionInsight Manager，选择“集群 > 服务 > Flume”，在“实例”下单击“Flume”角色。
3. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置 ”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

   说明

   - 每个Flume实例均可以上传单独的服务端配置文件。
   - 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。

4. 单击“保存”，单击“确定”。
5. 单击“完成”完成操作。

步骤3 验证日志是否传输成功。

1. 登录Kafka客户端：

   cd /*客户端安装目录*/Kafka/kafka

   kinit flume_kafka（输入密码）

2. 读取KafkaTopic中的数据（修改命令中的中文为实际参数）。

   bin/kafka-console-consumer.sh --topic *主题名称* --bootstrap-server *Kafka角色实例所在节点的业务IP地址*:21007 --consumer.config config/consumer.properties --from-beginning

   系统显示待采集文件目录下的内容：

   ```
   [root@host1 kafka]# bin/kafka-console-consumer.sh --topic test1 --bootstrap-server 192.168.101.10:21007 --consumer.config config/consumer.properties --from-beginning
   Welcome to flume
   ```

----结束

12.7.10.3 典型场景：从本地采集静态日志保存到 HDFS

操作场景

该任务指导用户使用Flume从本地（例如业务IP为192.168.108.11）采集静态日志保存到HDFS上“/flume/test”目录下。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。该配置可以只用一个Flume场景，例如Server:Spooldir Source+File Channel+HDFS Sink。

前提条件

- 已成功安装集群、HDFS及Flume服务。
- 确保集群网络环境安全。
- 已创建用户**flume_hdfs**并授权验证日志时操作的HDFS目录和数据。

操作步骤

步骤1 在FusionInsight Manager管理界面，选择“系统 > 权限 > 用户”，选择用户**flume_hdfs**，选择“更多 > 下载认证凭据”下载Kerberos证书文件并保存在本地。

步骤2 配置Flume角色客户端参数。

1. 使用FusionInsight Manager界面中的Flume来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用SpoolDir Source、File Channel和Avro Sink。

c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-136**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-136 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
spoolDir	待采集的文件所在的目录路径，此参数不能为空。该路径需存在，且对flume运行用户有读写执行权限。	/srv/BigData/hadoop/data1/zb
trackerDir	flume采集文件信息元数据保存路径。	/srv/BigData/hadoop/data1/tracker
batch-size	Flume一次发送数据的最大事件数。	61200
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hostname	要发送数据的主机名或者IP，此参数不能为空。须配置为与之相连的avro source所在的主机名或IP。	192.168.108.11
port	avro sink监听的端口，此参数不能为空。须配置为与之相连的avro source监听的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤3 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“ 服务 > Flume > 配置工具”。
   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。
      采用Avro Source、File Channel和HDFS Sink。

c. 双击对应的Source、Channel以及Sink，根据实际环境并参考**表12-137**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
- 不同的File Channel均需要配置一个不同的checkpoint目录。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-137 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
bind	avro source绑定的ip地址，此参数不能为空。须配置为服务端配置文件即将要上传的主机IP。	192.168.108.11
port	avro source监听的端口，此参数不能为空。须配置为未被使用的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数。建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hdfs.path	写入HDFS的目录，此参数不能为空。	hdfs://hacluster/flume/test
hdfs.inUsePrefix	正在写入HDFS的文件的前缀。	TMP_
hdfs.batchSize	一次写入HDFS的最大事件数目。	61200
hdfs.kerberosPrincipal	kerberos认证时用户，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	flume_hdfs
hdfs.kerberosKeytab	kerberos认证时keytab文件路径，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	/opt/test/conf/user.keytab **说明** user.keytab文件从下载用户**flume_hdfs**的kerberos证书文件中获取，另外，确保用于安装和运行Flume客户端的用户对user.keytab文件有读写权限。
hdfs.useLocalTimeStamp	是否使用本地时间，取值为"true"或者"false"。	true

2. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”，在“角色”下单击“Flume”角色。
3. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

   说明

   - 每个Flume实例均可以上传单独的服务端配置文件。
   - 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。
4. 单击“保存”，单击“确定”。

5. 单击“完成”完成操作。

步骤4 验证日志是否传输成功。

1. 以具有HDFS组件管理权限的用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击“NameNode(主)”对应的链接，打开HDFS WebUI，然后选择“Utilities > Browse the file system”。
2. 观察HDFS上“/flume/test”目录下是否有产生数据。

----结束

12.7.10.4 典型场景：从本地采集动态日志保存到 HDFS

操作场景

该任务指导用户使用Flume从本地(业务IP:192.168.108.11)采集动态日志保存到HDFS上“/flume/test”目录下。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。该配置可以只用一个Flume场景，例如Server:Taildir Source+File Channel +HDFS Sink。

前提条件

- 已成功安装集群、HDFS及Flume服务。
- 确保集群网络环境安全。
- 已创建用户**flume_hdfs**并授权验证日志时操作的HDFS目录和数据。

操作步骤

步骤1 在FusionInsight Manager管理界面，选择“系统 > 权限 > 用户”，选择“更多 > 下载认证凭据”下载用户**flume_hdfs**的kerberos证书文件并保存在本地。

步骤2 配置Flume角色客户端参数。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用Taildir Source、File Channel和Avro Sink。
   c. 双击对应的Source、Channel以及Sink，根据实际环境并参考**表12-138**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置Source/Channel/Sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-138 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
filegroups	文件分组列表名，此参数不能为空,以空格分隔。	epgtest
positionFile	保存当前采集文件信息（文件名和已经采集的位置），此参数不能为空。该文件不需要手工创建，但其上层目录需对flume运行用户可写。	/home/omm/flume/positionfile
batch-size	Flume一次发送数据的最大事件数。	61200
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hostname	要发送数据的主机名或者IP，此参数不能为空。须配置为与之相连的avro source所在的主机名或IP。	192.168.108.11
port	avro sink监听的端口，此参数不能为空。须配置为与之相连的avro source监听的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤3 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。
      采用Avro Source、File Channel和HDFS Sink。

c. 双击对应的source、channel以及sink，根据实际环境并参考表12-139设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置Source/Channel/Sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
- 不同的File Channel均需要配置一个不同的checkpoint目录。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-139 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
bind	avro source绑定的ip地址，此参数不能为空。须配置为服务端配置文件即将要上传的主机IP。	192.168.108.11
port	avro source监听的端口，此参数不能为空。须配置为未被使用的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数。建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hdfs.path	写入HDFS的目录，此参数不能为空。	hdfs://hacluster/flume/test
hdfs.inUsePrefix	正在写入HDFS的文件的前缀。	TMP_
hdfs.batchSize	一次写入HDFS的最大事件数目。	61200
hdfs.kerberosPrincipal	kerberos认证时用户，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	flume_hdfs
hdfs.kerberosKeytab	kerberos认证时keytab文件路径，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	/opt/test/conf/user.keytab **说明** user.keytab文件从下载用户**flume_hdfs**的kerberos证书文件中获取，另外，确保用于安装和运行Flume客户端的用户对user.keytab文件有读写权限。
hdfs.useLocalTimeStamp	是否使用本地时间，取值为"true"或者"false"。	true

2. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”，在“角色”下单击“Flume”角色。
3. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

   说明

   - 每个Flume实例均可以上传单独的服务端配置文件。
   - 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。
4. 单击“保存”，单击“确定”。

5. 单击“完成”完成操作。

步骤4 验证日志是否传输成功。

1. 以具有HDFS组件管理权限的用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击“NameNode(*节点名称*，主)”对应的链接，打开HDFS WebUI，然后选择“Utilities > Browse the file system”。
2. 观察HDFS上“/flume/test”目录下是否有产生数据。

----结束

12.7.10.5 典型场景：从 Kafka 采集日志保存到 HDFS

操作场景

该任务指导用户使用Flume从Kafka的Topic列表(test1)采集日志保存到HDFS上“/flume/test”目录下。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。该配置可以只用一个Flume场景，例如Server:Kafka Source+File Channel +HDFS Sink.

前提条件

- 已成功安装集群、HDFS、Kafka及Flume服务。
- 确保集群网络环境安全。
- 已创建用户**flume_hdfs**并授权验证日志时操作的HDFS目录和数据。

操作步骤

步骤1 在FusionInsight Manager管理界面，选择“系统 > 权限 > 用户”，选择“更多 > 下载认证凭据”下载用户**flume_hdfs**的kerberos证书文件并保存在本地。

步骤2 配置Flume角色客户端参数。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      例如采用Kafka Source、File Channel和Avro Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-140**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-140 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
kafka.topics	订阅的Kafka topic列表，用逗号分隔，此参数不能为空。	test1
kafka.consumer.group.id	从Kafka中获取数据的组标识，此参数不能为空。	flume
kafka.bootstrap.servers	Kafka的bootstrap地址端口列表,默认值为Kafka集群中所有的Kafka列表。如果集群安装有Kafka并且配置已经同步，可以不配置此项。	192.168.101.10:9092
batchSize	Flume一次发送的事件个数（数据条数）。	61200
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hostname	要发送数据的主机名或者IP，此参数不能为空。须配置为与之相连的avro source所在的主机名或IP。	192.168.108.11
port	avro sink监听的端口，此参数不能为空。须配置为与之相连的avro source监听的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤3 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。
      采用Avro Source、File Channel和HDFS Sink。

c. 双击对应的Source、Channel以及Sink，根据实际环境并参考**表12-141**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
- 不同的File Channel均需要配置一个不同的checkpoint目录。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-141 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
bind	avro source绑定的ip地址，此参数不能为空。须配置为服务端配置文件即将要上传的主机IP。	192.168.108.11
port	avro source监听的端口，此参数不能为空。须配置为未被使用的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 - true表示启用 - false表示不启用	false
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hdfs.path	写入HDFS的目录，此参数不能为空。	hdfs://hacluster/flume/test
hdfs.inUsePrefix	正在写入HDFS的文件的前缀。	TMP_
hdfs.batchSize	一次写入HDFS的最大事件数目。	61200
hdfs.kerberosPrincipal	kerberos认证时用户，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	flume_hdfs
hdfs.kerberosKeytab	kerberos认证时keytab文件路径，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	/opt/test/conf/user.keytab **说明** user.keytab文件从下载用户**flume_hdfs**的kerberos证书文件中获取，另外，确保用于安装和运行Flume客户端的用户对user.keytab文件有读写权限。
hdfs.useLocalTimeStamp	是否使用本地时间，取值为"true"或者"false"。	true

2. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”，在“角色”下单击“Flume”角色。
3. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

   说明

   - 每个Flume实例均可以上传单独的服务端配置文件。
   - 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。
4. 单击“保存”，单击“确定”。

5. 单击“完成”完成操作。

步骤4 验证日志是否传输成功。

1. 以具有HDFS组件管理权限的用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击“NameNode(*节点名称*，主)”对应的链接，打开HDFS WebUI，然后选择“Utilities > Browse the file system”。
2. 观察HDFS上“/flume/test”目录下是否有产生数据。

----结束

12.7.10.6 典型场景：从 Kafka 客户端采集日志经 Flume 客户端保存到 HDFS

操作场景

该任务指导用户使用Flume从Kafka客户端的Topic列表(test1)采集日志保存到HDFS上“/flume/test”目录下。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。

前提条件

- 已成功安装集群、HDFS、Kafka及Flume服务。
- 已创建用户**flume_hdfs**并授权验证日志时操作的HDFS目录和数据。
- 确保集群网络环境安全。

操作步骤

步骤1 在FusionInsight Manager管理界面，选择“系统 > 权限 > 用户”，选择“更多 > 下载认证凭据”下载用户**flume_hdfs**的kerberos证书文件并保存在本地。

步骤2 配置Flume角色客户端参数。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      例如采用Kafka Source、File Channel和HDFS Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-142**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，可以到“客户端安装目录/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-142 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
kafka.topics	订阅的Kafka topic列表，用逗号分隔，此参数不能为空。	test1
kafka.consumer.group.id	从Kafka中获取数据的组标识，此参数不能为空。	flume
kafka.bootstrap.servers	Kafka的bootstrap地址端口列表,默认值为Kafka集群中所有的Kafka列表。如果集群安装有Kafka并且配置已经同步，可以不配置此项。	192.168.101.10:21007
batchSize	Flume一次发送的事件个数（数据条数）。	61200
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hdfs.path	写入HDFS的目录，此参数不能为空。	hdfs://hacluster/flume/test
hdfs.inUsePrefix	正在写入HDFS的文件的前缀。	TMP_
hdfs.batchSize	一次写入HDFS的最大事件数目。	61200
hdfs.kerberosPrincipal	kerberos认证时用户，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	flume_hdfs
hdfs.kerberosKeytab	kerberos认证时keytab文件路径，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	/opt/test/conf/user.keytab **说明** user.keytab文件从下载用户**flume_hdfs**的kerberos证书文件中获取，另外，确保用于安装和运行Flume客户端的用户对user.keytab文件有读写权限。
hdfs.useLocalTimeStamp	是否使用本地时间，取值为"true"或者"false"	true

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。
3. Flume客户端连接到HDFS，还需要补充如下配置：

a. 通过“用户”下载用户**flume_hdfs**的kerberos证书文件获取krb5.conf配置文件，并上传至客户端所在节点安装目录的“fusioninsight-flume-1.9.0/conf/”下。

b. 新建jaas.conf配置文件到客户端所在节点安装目录的“fusioninsight-flume-1.9.0/conf/”下。

vi jaas.conf

```
KafkaClient {
com.sun.security.auth.module.Krb5LoginModule required
useKeyTab=true
keyTab="/opt/test/conf/user.keytab"
principal="flume_hdfs@<系统域名>"
useTicketCache=false
storeKey=true
debug=true;
};
```

参数keyTab和principal根据实际情况修改。

c. 从/opt/FusionInsight_Cluster_*<集群ID>*_Flume_ClientConfig/Flume/config目录下获取core-site.xml和hdfs-site.xml配置文件，并上传至客户端所在节点安装目录的“fusioninsight-flume-1.9.0/conf/”下。

4. 重启Flume服务。

步骤3 验证日志是否传输成功。

1. 以具有HDFS组件管理权限的用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击“NameNode(*节点名称*，主)”对应的链接，打开HDFS WebUI，然后选择“Utilities > Browse the file system”。
2. 观察HDFS上“/flume/test”目录下是否有产生数据。

----结束

12.7.10.7 典型场景：从本地采集静态日志保存到 HBase

操作场景

该任务指导用户使用Flume从本地(业务IP:192.168.108.11)采集静态日志保存到HBase表：flume_test。

本章节适用于MRS 3.x及之后版本。

说明

本配置默认集群网络环境是安全的，数据传输过程不需要启用SSL认证。如需使用加密方式，请参考**配置加密传输**。该配置可以只用一个Flume场景，例如Server:Spooldir Source+File Channel+HBase Sink。

前提条件

- 已成功安装集群、HBase及Flume服务。
- 确保集群网络环境安全。
- 已创建HBase表：**create 'flume_test', 'cf'**。
- 系统管理员已明确业务需求，并准备一个HBase管理员用户**flume_hbase**。

操作步骤

步骤1 在FusionInsight Manager管理界面，选择“系统 > 权限 > 用户”，选择“更多 > 下载认证凭据”下载用户**flume_hbase**的kerberos证书文件并保存在本地。

步骤2 配置Flume角色客户端参数。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用SpoolDir Source、File Channel和Avro Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-143**设置对应的配置参数。

      说明

      - 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
      - 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
   d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-143 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
spoolDir	待采集的文件所在的目录路径，此参数不能为空。该路径需存在，且对flume运行用户有读写执行权限。	/srv/BigData/hadoop/data1/zb
trackerDir	flume采集文件信息元数据保存路径。	/srv/BigData/hadoop/data1/tracker
batchSize	Flume一次发送的事件个数（数据条数）。增大会提升性能，降低实时性；反之降低性能，提升实时性。	61200

参数名称	参数值填写规则	参数样例
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data
checkpointDir	checkpoint信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hostname	要发送数据的主机名或者IP，此参数不能为空。须配置为与之相连的avro source所在的主机名或IP。	192.168.108.11
port	要发送数据的端口，此参数不能为空。须配置为与之相连的avro source监听的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false

2. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤3 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用Avro Source、File Channel和HBase Sink。
   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-144**设置对应的配置参数。

      说明

      - 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改非加密传输的相关配置项即可。
      - 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
      - 不同的File Channel均需要配置一个不同的checkpoint目录。
   d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-144 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
bind	avro source绑定的ip地址，此参数不能为空。须配置为服务端配置文件即将要上传的主机IP。	192.168.108.11
port	avro source监听的端口，此参数不能为空。须配置为未被使用的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用 ■ false表示不启用	false

参数名称	参数值填写规则	参数样例
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/data
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数。建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
table	HBase表名，此参数不能为空。	flume_test
columnFamily	HBase列族名，此参数不能为空。	cf
batchSize	Flume一次写入HBase中的最大事件数。	61200
kerberosPrincipal	kerberos认证时用户,在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	flume_hbase
kerberosKeytab	kerberos认证时文件路径，,在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	/opt/test/conf/user.keytab **说明** user.keytab文件从下载用户**flume_hbase**的kerberos证书文件中获取，另外，确保用于安装和运行Flume客户端的用户对user.keytab文件有读写权限。

2. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”，在“实例”下单击“Flume”角色。
3. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

说明

- 每个Flume实例均可以上传单独的服务端配置文件。
- 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。

4. 单击“保存”，单击“确定”。
5. 单击“完成”完成操作。

步骤4 验证日志是否传输成功。

1. 进入HBase客户端目录：

   cd /*客户端安装目录***/HBase/hbase**

   kinit flume_hbase（输入密码）
2. 执行**hbase shell**进入HBase客户端。
3. 执行语句：**scan 'flume_test'**，可以看到日志按行写入HBase列族里。

```
hbase(main):001:0> scan 'flume_test'
ROW                                                COLUMN
+CELL

2017-09-18 16:05:36,394 INFO  [hconnection-0x415a3f6a-shared--pool2-t1] ipc.AbstractRpcClient:
RPC Server Kerberos principal name for service=ClientService is hbase/hadoop.<系统域名>@<系统域名>
 default4021ff4a-9339-4151-a4d0-00f20807e76d                column=cf:pCol,
timestamp=1505721909388, value=Welcome to
flume
 incRow                                            column=cf:iCol, timestamp=1505721909461, value=
\x00\x00\x00\x00\x00\x00\x00\x01
2 row(s) in 0.3660 seconds
```

----结束

12.7.11 加密传输

12.7.11.1 配置加密传输

操作场景

该操作指导安装工程师在集群安装完成后，分别设置Flume服务（包括Flume角色和MonitorServer角色）的服务端和客户端参数，使其可以正常工作。

本章节适用于MRS 3.x及之后版本。

前提条件

已成功安装集群及Flume服务。

操作步骤

步骤1 分别生成Flume角色服务端和客户端的证书和信任列表。

1. 使用ECM远程以**omm**用户登录将要安装Flume服务端的节点。进入“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin”目录。

   cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin

说明

此处版本号8.1.0.1为示例，具体以实际环境的版本号为准。

2. 执行以下命令，生成并导出Flume角色服务端、客户端证书。

   sh geneJKS.sh -f *xxx* **-g** *xxx*

   生成的证书在“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf”路径下。其中：

   - “flume_sChat.jks”是Flume角色服务端的证书库，“flume_sChat.crt”是“flume_sChat.jks”证书的导出文件，“-f”配置项是证书和证书库的密码；
   - “flume_cChat.jks”是Flume角色客户端的证书库，“flume_cChat.crt”是“flume_cChat.jks”证书的导出文件，“-g”配置项是证书和证书库的密码；
   - “flume_sChatt.jks”和“flume_cChatt.jks”分别为Flume服务端、客户端SSL证书信任列表。

   说明

   本章节涉及到所有的用户自定义密码（如xxx），需满足以下复杂度要求：

   - 至少包含大写字母、小写字母、数字、特殊符号4种类型字符。
   - 至少8位，最多64位。
   - 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤2 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 使用ECM远程，以**omm**用户登录任意一个Flume角色所在的节点。执行以下命令进入“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin”。

   cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin

2. 执行以下命令，生成并得到Flume服务端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是*flume_sChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

3. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。

   a. 登录FusionInsight Manager，选择“ 服务 > Flume > 配置工具”。

   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      例如采用Avro Source、File Channel和HDFS Sink。

   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-145**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在FusionInsight Manager界面选择“ 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置”页面“flume.config.file”参数后的“下载文件”，可获取已有的服务端参数配置文件。然后选择“ 服务 > Flume > 导入”，将该文件导入后再修改加密传输的相关配置项即可。
- 导入配置文件时，建议配置Source/Channel/Sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-145 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl	是否启用SSL认证（基于安全要求，建议启用此功能）。 ■ true表示启用。 ■ false表示不启用。	true
keystore	服务端证书。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_sChat.jks
keystore-password	密钥库密码，获取keystore信息所需密码。 输入**步骤2.2**中获取的“password”值。	-
truststore	服务端的SSL证书信任列表。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_sChatt.jks
truststore-password	信任列表密码，获取truststore信息所需密码。 输入**步骤2.2**中获取的“password”值。	-

4. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”服务，在“角色”下单击“Flume”角色。
5. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

说明

- 每个Flume实例均可以上传单独的服务端配置文件。
- 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。

6. 单击“保存”，单击“确定”。单击“完成”完成操作。

步骤3 设置Flume角色客户端参数。

1. 执行以下命令将生成的客户端证书（flume_cChat.jks）和客户端信任列表（flume_cChatt.jks）复制到客户端目录下，如“/opt/flume-client/fusionInsight-flume-1.9.0/conf/”（要求已安装Flume客户端），其中**10.196.26.1**为客户端所在节点业务平面的IP地址。

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_cChat.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_cChatt.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

   说明

   复制过程中需要输入客户端所在主机（如**10.196.26.1**）**user**用户的密码。

2. 以**user**用户登录解压Flume客户端的节点。执行以下命令进入客户端目录“opt/flume-client/fusionInsight-flume-1.9.0/bin”。

   cd opt/flume-client/fusionInsight-flume-1.9.0/bin

3. 执行以下命令，生成并得到Flume客户端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是别名为*flumechatclient*的证书和*flume_cChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

   说明

   若产生以下错误提示，可执行命令**export JAVA_HOME=***JDK路径*进行处理。

   JAVA_HOME is null in current user,please install the JDK and set the JAVA_HOME

4. 执行**echo $SCC_PROFILE_DIR**检查SCC_PROFILE_DIR环境变量是否为空。
   - 是，执行**source .sccfile**。
   - 否，执行**步骤3.5**。

5. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      例如采用SpoolDir Source、File Channel和Avro Sink。

   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-146**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“客户端安装目录/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改加密传输的相关配置项即可。
- 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
- 不同的File Channel均需要配置一个不同的checkpoint目录。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-146 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 ■ true表示启用。 ■ false表示不启用。	true
keystore	客户端证书。	/opt/flume-client/fusionInsight-flume-1.9.0/conf/flume_cChat.jks
keystore-password	密钥库密码，获取keysore信息所需密码。 输入**步骤3.3**中获取的“password”值。	-
truststore	客户端的SSL证书信任列表。	/opt/flume-client/fusionInsight-flume-1.9.0/conf/flume_cChatt.jks
truststore-password	信任列表密码，获取truststore信息所需密码。 输入**步骤3.3**中获取的“password”值。	-

6. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤4 分别生成MonitorServer角色服务端和客户端的证书和信任列表。

1. 使用ECM以**omm**用户登录MonitorServer角色所在主机。

进入“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin”目录。

cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin

2. 执行以下命令，生成并导出MonitorServer角色服务端、客户端证书。

   sh geneJKS.sh -m *xxx* **-n** ***xxx***

   生成的证书在“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf”路径下，其中：

   - “ms_sChat.jks”是MonitorServer角色服务端的证书库，“ms_sChat.crt”是“ms_sChat.jks”证书的导出文件，“-m”配置项是证书和证书库的密码；
   - “ms_cChat.jks”是MonitorServer角色客户端的证书库，“ms_cChat.crt”是“ms_cChat.jks”证书的导出文件，“-n”配置项是证书和证书库的密码；
   - “ms_sChatt.jks”、“ms_cChatt.jks”分别为MonitorServer服务端、客户端SSL证书信任列表。

步骤5 配置MonitorServer角色服务端参数。

1. 执行以下命令，生成并得到MonitorServer服务端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是别名为*mschatserver*的证书和*ms_sChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

2. 使用以下命令打开“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/service/application.properties”文件。根据表12-147中的说明，修改相关参数，并保存退出。

   vi ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1**/install/FusionInsight-Flume-1.9.0/flume/conf/service/application.properties**

表 12-147 MonitorServer 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl_need_kspasswd_decrypt_key	是否开启自定义密钥加解密功能（基于安全要求，建议启用此功能）。 - true表示启用。 - false表示不启用。	true
ssl_server_enable	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 - true表示启用。 - false表示不启用。	true

参数名称	参数值填写规则	参数样例
ssl_server_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_sChat.jks
ssl_server_trust_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_sChatt.jks
ssl_server_key_store_password	keystore密码，根据具体制作证书的实际情况修改（生成证书的明文密钥） 输入**步骤5.1**中获取的“password”值。	-
ssl_server_trust_key_store_password	krustkeystore密码，根据具体制作证书的实际情况修改（生成信任列表的明文密钥）。 输入**步骤5.1**中获取的“password”值。	-
ssl_need_client_auth	是否启用客户端认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true

3. 重启MonitorServer实例。选择“ 服务 > Flume > 实例 > MonitorServer”，勾选配置的“MonitorServer”实例，选择“更多 > 重启实例”。输入管理员密码，单击“确定”，重启完成后单击“完成”完成操作。

步骤6 配置MonitorServer角色客户端参数。

1. 执行以下命令将生成的客户端证书（ms_cChat.jks）和客户端信任列表（ms_cChatt.jks）复制到客户端的“/opt/flume-client/fusionInsight-flume-1.9.0/conf/”目录下，其中**10.196.26.1**为客户端所在节点业务平面的IP地址。

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChat.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChatt.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

2. 以**user**用户登录Flume客户端所在的节点。执行以下命令进入客户端目录“/opt/flume-client/fusionInsight-flume-1.9.0/bin”。

   cd /opt/flume-client/fusionInsight-flume-1.9.0/bin

3. 执行以下命令，生成并得到MonitorServer客户端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是别名为*mschatclient*的证书和*ms_cChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

4. 使用以下命令打开“/opt/flume-client/fusionInsight-flume-1.9.0/conf/service/application.properties”文件（“/opt/flume-client/fusionInsight-flume-1.9.0”为客户端软件安装后的目录）。根据**表12-148**中的说明，修改相关参数，并保存退出。

   vi /opt/flume-client/fusionInsight-flume-1.9.0/flume/conf/service/application.properties

   表 12-148 MonitorServer 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl_need_kspasswd_decrypt_key	是否开启自定义密钥加解密功能（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true
ssl_client_enable	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true
ssl_client_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChat.jks
ssl_client_trust_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChatt.jks
ssl_client_key_store_password	keystore密码，根据具体制作证书的实际情况修改（生成证书的明文密钥）。 输入**步骤6.3**中获取的“password”值。	-

参数名称	参数值填写规则	参数样例
ssl_client_trust_key_store_password	trustkeystore密码，根据具体制作证书的实际情况修改（生成信任列表的明文密钥）。 输入步骤6.3中获取的“password”值。	-
ssl_need_client_auth	是否启用客户端认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true

----结束

12.7.11.2 典型场景：从本地采集静态日志保存到 HDFS

操作场景

该任务指导用户使用Flume从本地(业务IP:192.168.108.11)采集静态日志保存到HDFS上如下目录“/flume/test”。

本章节适用于MRS 3.x及之后版本。

前提条件

- 已成功安装集群、HDFS及Flume服务、Flume客户端。
- 已创建用户**flume_hdfs**并授权验证日志时操作的HDFS目录和数据。

操作步骤

步骤1 分别生成Flume角色服务端和客户端的证书和信任列表。

1. 以**omm**用户登录Flume服务端所在节点。进入“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin”目录。

   cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin

2. 执行以下命令，生成并导出Flume角色服务端、客户端证书。

   sh geneJKS.sh -f *密码* **-g** *密码*

   生成的证书在“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf”路径下。其中：

   – “flume_sChat.jks”是Flume角色服务端的证书库，“flume_sChat.crt”是“flume_sChat.jks”证书的导出文件，“-f”配置项是证书和证书库的密码；

   – “flume_cChat.jks”是Flume角色客户端的证书库，“flume_cChat.crt”是“flume_cChat.jks”证书的导出文件，“-g”配置项是证书和证书库的密码；

- “flume_sChatt.jks”和“flume_cChatt.jks”分别为Flume服务端、客户端SSL证书信任列表。

  说明

  本章节涉及到所有的用户自定义密码，需满足以下复杂度要求：
  - 至少包含大写字母、小写字母、数字、特殊符号4种类型字符
  - 至少8位，最多64位
  - 出于安全考虑，建议用户定期更换自定义密码（例如三个月更换一次），并重新生成各项证书和信任列表。

步骤2 在FusionInsight Manager管理界面，选择“系统 > 权限 > 用户”，选择“更多 > 下载认证凭据”下载用户**flume_hdfs**的kerberos证书文件并保存在本地。

步骤3 配置Flume角色的服务端参数，并将配置文件上传到集群。

1. 以**omm**用户登录任意一个Flume角色所在的节点。执行以下命令进入“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin”。

   cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin

2. 执行以下命令，生成并得到Flume服务端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是*flume_sChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

3. 使用FusionInsight Manager界面中的Flume配置工具来配置服务端参数并生成配置文件。

   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。

   b. “Agent名”选择“server”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用Avro Source、File Channel和HDFS Sink。

   c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-149**设置对应的配置参数。

      说明

      - 如果对应的Flume角色之前已经配置过服务端参数，为保证与之前的配置保持一致，在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例”，选择相应的Flume角色实例，单击“实例配置”页面“flume.config.file”参数后的“下载文件”按钮，可获取已有的服务端参数配置文件。然后选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改加密传输的相关配置项即可。
      - 导入配置文件时，建议配置source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。
      - 不同的File Channel均需要配置一个不同的checkpoint目录。

   d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-149 Flume 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
bind	avro source绑定的ip地址，此参数不能为空。须配置为服务端配置文件即将要上传的主机IP。	192.168.108.11
port	avro source监听的端口,此参数不能为空。须配置为未被使用的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用。 ■ false表示不启用。	true
keystore	服务端证书。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_sChat.jks
keystore-password	密钥库密码，获取keystore信息所需密码。 输入**步骤3.2**中获取的“password”值。	-
truststore	服务端的SSL证书信任列表。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_sChatt.jks
truststore-password	信任列表密码，获取truststore信息所需密码。 输入**步骤3.2**中获取的“password”值。	-

参数名称	参数值填写规则	参数样例
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/data
checkpointDir	checkpoint 信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flumeserver/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数。建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hdfs.path	写入HDFS的目录，此参数不能为空。	hdfs://hacluster/flume/test
hdfs.inUsePrefix	正在写入HDFS的文件的前缀。	TMP_
hdfs.batchSize	一次写入HDFS的最大事件数目。	61200
hdfs.kerberosPrincipal	kerberos认证时用户，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	flume_hdfs
hdfs.kerberosKeytab	kerberos认证时keytab文件路径，在安全版本下必须填写。安全集群需要配置此项，普通模式集群无需配置。	/opt/test/conf/user.keytab **说明** user.keytab文件从下载用户**flume_hdfs**的kerberos证书文件中获取，另外，确保用于安装和运行Flume客户端的用户对user.keytab文件有读写权限。
hdfs.useLocalTimeStamp	是否使用本地时间，取值为"true"或者"false"。	true

4. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”，在“角色”下单击“Flume”角色。

5. 选择准备上传配置文件的节点行的“Flume”角色，单击“实例配置”页面“flume.config.file”参数后的“上传文件”，选择“properties.properties”文件完成操作。

   说明

   - 每个Flume实例均可以上传单独的服务端配置文件。
   - 更新配置文件需要按照此步骤操作，后台修改配置文件是不规范操作，同步配置时后台做的修改将会被覆盖。

6. 单击“保存”，单击“确定”。
7. 单击“完成”完成操作。

步骤4 配置Flume角色客户端参数。

1. 执行以下命令将生成的客户端证书（flume_cChat.jks）和客户端信任列表（flume_cChatt.jks）复制到客户端目录下，如“/opt/flume-client/fusionInsight-flume-1.9.0/conf/”（要求已安装Flume客户端），其中**10.196.26.1**为客户端所在节点业务平面的IP地址。

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_cChat.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/flume_cChatt.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

   说明

   复制过程中需要输入客户端所在主机（如**10.196.26.1**）**user**用户的密码。

2. 以**user**用户登录解压Flume客户端的节点。执行以下命令进入客户端目录“/opt/flume-client/fusionInsight-flume-1.9.0/bin”。

   cd opt/flume-client/fusionInsight-flume-1.9.0/bin

3. 执行以下命令，生成并得到Flume客户端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是别名为*flumechatclient*的证书和*flume_cChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

   说明

   若产生以下错误提示，可执行命令**export JAVA_HOME=***JDK路径*进行处理。

   ```
   JAVA_HOME is null in current user,please install the JDK and set the JAVA_HOME
   ```

4. 执行**echo $SCC_PROFILE_DIR**检查SCC_PROFILE_DIR环境变量是否为空。
   - 是，执行**source .sccfile**。
   - 否，执行**步骤4.5**。

5. 使用FusionInsight Manager界面中的Flume配置工具来配置Flume角色客户端参数并生成配置文件。
   a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具”。
   b. “Agent名”选择“client”，然后选择要使用的source、channel以及sink，将其拖到右侧的操作界面中并将其连接。

      采用SpoolDir Source、File Channel和Avro Sink。

c. 双击对应的source、channel以及sink，根据实际环境并参考**表12-150**设置对应的配置参数。

说明

- 如果对应的Flume角色之前已经配置过客户端参数，为保证与之前的配置保持一致，可以到“*客户端安装目录*/fusioninsight-flume-1.9.0/conf/properties.properties”获取已有的客户端参数配置文件。然后登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置工具 > 导入”，将该文件导入后再修改加密传输的相关配置项即可。
- 导入配置文件时，建议配置中source/channel/sink的各自的个数都不要超过40个，否则可能导致界面响应时间过长。

d. 单击“导出”，将配置文件“properties.properties”保存到本地。

表 12-150 Flume 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
名称	不能为空，必须唯一。	test
spoolDir	待采集的文件所在的目录路径，此参数不能为空。该路径需存在，且对flume运行用户有读写执行权限。	/srv/BigData/hadoop/data1/zb
trackerDir	flume采集文件信息元数据保存路径。	/srv/BigData/hadoop/data1/tracker
batch-size	Flume一次发送数据的最大事件数。	61200
dataDirs	缓冲区数据保存目录，默认为运行目录。配置多个盘上的目录可以提升传输效率，多个目录使用逗号分隔。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/data，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/data

参数名称	参数值填写规则	参数样例
checkpointDir	checkpoint信息保存目录，默认在运行目录下。如果为集群内，则可以指定在如下目录/srv/BigData/hadoop/dataX/flume/checkpoint，dataX为data1~dataN。如果为集群外，则需要单独规划。	/srv/BigData/hadoop/data1/flume/checkpoint
transactionCapacity	事务大小：即当前channel支持事务处理的事件个数，建议和Source的batchSize设置为同样大小，不能小于batchSize。	61200
hostname	要发送数据的主机名或者IP，此参数不能为空。须配置为与之相连的avro source所在的主机名或IP。	192.168.108.11
port	avro sink监听的端口，此参数不能为空。须配置为与之相连的avro source监听的端口。	21154
ssl	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 只有“Avro”类型的Source才有此配置项。 ■ true表示启用。 ■ false表示不启用。	true
keystore	服务端生成的flume_cChat.jks证书。	/opt/flume-client/fusionInsight-flume-1.9.0/conf/flume_cChat.jks
keystore-password	密钥库密码，获取keystore信息所需密码。 输入**步骤4.3**中获取的“password”值。	-

参数名称	参数值填写规则	参数样例
truststore	服务端的SSL证书信任列表。	/opt/flume-client/fusionInsight-flume-1.9.0/conf/flume_cChatt.jks
truststore-password	信任列表密码，获取truststore信息所需密码。 输入步骤4.3中获取的“password”值。	-

6. 将“properties.properties”文件上传到Flume客户端安装目录下的“flume/conf/”下。

步骤5 分别生成MonitorServer角色服务端和客户端的证书和信任列表。

1. 以**omm**用户登录MonitorServer角色所在主机。

   进入“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin”目录。

   cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/bin

2. 执行以下命令，生成并导出MonitorServer角色服务端、客户端证书。

   sh geneJKS.sh -m *密码* **-n** *密码*

   生成的证书在“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf”路径下，其中：

   - “ms_sChat.jks”是MonitorServer角色服务端的证书库，“ms_sChat.crt”是“ms_sChat.jks”证书的导出文件，“-m”配置项是证书和证书库的密码；
   - “ms_cChat.jks”是MonitorServer角色客户端的证书库，“ms_cChat.crt”是“ms_cChat.jks”证书的导出文件，“-n”配置项是证书和证书库的密码；
   - “ms_sChatt.jks”、“ms_cChatt.jks”分别为MonitorServer服务端、客户端SSL证书信任列表。

步骤6 配置MonitorServer角色服务端参数。

1. 执行以下命令，生成并得到MonitorServer服务端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是别名为*mschatserver*的证书和*ms_sChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

2. 使用以下命令打开“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/service/application.properties”文件。根据表12-151中的说明，修改相关参数，并保存退出。

   vi ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1**/install/FusionInsight-Flume-1.9.0/flume/conf/service/application.properties**

表 12-151 MonitorServer 角色服务端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl_need_kspasswd_decrypt_key	是否开启自定义密钥加解密功能（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true
ssl_server_enable	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true
ssl_server_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_sChat.jks
ssl_server_trust_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_sChatt.jks
ssl_server_key_store_password	keystore密码，根据具体制作证书的实际情况修改（生成证书的明文密钥）。 输入**步骤6.1**中获取的“password”值。	-
ssl_server_trust_key_store_password	krustkeystore密码，根据具体制作证书的实际情况修改（生成信任列表的明文密钥）。 输入**步骤6.1**中获取的“password”值。	-
ssl_need_client_auth	是否启用客户端认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true

3. 重启MonitorServer实例。选择“集群 > *待操作集群的名称* > 服务 > Flume > 实例 > MonitorServer”，勾选配置的“MonitorServer”实例，选择“更多 > 重启实例”。输入管理员密码，单击“确定”，重启完成后单击“完成”完成操作。

步骤7 配置MonitorServer角色客户端参数。

1. 执行以下命令将生成的客户端证书（ms_cChat.jks）和客户端信任列表（ms_cChatt.jks）复制到客户端的“/opt/flume-client/fusionInsight-flume-1.9.0/conf/”目录下，其中**10.196.26.1**为客户端所在节点业务平面的IP地址。

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChat.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

   scp ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChatt.jks user@10.196.26.1:/opt/flume-client/fusionInsight-flume-1.9.0/conf/

2. 以**user**用户登录Flume客户端所在的节点。执行以下命令进入客户端目录“/opt/flume-client/fusionInsight-flume-1.9.0/bin”。

   cd /opt/flume-client/fusionInsight-flume-1.9.0/bin

3. 执行以下命令，生成并得到MonitorServer客户端密钥库密码、信任列表密码和keystore-password加密的私钥信息。连续输入两次密码并确认，该密码是别名为*mschatclient*的证书和*ms_cChat.jks*证书库的密码。

   ./genPwFile.sh

   cat password.property

4. 使用以下命令打开“/opt/flume-client/fusionInsight-flume-1.9.0/conf/service/application.properties”文件（“/opt/flume-client/fusionInsight-flume-1.9.0”为客户端安装后的目录）。根据**表12-152**中的说明，修改相关参数，并保存退出。

   vi /opt/flume-client/fusionInsight-flume-1.9.0/conf/service/application.properties

   表 12-152 MonitorServer 角色客户端所需修改的参数列表

参数名称	参数值填写规则	参数样例
ssl_need_kspasswd_decrypt_key	是否开启自定义密钥加解密功能（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true
ssl_client_enable	是否启用SSL认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true
ssl_client_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChat.jks

参数名称	参数值填写规则	参数样例
ssl_client_trust_key_store	根据具体的存放位置进行修改。	${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Flume-1.9.0/flume/conf/ms_cChatt.jks
ssl_client_key_store_password	keystore密码，根据具体制作证书的实际情况修改（生成证书的明文密钥）。 输入**步骤7.3**中获取的“password”值。	-
ssl_client_trust_key_store_password	trustkeystore密码，根据具体制作证书的实际情况修改（生成信任列表的明文密钥）。 输入**步骤7.3**中获取的“password”值。	-
ssl_need_client_auth	是否启用客户端认证（基于安全要求，建议用户启用此功能）。 – true表示启用。 – false表示不启用。	true

步骤8 验证日志是否传输成功。

1. 以具有HDFS组件管理权限的用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。在FusionInsight Manager界面选择“集群 > *待操作集群的名称* > 服务 > HDFS”，单击“NameNode(*节点名称*，主)”对应的链接，打开HDFS WebUI，然后选择“Utilities > Browse the file system”
2. 观察HDFS上“/flume/test”目录下是否有产生数据。

----结束

12.7.12 查看 Flume 客户端监控信息

操作场景

集群外的Flume客户端也是端到端数据采集的一环，与集群内Flume服务端一起都需要监控，用户通过FusionInsight Manager可以对Flume客户端进行监控，可以查看客户端的Source、Sink、Channel的监控指标以及客户端的进程状态。

本章节适用于MRS 3.x及之后版本。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Flume > Flume管理”，即可查看当前Flume客户端列表及进程状态。

步骤3 选择“实例ID”，进入客户端监控列表，在“实时”区域框中，可查看客户端的各监控指标。

步骤4 选择“历史”进入历史监控数据查询界面。筛选时间段，单击“查看”可显示该时间段内的监控数据。

----结束

12.7.13 Flume 对接安全 Kafka 指导

操作场景

使用Flume客户端对接安全kafka。

本章节适用于MRS 3.x及之后版本。

操作步骤

步骤1 新增jaas.conf文件，并保存到“${Flume客户端安装目录}/conf”下，jaas.conf文件内容如下：

```
KafkaClient {
com.sun.security.auth.module.Krb5LoginModule required
useKeyTab=true
keyTab="/opt/test/conf/user.keytab"
principal="flume_hdfs@<系统域名>"
useTicketCache=false
storeKey=true
debug=true;
 };
```

其中**keyTab**和**principal**的值请按照实际情况配置，所配置的**principal**需要有相应的kafka的权限。

步骤2 配置业务，其中**kafka.bootstrap.servers**的端口号使用21007，**kafka.security.protocol**使用**SASL_PLAINTEXT**。

步骤3 如果Kafka所在集群的域名发生了更改，需要对${Flume客户端安装目录}/conf/flume-env.sh文件中的**-Dkerberos.domain.name**项的值做修改，具体请根据实际域名进行配置。

步骤4 上传所配置的**properties.properties**文件到${Flume客户端安装目录}/conf目录下。

----结束

12.7.14 Flume 对接安全 Hive 指导

操作场景

使用Flume对接集群中的Hive（3.1.0版本）。

本章节适用于MRS 3.x及之后版本。

前置条件

集群正确安装了Flume服务和Hive服务，且服务正常无告警异常。

操作步骤

步骤1 使用**omm**用户将如下jar包导入到需要测试的Flume实例的lib目录下（客户端/服务端），列表如下：

- antlr-2.7.7.jar
- antlr-runtime-3.4.jar
- calcite-core-1.16.0.jar
- hadoop-mapreduce-client-core-3.1.1.jar
- hive-beeline-3.1.0.jar
- hive-cli-3.1.0.jar
- hive-common-3.1.0.jar
- hive-exec-3.1.0.jar
- hive-hcatalog-core-3.1.0.jar
- hive-hcatalog-***-adapter-3.1.0.jar
- hive-hcatalog-server-extensions-3.1.0.jar
- hive-hcatalog-streaming-3.1.0.jar
- hive-metastore-3.1.0.jar
- hive-service-3.1.0.jar
- libfb303-0.9.3.jar
- hadoop-plugins-1.0.jar

相关jar包可从Hive安装目录中获取，重启对应的Flume进程，保证jar包加载到运行环境中。

步骤2 配置Hive配置项。

在FusionInsight Manager界面，选择“集群 > 服务 > Hive > 配置 > 全部配置 > HiveServer > 自定义 > hive.server.customized.configs”。

配置项如下：

名称	值
hive.support.concurrency	true
hive.exec.dynamic.partition.mode	nonstrict
hive.txn.manager	org.apache.hadoop.hive.ql.lockmgr.DbTxnManager
hive.compactor.initiator.on	true
hive.compactor.worker.threads	1

步骤3 准备具备supergroup和Hive权限的系统用户flume_hive，安装客户端并创建所需的Hive表。

示例如下：

1. 正确安装集群客户端，例如安装目录为“/opt/client”。
2. 执行以下命令完成用户认证。

   cd /opt/client

   source bigdata_env

   kinit flume_hive
3. 执行**beeline**命令，然后执行以下建表语句。

```
create table flume_multi_type_part(id string, msg string)
partitioned by (country string, year_month string, day string)
clustered by (id) into 5 buckets
stored as orc TBLPROPERTIES('transactional'='true');
```

4. 执行**select * from** *表名*命令**;**，查询表中数据。

   此时表中数据量为0行。

步骤4 准备相关配置文件，假设下载的客户端安装包在“/opt/FusionInsight_Cluster_1_Services_ClientConfig”。

1. 从“${*客户端解压目录*}/Hive/config”目录获取以下文件：
   - hivemetastore-site.xml
   - hive-site.xml
2. 从“${*客户端解压目录*}/HDFS/config”目录下获取以下文件：

   core-site.xml
3. 在Flume实例启动的机器上创建目录，将准备好的上述文件放置在创建的目录下。

   例如：“/opt/hivesink-conf/hive-site.xml”。
4. 将“hivemetastore-site.xml”文件中的所有property配置，拷贝至“hive-site.xml”，并保证处于原有配置之前。

   因为hive内部加载有顺序。

   说明

   保证配置文件所在的目录对于Flume运行用户**omm**有读写权限。

步骤5 结果观察。

在Hive的客户端执行，**select * from** *表名***;**查看对应的数据是否已经写入到Hive表中。

----结束

参考实例

Flume配置参考示例（SpoolDir--Mem--Hive）：

```
server.sources = spool_source
server.channels = mem_channel
server.sinks = Hive_Sink

#config the source
server.sources.spool_source.type = spooldir
server.sources.spool_source.spoolDir = /tmp/testflume
server.sources.spool_source.montime =
server.sources.spool_source.fileSuffix =.COMPLETED
server.sources.spool_source.deletePolicy = never
server.sources.spool_source.trackerDir =.flumespool
server.sources.spool_source.ignorePattern = ^$
server.sources.spool_source.batchSize = 20
server.sources.spool_source.inputCharset =UTF-8
server.sources.spool_source.selector.type = replicating
```

```
server.sources.spool_source.fileHeader = false
server.sources.spool_source.fileHeaderKey = file
server.sources.spool_source.basenameHeaderKey= basename
server.sources.spool_source.deserializer = LINE
server.sources.spool_source.deserializer.maxBatchLine= 1
server.sources.spool_source.deserializer.maxLineLength= 2048
server.sources.spool_source.channels = mem_channel

#config the channel
server.channels.mem_channel.type = memory
server.channels.mem_channel.capacity =10000
server.channels.mem_channel.transactionCapacity= 2000
server.channels.mem_channel.channelfullcount= 10
server.channels.mem_channel.keep-alive = 3
server.channels.mem_channel.byteCapacity =
server.channels.mem_channel.byteCapacityBufferPercentage= 20

#config the sink
server.sinks.Hive_Sink.type = hive
server.sinks.Hive_Sink.channel = mem_channel
server.sinks.Hive_Sink.hive.metastore = thrift://${任意metastore业务IP}:21088
server.sinks.Hive_Sink.hive.hiveSite = /opt/hivesink-conf/hive-site.xml
server.sinks.Hive_Sink.hive.coreSite = /opt/hivesink-conf/core-site.xml
server.sinks.Hive_Sink.hive.metastoreSite = /opt/hivesink-conf/hivemeatastore-site.xml
server.sinks.Hive_Sink.hive.database = default
server.sinks.Hive_Sink.hive.table = flume_multi_type_part
server.sinks.Hive_Sink.hive.partition = Tag,%Y-%m,%d
server.sinks.Hive_Sink.hive.txnsPerBatchAsk= 100
server.sinks.Hive_Sink.hive.autoCreatePartitions= true
server.sinks.Hive_Sink.useLocalTimeStamp = true
server.sinks.Hive_Sink.batchSize = 1000
server.sinks.Hive_Sink.hive.kerberosPrincipal= super1
server.sinks.Hive_Sink.hive.kerberosKeytab= /opt/mykeytab/user.keytab
server.sinks.Hive_Sink.round = true
server.sinks.Hive_Sink.roundValue = 10
server.sinks.Hive_Sink.roundUnit = minute
server.sinks.Hive_Sink.serializer = DELIMITED
server.sinks.Hive_Sink.serializer.delimiter= ";"
server.sinks.Hive_Sink.serializer.serdeSeparator= ';'
server.sinks.Hive_Sink.serializer.fieldnames= id,msg
```

12.7.15 Flume 业务模型配置指导

12.7.15.1 概述

本章节适用于MRS 3.x及之后版本。

本任务旨在提供Flume常用模块的性能差异，用于指导用户进行合理的Flume业务配置，避免出现前端Source和后端Sink性能不匹配进而导致整体业务性能不达标的场景。

本任务只针对于单通道的场景进行比较说明。

12.7.15.2 业务模型配置指导

本章节适用于MRS 3.x及之后版本。

Flume业务配置及模块选择过程中，一般要求Sink的极限吞吐量需要大于Source的极限吞吐量，否则在极限负载的场景下，Source往Channel的写入速度大于Sink从Channel取出的速度，从而导致Channel频繁被写满，进而影响性能表现。

Avro Source和Avro Sink一般都是成对出现，用于多个Flume Agent间进行数据中转，因此一般场景下Avro Source和Avro Sink都不会成为性能瓶颈。

模块间性能

根据模块间极限性能对比，可以看到对于前端是SpoolDir Source的场景下，Kafka Sink和HDFS Sink都能满足吞吐量要求，但是HBase Sink由于自身写入性能较低的原因，会成为性能瓶颈，会导致数据都积压在Channel中。但是如果有必须使用HBase Sink或者其他性能容易成为瓶颈的Sink的场景时，可以选择使用**Channel Selector**或者**Sink Group**来满足性能要求。

Channel Selector

Channel Selector可以允许一个Source对接多个Channel，通过选择不同的Selector类型来将Source的数据进行分流或者复制，目前Flume提供的Channel Selector有两种：Replicating和Multiplexing。

Replicating：表示Source的数据同步发送给所有Channel。

Multiplexing：表示根据Event中的Header的指定字段的值来进行判断，从而选择相应的Channel进行发送，从而起到根据业务类型进行分流的目的。

- Replicating配置样例：

```
client.sources = kafkasource
client.channels = channel1 channel2
client.sources.kafkasource.type = org.apache.flume.source.kafka.KafkaSource
client.sources.kafkasource.kafka.topics = topic1,topic2
client.sources.kafkasource.kafka.consumer.group.id = flume
client.sources.kafkasource.kafka.bootstrap.servers = 10.69.112.108:21007
client.sources.kafkasource.kafka.security.protocol = SASL_PLAINTEXT
client.sources.kafkasource.batchDurationMillis = 1000
client.sources.kafkasource.batchSize = 800
client.sources.kafkasource.channels = channel1 c el2

client.sources.kafkasource.selector.type = replicating
client.sources.kafkasource.selector.optional = channel2
```

表 12-153 Replicating 配置样例参数说明

选项名称	默认值	描述
Selector.type	replicating	Selector类型，应配置为replicating
Selector.optional	-	可选Channel，可以配置为列表

- Multiplexing配置样例：

```
client.sources = kafkasource
client.channels = channel1 channel2
client.sources.kafkasource.type = org.apache.flume.source.kafka.KafkaSource
client.sources.kafkasource.kafka.topics = topic1,topic2
client.sources.kafkasource.kafka.consumer.group.id = flume
client.sources.kafkasource.kafka.bootstrap.servers = 10.69.112.108:21007
client.sources.kafkasource.kafka.security.protocol = SASL_PLAINTEXT
client.sources.kafkasource.batchDurationMillis = 1000
client.sources.kafkasource.batchSize = 800
client.sources.kafkasource.channels = channel1 channel2

client.sources.kafkasource.selector.type = multiplexing
client.sources.kafkasource.selector.header = myheader
client.sources.kafkasource.selector.mapping.topic1 = channel1
client.sources.kafkasource.selector.mapping.topic2 = channel2
client.sources.kafkasource.selector.default = channel1
```

表 12-154 Multiplexing 配置样例参数说明

选项名称	默认值	描述
Selector.type	replicating	Selector类型，应配置为multiplexing
Selector.header	Flume.selector.header	-
Selector.default	-	-
Selector.mapping.*	-	-

Multiplexing类型的Selector的样例中，选择Event中Header名称为topic的字段来进行判断，当Header中topic字段的值为topic1时，向channel1发送该Event，当Header中topic字段的值为topic2时，向channel2发送该Event。

这种Selector需要借助Source中Event的特定Header来进行Channel的选择，需要根据业务场景选择合理的Header来进行数据分流。

SinkGroup

当后端单Sink性能不足、需要高可靠性保证或者异构输出时可以使用Sink Group来将指定的Channel和多个Sink对接，从而满足相应的使用场景。目前Flume提供了两种Sink Processor用于对Sink Group中的Sink进行管理：Load Balancing和Failover。

Failover：表示在Sink Group中同一时间只有一个Sink处于活跃状态，其他Sink作为备份处于非活跃状态，当活跃状态的Sink故障时，根据优先级从非活跃状态的Sink中选择一个来接管业务，保证数据不会丢失，多用于高可靠性场景。

Load Balancing：表示在Sink Group中所有Sink都处于活跃状态，每个Sink都会从Channel中去获取数据并进行处理，并且保证在运行过程中该Sink Group的所有Sink的负载是均衡的，多用于性能提升场景。

- Load Balancing配置样例：

```
client.sources = source1
client.sinks = sink1 sink2
client.channels = channel1

client.sinkgroups = g1
client.sinkgroups.g1.sinks = sink1 sink2
client.sinkgroups.g1.processor.type = load_balance
client.sinkgroups.g1.processor.backoff = true
client.sinkgroups.g1.processor.selector = random

client.sinks.sink1.type = logger
client.sinks.sink1.channel = channel1

client.sinks.sink2.type = logger
client.sinks.sink2.channel = channel1
```

表 12-155 Load Balancing 配置样例参数说明

选项名称	默认值	描述
sinks	-	Sink Group的sink列表，多个以空格分隔

选项名称	默认值	描述
processor.type	default	Processor的类型，应配置为load_balance
processor.backoff	false	是否以指数的形式退避失败的Sinks
processor.selector	round_robin	选择机制。必须是round_robin,random或者自定义的类，且该类继承了AbstractSinkSelector
processor.selector.maxTimeOut	30000	屏蔽故障sink的时间，默认是30000毫秒

- Failover配置样例：

```
client.sources = source1
client.sinks = sink1 sink2
client.channels = channel1

client.sinkgroups = g1
client.sinkgroups.g1.sinks = sink1 sink2
client.sinkgroups.g1.processor.type = failover
client.sinkgroups.g1.processor.priority.sink1 = 10
client.sinkgroups.g1.processor.priority.sink2 = 5
client.sinkgroups.g1.processor.maxpenalty = 10000

client.sinks.sink1.type = logger
client.sinks.sink1.channel = channel1

client.sinks.sink2.type = logger
client.sinks.sink2.channel = channel1
```

表 12-156 Failover 配置样例参数说明

选项名称	默认值	描述
sinks	-	Sink Group的sink列表，多个以空格分隔
processor.type	default	Processor的类型，应配置为failover
processor.priority.<sinkName>	-	优先级值。<sinkName>必须是sinks中有定义的。优先级值高Sink会更早被激活。值越大，优先级越高。**注**：多个sinks的话，优先级的值不要相同，如果优先级相同的话，只会有一个生效。
processor.maxpenalty	30000	失败的Sink最大的退避时间(单位：毫秒)

Interceptors

Flume的拦截器（Interceptor）支持在数据传输过程中修改或丢弃传输的基本单元Event。用户可以通过在配置中指定Flume内建拦截器的类名列表，也可以开发自定义的拦截器来实现Event的修改或丢弃。Flume内建支持的拦截器如下表所示，本章节会选取一个较为复杂的作为示例。其余的用户可以根据需要自行配置使用。官网参考：

http://flume.apache.org/releases/content/1.9.0/FlumeUserGuide.html

说明

1. 拦截器用在Flume的Source、Channel之间，大部分的Source都带有Interceptor参数。用户可以依据需要配置。
2. Flume支持一个Source配置多个拦截器，各拦截器名称用空格分开。
3. 指定拦截器的顺序就是它们被调用的顺序。
4. 使用拦截器在Header中插入的内容，都可以在Sink中读取并使用。

表 12-157 Flume 内建支持的拦截器类型

拦截器类型	简要描述
Timestamp Interceptor	该拦截器会在Event的Header中插入一个时间戳。
Host Interceptor	该拦截器会在Event的Header中插入当前Agent所在节点的IP或主机名。
Remove Header Interceptor	该拦截器会依据Header中包含的符合正则匹配的字符串，丢弃掉对应的Event。
UUID Interceptor	该拦截器会为每个Event的Header生成一个UUID字符串。
Search and Replace Interceptor	该拦截器基于Java正则表达式提供简单的基于字符串的搜索和替换功能。与Java Matcher.replaceAll() 的规则相同。
Regex Filtering Interceptor	该拦截器通过将Event的Body体解释为文本文件，与配置的正则表达式进行匹配来选择性的过滤Event。提供的正则表达式可用于排除或包含事件。
Regex Extractor Interceptor	该拦截器使用正则表达式抽取原始events中的内容，并将该内容加入events的header中。

下面以Regex Filtering Interceptor 为例说明Interceptor使用(其余的可参考官网配置)：

表 12-158 Regex Filtering Interceptor 配置参数说明

选项名称	默认值	描述
type	-	组件类型名称，必须写为regex_filter。
regex	-	用于匹配事件的正则表达式。

选项名称	默认值	描述
excludeEvents	false	默认收集匹配到的Event。设置为true，则会删除匹配的Event，保留不匹配的。

配置示例(为了方便观察，此模型使用了netcat tcp作为Source源，logger作为Sink)。配置好如下参数后，在Linux的配置的主机节点上执行Linux命令“telnet *主机名或IP* 44444”，并任意敲入符合正则和不符合正则的字符串。会在日志中观察到，只有匹配到的字符串被传输了。

```
#define the source、channel、sink
server.sources = r1

server.channels = c1
server.sinks = k1

#config the source
server.sources.r1.type = netcat
server.sources.r1.bind = ${主机IP}
server.sources.r1.port = 44444
server.sources.r1.interceptors= i1
server.sources.r1.interceptors.i1.type= regex_filter
server.sources.r1.interceptors.i1.regex= (flume)|(myflume)
server.sources.r1.interceptors.i1.excludeEvents= false
server.sources.r1.channels = c1

#config the channel
server.channels.c1.type = memory
server.channels.c1.capacity = 1000
server.channels.c1.transactionCapacity = 100
#config the sink
server.sinks.k1.type = logger
server.sinks.k1.channel = c1
```

12.7.16 Flume 日志介绍

日志描述

日志路径：Flume相关日志的默认存储路径为“/var/log/Bigdata/角色名”。

- FlumeServer：“/var/log/Bigdata/flume/flume”
- FlumeClient：“/var/log/Bigdata/flume-client-n/flume”
- MonitorServer：“/var/log/Bigdata/flume/monitor”

日志归档规则：Flume日志启动了自动压缩归档功能，缺省情况下，当日志大小超过50MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件，压缩文件保留个数可以在Manager界面中配置。

表 12-159 Flume 日志列表

日志类型	日志文件名	描述
运行日志	/flume/flumeServer.log	FlumeServer运行环境信息日志。

日志类型	日志文件名	描述
	/flume/install.log	FlumeServer安装日志。
	/flume/flumeServer-gc.log.<编号>	FlumeServer进程的GC日志。
	/flume/prestartDvietail.log	Flume启动前的工作日志。
	/flume/startDetail.log	Flume进程启动工作日志。
	/flume/stopDetail.log	Flume进程停止日志。
	/monitor/monitorServer.log	MonitorServer运行环境信息日志。
	/monitor/startDetail.log	MonitorServer进程启动工作日志。
	/monitor/stopDetail.log	MonitorServer进程停止日志。
	function.log	外部函数调用日志。

日志级别

Flume提供了如表12-160所示的日志级别。

运行日志的级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-160 日志级别

日志类型	级别	描述
运行日志	FATAL	FATAL表示系统运行的致命错误信息。
	ERROR	ERROR表示系统运行的错误信息。
	WARN	WARN表示当前事件处理存在异常信息。
	INFO	INFO表示记录系统及各事件正常运行状态信息。
	DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 请参考**修改集群服务配置参数**，进入Flume的“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

----结束

说明

配置完成后即生效，不需要重启服务。

日志格式

Flume的日志格式如下所示：

表 12-161 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level><产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2014-12-12 11:54:57,316	INFO	[main]	log4j dynamic load is start.	org.apache.flume.tools.LogDynamicLoad.start(LogDynamicLoad.java:59)
	<yyyy-MM-dd HH:mm:ss,SSS><User Name><User IP><Time><Operation><Resource><Result><Detail>	2014-12-12 23:04:16,572	INFO	[SinkRunner-PollingRunner-DefaultSinkProcessor]	SRCIP=null OPERATION=close

12.7.17 Flume 客户端 Cgroup 使用指导

操作场景

该操作指导用户加入、退出Cgroup，查询Cgroup状态以及更改Cgroup cpu阈值。

本章节适用于MRS 3.x及之后版本。

操作步骤

- **加入Cgroup**

  执行以下命令，加入Cgroup，假设Flume客户端安装路径为“/opt/FlumeClient”，Cgroup cpu阈值设置为50%：

  cd /opt/FlumeClient/fusioninsight-flume-1.9.0/bin

  ./flume-manage.sh cgroup join 50

说明

- 该命令不仅可以加入Cgroup，同时也可以更改Cgroup cpu阈值。
- Cgroup cpu阈值取值范围为1~100*N之间的整数，N表示机器cpu核数。

- **查询Cgroup状态**

  执行以下命令，查询Cgroup状态，假设Flume客户端安装路径为“/opt/FlumeClient”：

  cd /opt/FlumeClient/fusioninsight-flume-1.9.0/bin

  ./flume-manage.sh cgroup status

- **退出Cgroup**

  执行以下命令，退出Cgroup，假设Flume客户端安装路径为“/opt/FlumeClient”：

  cd /opt/FlumeClient/fusioninsight-flume-1.9.0/bin

  ./flume-manage.sh cgroup exit

说明

- 客户端安装完成后，会自动创建默认Cgroup。若安装客户端时未配置“-s”参数，则默认值为“-1”，表示agent进程不受cpu使用率限制。
- 加入、退出Cgroup时，agent进程不受影响。若agent进程未启动，加入、退出Cgroup仍然可以成功执行，待下一次agent启动时生效。
- 客户端卸载完成后，安装时期创建的Cgroup会自动删除。

12.7.18 Flume 第三方插件二次开发指导

操作场景

该操作指导用户进行第三方插件二次开发。

本章节适用于MRS 3.x及之后版本。

前提条件

- 第三方jar包。
- 已成功安装Flume服务端或者客户端。

操作步骤

步骤1 将自主研发的代码打成jar包。

步骤2 建立插件目录布局。

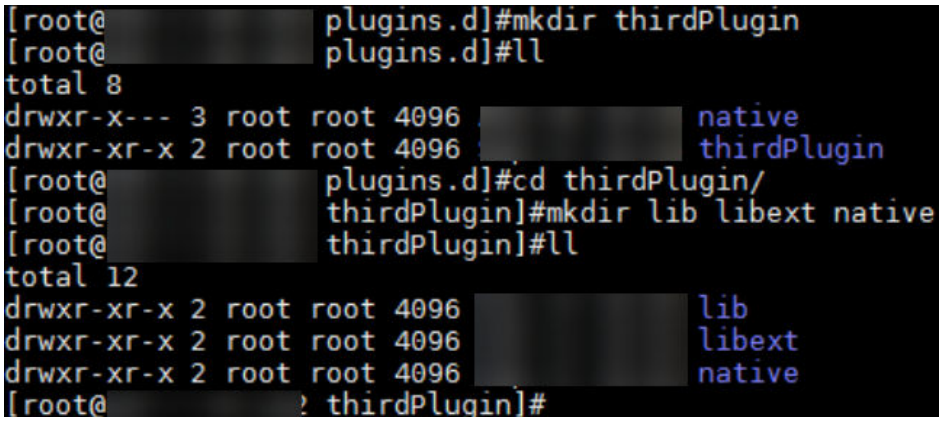
1. 进入$FLUME_HOME/plugins.d路径下，使用以下命令建立目录：

   mkdir thirdPlugin

   cd thirdPlugin

   mkdir lib libext native

   显示结果如下：



2. 将第三方jar包放入**$FLUME_HOME/plugins.d/thirdPlugin/lib**路径下，若该jar包依赖其他jar包，则将所依赖的jar包放入**$FLUME_HOME/plugins.d/thirdPlugin/libext**文件夹中，**$FLUME_HOME/plugins.d/thirdPlugin/native**放置本地库文件。

步骤3 配置**$FLUME_HOME/conf/properties.properties**文件。

具体**properties.properties**参数配置方法，参考**非加密传输**和**加密传输**对应典型场景中**properties.properties**文件参数列表的说明。

说明

- **$FLUME_HOME**表示Flume安装路径，配置第三方插件时，根据实际情况（服务端/客户端）指定。
- **thirdPlugin**根据实际业务进行命名，无固定名称。

----结束

12.7.19 Flume 常见问题

Flume日志保存在/var/log/Bigdata/flume/flume/flumeServer.log 里。绝大多数数据传输异常、数据传输不成功，在日志里都可以看到提示。可以直接输入以下命令查看：

tailf /var/log/Bigdata/flume/flume/flumeServer.log

- 问题：当配置文件上传后，发现异常，重新上传配置文件，发现仍然没有满足场景要求，但日志上没有任何异常。

  解决方法：重启此flume进程，**kill -9** *进程代码*，再看日志。
- 问题：连接HDFS出现java.lang.IllegalArgumentException: Keytab is not a readable file: /opt/test/conf/user.keytab。

  解决方法：添加Flume运行用户读写权限。
- 问题：执行Flume客户端连接Kafka报如下错误：

  ```
  Caused by: java.io.IOException: /opt/FlumeClient/fusioninsight-flume-1.9.0/cof//jaas.conf (No such file or directory)
  ```

  解决方法：新增jaas.conf配置文件并保存到flume client的conf路径下。

  vi jaas.conf

  ```
  KafkaClient {
  com.sun.security.auth.module.Krb5LoginModule required
  useKeyTab=true
  keyTab="/opt/test/conf/user.keytab"
  principal="flume_hdfs@<系统域名>"
  ```

```
useTicketCache=false
storeKey=true
debug=true;
};
```

参数keyTab和principal根据实际情况修改。

- 问题：执行Flume客户端连接HBase报如下错误：

```
Caused by: java.io.IOException: /opt/FlumeClient/fusioninsight-flume-1.9.0/cof//jaas.conf (No such file or directory)
```

解决方法：新增jaas.conf配置文件并保存到flume client的conf路径下。

vi jaas.conf

```
Client {
com.sun.security.auth.module.Krb5LoginModule required
useKeyTab=true
keyTab="/opt/test/conf/user.keytab"
principal="flume_hbase@<系统域名>"
useTicketCache=false
storeKey=true
debug=true;
};
```

参数keyTab和principal根据实际情况修改。

- 问题：一旦提交配置文件后，flume agent即在占用资源运行，如何恢复到没有上传配置文件的状态?

  解决方法：提交一个内容为空的properties.properties文件。

12.8 使用 HBase

12.8.1 从零开始使用 HBase

HBase是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。本章节提供从零开始使用HBase的操作指导，在集群Master节点中更新客户端，通过客户端实现创建表，往表中插入数据，修改表，读取表数据，删除表中数据以及删除表的功能。

背景信息

假定用户开发一个应用程序，用于管理企业中的使用A业务的用户信息，使用HBase客户端实现A业务操作流程如下：

- 创建用户信息表user_info。
- 在用户信息中新增用户的学历、职称信息。
- 根据用户编号查询用户姓名和地址。
- 根据用户姓名进行查询。
- 用户销户，删除用户信息表中该用户的数据。
- A业务结束后，删除用户信息表。

表 12-162 用户信息

编号	姓名	性别	年龄	地址
12005000201	A	男	19	A城市
12005000202	B	女	23	B城市

编号	姓名	性别	年龄	地址
12005000203	C	男	26	C城市
12005000204	D	男	18	D城市
12005000205	E	女	21	E城市
12005000206	F	男	32	F城市
12005000207	G	女	29	G城市
12005000208	H	女	30	H城市
12005000209	I	男	26	I城市
12005000210	J	男	25	J城市

前提条件

已安装客户端，例如安装目录为“/opt/client”。以下操作的客户端目录只是举例，请根据实际安装目录修改。在使用客户端前，需要先下载并更新客户端配置文件，确认Manager的主管理节点后才能使用客户端。

操作步骤

MRS 3.x以前版本集群执行以下操作：

步骤1 下载客户端配置文件。

1. 登录MRS Manager页面，具体请参见**访问集群Manager**，然后选择“服务管理”。
2. 单击“下载客户端”。

   “客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤2 登录MRS Manager的主管理节点。

1. 在集群详情的“节点信息”页签中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。

   MRS Manager的主备管理节点默认安装在集群Master节点上。在主备模式下，由于Master1和Master2之间会切换，Master1节点不一定是MRS Manager的主管理节点，需要在Master1节点中执行命令，确认MRS Manager的主管理节点。命令请参考**步骤2.4**。
2. 以**root**用户使用密码方式登录Master1节点。
3. 切换至**omm**用户。

   sudo su - root

   su - omm
4. 执行以下命令确认MRS Manager的主管理节点。

   sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh

回显信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“mgtomsdat-sh-3-01-1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“mgtomsdat-sh-3-01-2”为备管理节点）。

```
Ha mode
double
NodeName            HostName                HAVersion         StartTime             HAActive
HAAllResOK          HARunPhase
192-168-0-30        mgtomsdat-sh-3-01-1     V100R001C01       2021-11-18 23:43:02
active              normal                  Actived
192-168-0-24        mgtomsdat-sh-3-01-2     V100R001C01       2021-11-21 07:14:02
standby             normal                  Deactived
```

5. 使用**root**用户登录MRS Manager的主管理节点，例如“192-168-0-30”节点，并执行以下命令切换到**omm**用户。

   sudo su - omm

步骤3 执行以下命令切换到客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，更新主管理节点的客户端配置。

sh refreshConfig.sh /opt/client *客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/client /tmp/MRS-client/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

步骤5 在Master节点使用客户端。

1. 在已更新客户端的主管理节点，例如“192-168-0-30”节点，执行以下命令切换到客户端目录。

   cd /opt/client

2. 执行以下命令配置环境变量。

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit MRS 集群用户

   例如，**kinit hbaseuser**。

4. 直接执行HBase组件的客户端命令。

   hbase shell

步骤6 运行HBase客户端命令，实现A业务。

1. 根据**表12-162**创建用户信息表user_info并添加相关数据。

   create '*user_info*',{**NAME** => 'i'}

   以增加编号12005000201的用户信息为例，其他用户信息参照如下命令依次添加：

   put '*user_info*','*12005000201*','***i:name***','*A*'

   put '*user_info*','*12005000201*','***i:gender***','*Male*'

put '*user_info*','*12005000201*','***i:age***','*19*'

put '*user_info*','*12005000201*','***i:address***','*City A*'

2. 在用户信息表user_info中新增用户的学历、职称信息。

   以增加编号为12005000201的用户的学历、职称信息为例，其他用户类似。

   put '*user_info*','*12005000201*','**i:degree**','*master*'

   put '*user_info*','*12005000201*','**i:pose**','*manager*'

3. 根据用户编号查询用户姓名和地址。

   以查询编号为12005000201的用户姓名和地址为例，其他用户类似。

   scan'*user_info*',
   {**STARTROW**=>'*12005000201*',**STOPROW**=>'*12005000201*',**COLUMNS**=>['**i:name**','**i:address**']}

4. 根据用户姓名进行查询。

   以查询A用户信息为例，其他用户类似。

   scan'*user_info*',{**FILTER**=>"SingleColumnValueFilter('i','name',=,'binary:*A*')"}

5. 删除用户信息表中该用户的数据。

   所有用户的数据都需要删除，以删除编号为12005000201的用户数据为例，其他用户类似。

   delete'*user_info*','*12005000201*','i'

6. 删除用户信息表。

   disable'*user_info*'

   drop '*user_info*'

----结束

MRS 3.x及之后版本集群执行以下操作：

步骤1 在主管理节点使用客户端。

1. 以客户端安装用户登录客户端安装节点，执行以下命令切换到客户端目录。

   cd /opt/client

2. 执行以下命令配置环境变量。

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit hbaseuser**。

4. 直接执行HBase组件的客户端命令。

   hbase shell

步骤2 运行HBase客户端命令，实现A业务。

1. 根据**表12-162**创建用户信息表user_info并添加相关数据。

   create '*user_info*',{**NAME** => 'i'}

   以增加编号12005000201的用户信息为例，其他用户信息参照如下命令依次添加：

   put '*user_info*','*12005000201*','***i:name***','*A*'

put '*user_info*','*12005000201*','***i:gender***','*Male*'

put '*user_info*','*12005000201*','***i:age***','*19*'

put '*user_info*','*12005000201*','***i:address***','*City A*'

2. 在用户信息表user_info中新增用户的学历、职称信息。

   以增加编号为12005000201的用户的学历、职称信息为例，其他用户类似。

   put '*user_info*','*12005000201*','**i:degree**','*master*'

   put '*user_info*','*12005000201*','**i:pose**','*manager*'

3. 根据用户编号查询用户姓名和地址。

   以查询编号为12005000201的用户姓名和地址为例，其他用户类似。

   scan'*user_info*',
   {**STARTROW**=>'*12005000201*',**STOPROW**=>'*12005000201*',**COLUMNS**=>['**i:name**','**i:address**']}

4. 根据用户姓名进行查询。

   以查询A用户信息为例，其他用户类似。

   scan'*user_info*',{**FILTER**=>"SingleColumnValueFilter('i','name',=,'binary:*A*')"}

5. 删除用户信息表中该用户的数据。

   所有用户的数据都需要删除，以删除编号为12005000201的用户数据为例，其他用户类似。

   delete'*user_info*','*12005000201*','i'

6. 删除用户信息表。

   disable'*user_info*'

   drop '*user_info*'

----结束

12.8.2 使用 HBase 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用HBase客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。

  “机机”用户需要下载keytab文件，“人机”用户第一次登录时需修改密码。
- 非**root**用户使用HBase客户端，请确保该HBase客户端目录的属主为该用户，否则请参考如下命令修改属主。

  chown user:group -R *客户端安装目录***/HBase**

使用 Hbase 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令切换到客户端目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

步骤5 直接执行HBase组件的客户端命令。

hbase shell

----结束

使用 HBase 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令切换到客户端目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若安装了HBase多实例，在使用客户端连接具体HBase实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载HBase2实例变量：

source HBase2/component_env

步骤5 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

步骤6 直接执行HBase组件的客户端命令。

hbase shell

----结束

HBase 客户端常用命令

常用的HBase客户端命令如下表所示。更多命令可参考**http://hbase.apache.org/2.2/book.html**

表 12-163 HBase 客户端命令

命令	说明
create	创建一张表，例如**create 'test', 'f1', 'f2', 'f3'**。

命令	说明
disable	停止指定的表，例如**disable 'test'**。
enable	启动指定的表，例如**enable 'test'**。
alter	更改表结构。可以通过alter命令增加、修改、删除列族信息以及表相关的参数值，例如**alter 'test', {NAME => 'f3', METHOD => 'delete'}**。
describe	获取表的描述信息，例如**describe 'test'**。
drop	删除指定表。删除前表必须已经是停止状态，例如**drop 'test'**。
put	写入指定cell的value。Cell的定位由表、rowk、列组合起来唯一决定，例如**put 'test','r1','f1:c1','myvalue1'**。
get	获取行的值或者行的指定cell的值。例如**get 'test','r1'**。
scan	查询表数据。参数中指定表名和scanner，例如**scan 'test'**。

12.8.3 创建 HBase 角色

操作场景

该任务指导系统管理员在Manager创建并设置HBase的角色。HBase角色可设置HBase管理员权限以及HBase表和列族的读（R）、写（W）、创建（C）、执行（X）或管理（A）权限。

用户需要在HBase中对指定的数据库或表设置权限，才能够创建表、查询数据、删除数据、插入数据、更新数据以及授权他人访问HBase表。

说明

- 本章节适用于MRS 3.x及之后版本。
- 安全模式支持创建HBase角色，普通模式不支持创建HBase角色。
- 如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加HBase的Ranger访问权限策略**。

前提条件

- 系统管理员已明确业务需求。
- 已登录Manager。

操作步骤

步骤1 在Manager界面，选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

步骤3 设置角色“配置资源权限”请参见**表12-164**。

HBase权限：

- HBase Scope：对HBase表授权，最小支持设置列的读（R）和写（W）权限。
- HBase管理员权限：HBase管理员权限。

说明

用户对自己创建的表具有读（R）、写（W）、创建（C）、执行（X）或管理（A）权限。

表 12-164 设置角色

任务场景	角色授权操作
设置HBase管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase”，勾选“HBase管理员权限”。
设置用户创建表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope”。 2. 单击“global”。 3. 在指定命名空间的“权限”列，勾选“创建”和“执行”。例如勾选默认命名空间“default”的“创建”和“执行”。
设置用户写入数据的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global”。 2. 在指定命名空间的“权限”列，勾选“写”。例如勾选默认命名空间“default”的“写”。HBase子对象默认可从父对象继承权限，此时已授予向命名空间中的表写入数据的权限。
设置用户读取数据的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global”。 2. 在指定命名空间的“权限”列，勾选“读”。例如勾选默认命名空间“default”的“读”。HBase子对象默认可从父对象继承权限，此时已授予从命名空间中的表读取数据的权限。
设置用户管理命名空间或表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global”。 2. 在指定命名空间的“权限”列，勾选“管理”。例如勾选默认命名空间“default”的“管理”。

任务场景	角色授权操作
设置列的读取或写入权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global”，单击指定命名空间显示命名空间的表。 2. 单击指定的表。 3. 单击指定的列族。 4. 确认是否是新建角色? – 是，在“资源名称”的输入框输入列名称，多个列用英文逗号分隔，勾选“读”或“写”。如果HBase表中不存在同名的列，则创建同名的列后角色将拥有该列的权限。列权限设置完成。 – 否，修改已有HBase角色的列权限，表格将显示已单独设置权限的列，执行**步骤3.5**。 5. 角色新增列权限，在“资源名称”的输入框输入列名称并设置列的权限。角色修改列权限，可以在“资源名称”的输入框输入列名称并设置列权限，也可以在表格中直接修改列的权限。若在表格中修改了列权限，又同时增加了同名的列权限，则无法保存。角色修改列权限，建议直接修改列的权限。支持搜索功能。

步骤4 单击“确定”完成，返回“角色”。

----结束

12.8.4 配置 HBase 备份

操作场景

HBase集群备份作为提高HBase集群系统高可用性的一个关键特性，为HBase提供了实时的异地数据备份功能。它对外提供了基础的运维工具，包含主备关系维护、重建，数据校验，数据同步进展查看等功能。为了实现数据的实时备份，可以把本HBase集群中的数据备份到另一个集群。

前提条件

- 主备集群都已经安装并启动成功（在Console页面“现有集群”页签，查看集群状态为“运行中”），且获取集群的管理员权限。
- 必须保证主备集群间的网络畅通和端口的使用。
- 主备集群必须已配置跨集群互信。
- 如果主集群上有历史数据，需要同步到备集群上，那么主备集群必须配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 必须在主备集群中的“/etc/hosts”文件中，配置**主备集群所有机器**的机器名与业务IP地址的对应关系。配置方式为在hosts文件中追加"192.***.***.*** host1"。

- 主备集群间的网络带宽需要根据业务流量而定，不应少于最大的可能业务流量。

使用约束

- 尽管备份提供了实时的数据复制功能，但实际的数据同步进展，由多方面的因素决定的，例如，当前主集群业务的繁忙程度，备集群进程的健康状态等。因此，在正常情形下，备集群不应该接管业务。极端情形下是否可以接管业务，可由系统维护人员以及决策人员根据当前的数据同步指标来决定。
- 备份功能当前仅支持一主一备。
- 通常情况下，不允许对备集群的同步表进行表级别的操作，例如修改表属性、删除表等，一旦误操作备集群后会造成主集群数据同步失败、备集群对应表的数据丢失。
- 主集群的HBase表已启用备份功能同步数据，用户每次修改表的结构时，需要手动修改备集群的同步表结构，保持与主集群表结构一致。

操作步骤

启用主集群的备份功能来同步put方式写入的数据

步骤1 登录MRS控制台，单击集群名称，选择“组件管理”。

步骤2 进入HBase服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤3 （可选）如**表12-165**所示，为HBase备份操作过程中的可选配置项，您可以根据描述来进行参数配置，或者使用缺省提供的值。

表 12-165 可选配置项

配置入口	配置项	默认值	描述
“HMaster > 性能”	hbase.master.logcleaner.ttl	600000	HLog文件生存时间。如果配置值为“604800000”（单位：毫秒），表示HLog的保存期限为7天。
	hbase.master.cleaner.interval	60000	HMaster清理过去HLog文件的周期，即超过设置的时间的HLog会被自动删除。建议尽可能配置大的值来保留更多的HLog。
“RegionServer > Replication”	replication.source.size.capacity	16777216	edits最大大小。单位为byte。如果edit大小超过这个值Hlog edits将会发送到备集群。

配置入口	配置项	默认值	描述
	replication.source.nb.capacity	25000	edits最大数目，是另一个触发Hlog edits到备集群的条件。当主集群同步数据到备集群中时，主集群会从HLog中读取数据，此时会根据本参数配置的个数读取并发送。与“replication.source.size.capacity”一起配置使用。
	replication.source.maxretriesmultiplier	10	replication出现异常时的最大重试次数。
	replication.source.sleepforretries	1000	每次重试的sleep时间。（单位：毫秒）
	hbase.regionserver.replication.handler.count	6	RegionServer上的replication RPC服务器实例数。

启用主集群备份功能来同步Bulkload方式写入的数据

步骤4 是否启用Bulkload写数据备份功能?

说明

当使用了HBase 的Bulkload导入数据的特性时且需要同步这些数据时，需要开启批量写数据备份功能。

是，执行**步骤5**。

否，执行**步骤9**。

步骤5 参考**修改集群服务配置参数**进入HBase服务参数“全部配置”界面。

步骤6 在主、备集群的HBase配置界面，搜索并修改“hbase.replication.cluster.id”参数，表示主、备集群HBase的id，例如主集群HBase的id配置为“replication1”，备集群HBase的id配置为“replication2”，用于主备集群的连接。为了节省数据开销建议参数值长度不超过30。

步骤7 在备集群的HBase配置界面，搜索并修改“hbase.replication.conf.dir”参数，表示备集群中所使用主集群客户端的HBase配置，用于启用bulkload数据备份功能时的数据备份。参数值为路径名，例如“/home”。

说明

- MRS 3.x之前的版本无需配置此参数，可跳过步骤**步骤7**。
- 当启用bulkload数据备份功能时，需在备集群的所有RegionServer节点上手动放置主集群中HBase相应客户端配置文件(core-site.xml, hdfs-site.xml, hbase-site.xml)，放置配置文件的实际路径为“${hbase.replication.conf.dir}/${hbase.replication.cluster.id}”。例如备集群的hbase.replication.conf.dir配置为“/home”，主集群的hbase.replication.cluster.id配置为“replication1”，则配置文件放置在备集群中实际的路径为“/home/replication1”。并修改对应目录及文件相应权限，可执行如下命令**chown -R omm:wheel /home/replication1**。
- 客户端配置文件可从主集群中的的客户端中获取，例如，路径为“/opt/client/HBase/hbase/conf”。

步骤8 在主集群的HBase配置界面，搜索并修改“hbase.replication.bulkload.enabled”参数，将配置项的值修改为“true”，启用Bulkload写数据备份功能。

重启HBase服务并安装客户端。

步骤9 保存配置，并重启HBase服务。

步骤10 在主备集群，更新客户端配置文件。

同步主集群表数据。（主集群无数据可不执行）

步骤11 以“hbase”用户进入集群的HBase shell界面。

1. 在已更新客户端的主管理节点，执行以下命令切换到客户端目录。

   cd /opt/client

2. 执行以下命令配置环境变量。

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit hbase

   说明

   执行**kinit hbase**后系统提示输入密码，hbase用户默认密码是Hbase@123。

4. 直接执行HBase组件的客户端命令。

   hbase shell

步骤12 检查备集群上是否已有历史数据。如果有历史数据且需要保持主备集群上的数据完全一致，需要先清理备集群上的数据。

1. 在备集群的hbase shell界面中，执行**list**命令查看备集群中已经存在的表。
2. 根据输出列表删除备集群上的数据表。

   disable '*tableName*'

   drop '*tableName*'

步骤13 检查配置HBase备份并启用数据同步后，主集群是否已存在表及数据，且历史数据需要同步到备集群。

执行**list**命令查看主集群中已经存在的表，使用**scan** '*tableName*'命令查看表中是否已经有历史数据。

- 是，存在表且需要同步数据，执行**步骤14**。
- 否，不需要同步数据，任务结束。

步骤14 配置HBase备份时不支持自动同步表中的历史数据，需要对主集群的历史数据进行备份，然后再手动同步历史数据到备集群中。

手动同步即单表的同步，单表手动同步通过Export、distcp、Import来完成。

单表手动同步操作步骤：

1. 从主集群导出表中数据。

   hbase org.apache.hadoop.hbase.mapreduce.Export -Dhbase.mapreduce.include.deleted.rows=true *表名 保存源数据的目录*

   例如，**hbase org.apache.hadoop.hbase.mapreduce.Export -Dhbase.mapreduce.include.deleted.rows=true t1 /user/hbase/t1**

2. 把导出的数据复制到备集群。

   hadoop distcp *主集群保存源数据的目录 hdfs://ActiveNameNodeIP:9820/备集群保存源数据的目录*

   其中，ActiveNameNodeIP是备集群中主NameNode节点的IP地址。

   例如，**hadoop distcp /user/hbase/t1 hdfs://192.168.40.2:9820/user/hbase/t1**

3. 使用备集群HBase表用户，在备集群中导入数据。

   hbase org.apache.hadoop.hbase.mapreduce.Import *-Dimport.bulk.output=备集群保存输出的目录 表名 备集群保存源数据的目录*

   hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles *备集群保存输出的目录 表名*

   例如，**hbase org.apache.hadoop.hbase.mapreduce.Import -Dimport.bulk.output=/user/hbase/output_t1 t1 /user/hbase/t1**

   hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /user/hbase/output_t1 t1

添加主备集群备份关系。

步骤15 在HBase shell中执行如下命令，创建主集群HBase与备集群HBase之间的备份同步关系。

add_peer '*备集群ID', CLUSTER_KEY => '备集群ZooKeeper地址信息*',**{HDFS_CONFS => true}**

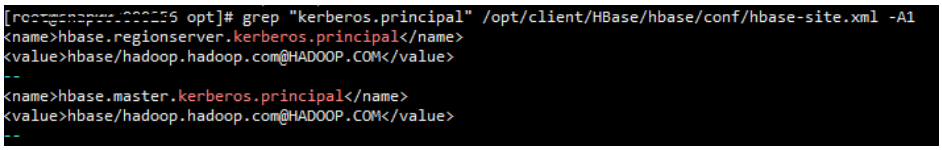
- 备集群ID表示主集群识别备集群使用的id，建议使用字母与数字。
- 备集群ZooKeeper地址信息包含ZooKeeper业务IP地址、侦听客户端连接的端口和备集群的HBase在ZooKeeper上的根目录。
- **{HDFS_CONFS => true}**表示将主集群的默认HDFS配置信息同步到对应集群，用于备集群的HBase访问主集群的HDFS。如果不启用Bulkload批量写数据备份，可以不使用此参数。

  例如，添加包含BulkLoad数据的主备集群备份关系，若备集群ID为“replication2”，备集群ZooKeeper地址信息为“**192.168.40.2,192.168.40.3,192.168.40.4:2181:/hbase**”。

  - 安全集群请执行：**add_peer 'replication2',CLUSTER_KEY => '192.168.40.2,192.168.40.3,192.168.40.4:2181:/hbase',CONFIG => { "hbase.regionserver.kerberos.principal" => "<val>", "hbase.master.kerberos.principal" => "<val2>" }**，普通集群请执行**add_peer 'replication2',CLUSTER_KEY => '192.168.40.2,192.168.40.3,192.168.40.4:2181:/hbase'**

    其中参数“hbase.master.kerberos.principal”和“hbase.regionserver.kerberos.principal”为安全集群中hbase的kerberos用户，可搜索客户端中hbase-site.xml文件得到参数值。例如，客户端安装在master节点的“/opt/client”下，则可使用命令**grep "kerberos.principal" /opt/client/HBase/hbase/conf/hbase-site.xml -A1**获取，如下图所示。

    图 12-14 获取 hbase 的 principal

说明

1. 获取ZooKeeper业务IP地址。

   登录MRS控制台，单击集群名称，选择“组件管理 > ZooKeeper > 实例”，获取ZooKeeper业务IP地址。
2. 在ZooKeeper服务参数“全部配置”界面，搜索获取clientPort，即为客户端连接服务器的端口。
3. 执行**list_peers**命令判断主备备份关系添加结果，当界面提示以下信息表示成功。

   ```
   hbase(main):003:0> list_peers
   PEER_ID CLUSTER_KEY ENDPOINT_CLASSNAME STATE REPLICATE_ALL NAMESPACES
   TABLE_CFS BANDWIDTH SERIAL
   replication2 192.168.0.13,192.168.0.177,192.168.0.25:2181:/hbase ENABLED  true   0 false
   ```

指定主备集群写数据状态。

步骤16 在主集群HBase shell界面，执行以下命令保持写数据状态。

set_clusterState_active

界面提示以下信息表示执行成功：

```
hbase(main):001:0> set_clusterState_active
=> true
```

步骤17 在备集群HBase shell界面，执行以下命令保持只读数据状态。

set_clusterState_standby

界面提示以下信息表示执行成功：

```
hbase(main):001:0> set_clusterState_standby
=> true
```

启用HBase备份功能同步数据。

步骤18 检查备集群的HBase服务实例中，是否已存在一个命名空间，与待启用备份功能的HBase表所属的命名空间名称相同？

在备集群的HBase shell中，执行**list_namespace**命令，查询命名空间。

- 是，存在同名的命令空间，执行**步骤19**。
- 否，不存在同名的命令空间，需先在备集群的HBase shell中，执行**create_namespace 'ns1'**创建同名的命名空间，然后执行**步骤19**。

步骤19 在主集群的HBase shell中，执行以下命令，启用主集群表的数据实时备份功能，确保后续主集群中修改的数据能够实时同步到备集群中。

一次只能针对一个HTable进行数据同步。

enable_table_replication '*表名*'

说明

- 若备集群中不存与要开启实时同步的表同名的表，则该表会自动创建。
- 若备集群中存在与要开启实时同步的表同名的表，则两个表的结构必须一致。
- 若'表名'设置了加密算法SMS4或AES，则不支持对此HBase表启用将数据从主集群实时同步到备集群的功能。
- 若备集群不在线，或备集群中已存在同名但结构不同的表，启用备份功能将失败。

  若备集群不在线，请启动备集群。

  若备集群中已存在同名但结构不同的表，请修改备集群的表结构为相同的表结构。在备集群的HBase shell中，执行***alter***命令，参考示例修改。

步骤20 在主集群的HBase shell中，执行以下命令，启用主集群的实时备份功能，同步HBase的权限表。

enable_table_replication 'hbase:acl'

说明

主集群HBase源数据表修改权限时，如果备集群需要正常读取数据，请修改备集群角色的权限。

检验主备集群数据同步状态。

步骤21 在HBase客户端执行以下命令，校验主备集群同步的数据。启用备份同步功能后，也可以执行该命令检验新的同步数据是否一致。

hbase org.apache.hadoop.hbase.mapreduce.replication.VerifyReplication --starttime=*开始时间* **--endtime**=*结束时间 列族名称 备集群ID 表名*

说明

- 开始时间必须早于结束时间。
- 开始时间和结束时间需要填写时间戳的格式，例如执行**date -d "2015-09-30 00:00:00" +%s**将普通时间转化为时间戳格式。因此命令返回的为10位数字（精确到秒），而HBase识别的为13位（精确到毫秒），所以需要在date命令返回的结果后补上3个0。

主备集群发生倒换

说明

1. 当备集群需要被倒换为主集群时，请参见**步骤2**~**步骤10**和**步骤15**~**步骤20**重新配置主备关系。
2. 勿需执行“同步集群表数据”操作，即**步骤11**~**步骤14**。

----结束

相关命令

表 12-166 HBase 备份

操作	命令	描述
建立主备关系	**add_peer** *'备集群ID', '备集群地址信息'* 示例： **add_peer '1', 'zk1,zk2,zk3:2181:/hbase'** **add_peer '1', 'zk1,zk2,zk3:2181:/hbase1'**	建立主集群与备集群的关系，让其互相对应。如果启用Bulkload批量写数据备份，则命令为**add_peer** *'备集群ID'**,CLUSTER_KEY =>** '备集群地址信息'*，并配置参数*"hbase.replication.conf.dir"*，同时手动拷贝主集群的hbase相应客户端配置文件到备集群的所有RegionServer节点，详情请参考步骤4~11。
移除主备关系	**remove_peer** *'备集群ID'* 示例： **remove_peer '1'**	在主集群中移除备集群的信息。
查询主备关系	**list_peers**	在主集群中查询已经设置的备集群的信息，主要为Zookeeper信息。
启用用户表实时同步	**enable_table_replication** *'表名'* 示例： **enable_table_replication 't1'**	在主集群中，设置已存在的表同步到备集群。
禁用用户表实时同步	**disable_table_replication** *'表名'* 示例： **disable_table_replication 't1'**	在主集群中，设置已存在的表不同步到备集群。
主备集群数据校验	**bin/hbase org.apache.hadoop.hbase.mapreduce.replication.VerifyReplication --starttime**=*开始时间* **--endtime**=*结束时间 列族名称 备集群ID 表名*	检查指定的表在主备集群间的数据是否一致。 命令行中参数说明如下： • 开始时间：如果未设置，则取默认的开始时间为0。 • 结束时间：如果未设置，则取默认的结束时间为当前操作提交的时间。 • 表名：如果未输入表名，则默认校验所有的启用了实时同步的用户表。
切换数据写入状态	**set_clusterState_active** **set_clusterState_standby**	设置集群HBase表是否可写入数据。

操作	命令	描述
新增或更新已经在对端集群保存的主集群中HDFS配置	**set_replication_hdfs_confs 'PeerId', {'key1' => 'value1', 'key2' => 'value2'}**	启用包含Bulkload数据的备份，在主集群修改HDFS参数时，新的参数值默认不会从主集群自动同步到备集群，需要手动执行命令同步。受影响的参数如下： ● “fs.defaultFS” ● “dfs.client.failover.proxy.provider.hacluster” ● “dfs.client.failover.connection.retries.on.timeouts” ● “dfs.client.failover.connection.retries” 例如，“fs.defaultFS”修改为“hdfs://hacluster_sale”，同步HDFS配置到id为1的备集群时执行：**set_replication_hdfs_confs '1', {'fs.defaultFS' => 'hdfs://hacluster_sale'}**

12.8.5 配置 HBase 参数

说明

该章节操作仅适用于MRS 3.x之前版本集群。

当MRS服务中默认的参数配置不足以满足用户需要时，用户可以自定义修改参数配置来适应自身需求。

步骤1 登录集群详情页面，选择“组件管理”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 选择“HBase > 服务配置”，将“基础配置”切换为“全部配置”，进入HBase配置界面修改参数配置。

表 12-167 HBase 参数说明

参数	参数说明	参数值
hbase.regionserver.hfile.durable.sync	设置是否启用Hfile耐久性以将数据持久化到磁盘。若将该参数设置为true，由于每个Hfile写入HBase时都会被hadoop fsync同步到磁盘上，则HBase性能将受到影响。 该参数仅在MRS 1.9.2及之前版本存在。	取值范围： • true • false 默认值为true
hbase.regionserver.wal.durable.sync	设置是否启用WAL文件耐久性以将WAL数据持久化到磁盘。若将该参数设置为true，由于每个WAL的编辑都会被hadoop fsync同步到磁盘上，则HBase性能将受到影响。 该参数仅在MRS 1.9.2及之前版本存在。	取值范围： • true • false 默认值为true

----结束

12.8.6 启用集群间拷贝功能

操作场景

当用户需要将保存在HDFS中的数据从当前集群备份到另外一个集群时，需要使用DistCp工具。DistCp工具依赖于集群间拷贝功能，该功能默认未启用。两个集群都需要配置。

该任务指导系统管理员在MRS修改参数以启用集群间拷贝功能。

对系统的影响

启用集群间复制功能需要重启Yarn，服务重启期间无法访问。

前提条件

两个集群HDFS的参数“hadoop.rpc.protection”需使用相同的数据传输方式。设置为“privacy”表示加密，“authentication”表示不加密。

说明

参考**修改集群服务配置参数**，进入HDFS服务参数“全部配置”界面“，搜索hadoop.rpc.protection查看。

针对MRS 3.x之前版本，在集群详情页选择“组件管理 > HDFS > 服务配置”，将“基础配置”切换为“全部配置”，搜索hadoop.rpc.protection查看。

操作步骤

步骤1 进入Yarn服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 左边菜单栏中选择“Yarn > 集群间拷贝”。

步骤3 设置“dfs.namenode.rpc-address”参数的“haclusterX.remotenn1”值为对端集群其中一个NameNode实例的业务IP和RPC端口，设置“haclusterX.remotenn2”值为对端集群另外一个NameNode实例的业务IP和RPC端口。按照“*IP:port*”格式填写。

说明

针对MRS 3.x版本集群，登录FusionInsight Manager页面，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，获取NameNode实例的业务IP。

针对MRS 3.x之前版本，在集群详情页选择“组件管理 > HDFS > 实例”，获取NameNode实例的业务IP。

“dfs.namenode.rpc-address.haclusterX.remotenn1”和“dfs.namenode.rpc-address.haclusterX.remotenn2”不区分主备NameNode。NameNode RPC端口默认为“9820”，不支持通过Manager修改。

修改后参数值例如：“10.1.1.1:9820”和“10.1.1.2:9820”。

步骤4 保存配置并在概览页面选择“更多 > 重启服务”，重启Yarn服务。

界面提示“操作成功。”，单击“完成”，Yarn服务启动成功。

步骤5 登录另外一个集群，重复以上操作。

----结束

12.8.7 使用 ReplicationSyncUp 工具

前提条件

1. 主备集群已经安装并且启动。
2. 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
3. 当主集群HBase服务关闭时，Zookeeper和HDFS服务应该启动并运行。
4. 该工具应该由启动HBase进程的系统用户运行。
5. 如果处于安全模式，请确保备用集群的HBase系统用户具有主集群HDFS的读取权限。因为它将更新HBase系统Zookeeper节点和HDFS文件。
6. 主集群HBase故障后，主集群的Zookeeper，文件系统和网络依然可用。

场景介绍

Replication机制可以使用WAL将一个集群的状态与另一个集群的状态保持同步。启用HBase备份后，若主集群出现故障，ReplicationSyncUp工具会使用来自zookeeper的信息将主集群中的启用HBase备份功能的数据增量同步到备集群中。数据同步完成后，备集群可以作为主集群使用。

参数配置

参数	描述	默认值
hbase.replication.bulkload.enabled	是否开启批量加载数据复制功能。参数值类型为Boolean。开启批量加载数据复制功能后该参数须在主集群中设置为true。	false
hbase.replication.cluster.id	源HBase集群ID。开启批量加载数据复制功能是必须设置该参数，在源集群定义。参数值类型为String。	-

工具使用

在主集群client上输入如下命令使用：

hbase org.apache.hadoop.hbase.replication.regionserver.ReplicationSyncUp -Dreplication.sleep.before.failover=1

说明

replication.sleep.before.failover是指在RegionServer启动失败时备份其剩余数据前需要的休眠时间。由于30秒（默认值）的睡眠时间没有任何意义，因此将其设置为1（s），使备份过程更快触发。

注意事项

1. 当主集群关闭时，此工具将从ZooKeeper节点（RS znode）获得WAL的处理进度以及WAL的处理队列，并将未复制的队列复制到备集群中。
2. 每个主集群的RegionServer在备集群ZooKeeper上的replication节点下都有自己的znode。它包含每个对等集群的一个znode。
3. 当Regionserver故障时，主集群的每个RegionServer都会通过watcher收到通知，并尝试锁定故障RegionServer的znode，包含它的队列。成功创建的RegionServer会将所有队列转移到自己队列的znode下。队列传输后，它们将从旧位置删除。
4. 在主集群关闭期间，ReplicationSyncUp工具将使用来自ZooKeeper节点的信息同步主备集群的数据，并且RegionServer znode的wals将被移动到备集群下。

限制和约束

如果备集群处于关闭状态或关闭了对等关系，该工具正常运行，只有该对等关系复制不会发生。

12.8.8 使用 HIndex

12.8.8.1 HIndex 介绍

场景介绍

HBase是基于Key-Value的分布式存储数据库，基于rowkeys对表中的数据按照字典进行排序。如果您根据指定的rowkey查询数据，或者扫描指定rowkey范围内的数据，

HBase可以快速查找到需要读取的数据，从而提高效率。在大多数实际情况下，会需要查询列值为XXX的数据。HBase提供了Filter功能来查询具有特定列值的数据：所有数据按RowKey的顺序进行扫描，然后将数据与特定的列值进行匹配，直到找到所需的数据。过滤器功能会scan一些不必要的数据以获取所需的数据。基于前面的描述，Filter功能不能满足高性能标准频繁查询的要求。

这就是HBase HIndex产生的背景。HBase HIndex为HBase提供了能够根据特定的列值进行索引的能力，使得查询会变得更快。

说明

- 索引数据不支持滚动升级。
- 复合索引：用户必须将所有参与复合索引的列全部放入/删除，否则会导致数据不一致。
- 用户不应将任何split policy显式地配置到已建立索引的数据表中。
- 不支持mutation操作，如increment,append。
- 不支持列索引的版本maxVersions> 1。
- 添加索引的列值不应超过32KB。
- 当用户数据由于列族级TTL失效而被删除时，相应的索引数据不会立即删除。索引数据将在major compaction期间被删除。
- 创建索引后，不应更改用户列族的TTL。
  - 如果在创建索引后将列族TTL更改为更高值，则应删除并重新创建索引，否则某些已生成的索引数据可能比用户数据先删除。
  - 如果在创建索引后将列族TTL更改为较低值，则索引可能会晚于用户数据被删除。
- HBase表启动容灾之后，主集群新建二级索引，索引表变更不会自动同步到备集群。要实现该容灾场景，必须执行以下操作：
  1. 在主表创建二级索引之后，需要在备集群使用相同方法创建结构、名称完全相同的二级索引。
  2. 在主集群手动将索引列族（默认是d）的REPLICATION_SCOPE设置为1。

参数配置

1. 登录MRS控制台，单击集群名称，选择“组件管理”。
2. 进入HBase服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**。
3. 在HBase全部配置界面查看参数。

配置入口	配置项	默认值	描述
“HMaster > 系统”	hbase.coprocessor.master.classes	org.apache.hadoop.hbase.hindex.server.master.HIndexMasterCoprocessor,com.xxx.hadoop.hbase.backup.services.RecoveryCoprocessor,org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor,org.apache.hadoop.hbase.security.access.ReadOnlyClusterEnabler,org.apache.hadoop.hbase.rsgroup.RSGroupAdminEndpoint	该协处理器用于在启用Hindex功能后处理Master级的操作，比如创建索引meta表，添加索引，删除索引，删除表删除索引元数据。
“RegionServer > RegionServer”	hbase.coprocessor.regionserver.classes	org.apache.hadoop.hbase.hindex.server.regionserver.HIndexRegionServerCoprocessor,org.apache.hadoop.hbase.JMXListener,org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor	该协处理器用于在启用Hindex功能后实际上处理master下发到Regionserver上的操作。

配置入口	配置项	默认值	描述
	hbase.coprocessor.region.classes	org.apache.hadoop.hbase.hindex.server.regionserver.HIndexRegionCoprocessor,org.apache.hadoop.hbase.security.token.TokenProvider,com.xxx.hadoop.hbase.backup.services.RecoveryCoprocessor,org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor,org.apache.hadoop.hbase.security.access.SecureBulkLoadEndpoint,org.apache.hadoop.hbase.security.access.ReadOnlyClusterEnabler,org.apache.hadoop.hbase.coprocessor.MetaTableMetrics	该协处理器用于在启用Hindex功能后实际上操作Region上的数据。

配置入口	配置项	默认值	描述
	hbase.coprocessor.wal.classes	org.apache.hadoop.hbase.hindex.server.regionserver.HIndexRegionServerCoprocessor,org.apache.hadoop.hbase.JMXListener,org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor	该协处理器用于Replication，其会过滤掉索引数据以避免索引数据发送到对等集群中，对等集群中的数据索引数据将会自己生成。 该参数仅MRS 3.x之前版本支持。

说明

1.上述默认值为启用HBase HIndex功能后需额外配置的值，当前支持HBase HIndex功能的MRS集群默认已配置。

2.必须确保master参数配置在hmster上，region/regionserver参数配置在regonserver上。

相关接口

使用HIndex的API都在类org.apache.hadoop.hbase.hindex.client.HIndexAdmin中，相关接口介绍如下：

操作	接口	描述	注意事项
添加索引	addIndices()	将索引添加到没有数据的表中。调用此接口会将用户指定的索引添加到表中，但会跳过生成索引数据。因此，在此操作之后，索引不能用于scan/filter操作。它的使用场景为用户想要在具有大量预先存在用户数据的表上批量添加索引，其具体操作为使用诸如TableIndexer工具之类的外部工具来构建索引数据。	• 索引一旦添加则不能修改。若要修改，则需先删除旧的索引然后重新创建。 • 用户应注意不要在具有不同索引名称的相同列上创建两个索引。如果这样做，将会导致存储和处理的浪费。 • 索引不能添加到系统表中。 • 向索引列put数据时不支持append和increment操作。 • 如果客户端出现任何故障，除非发生DoNotRetryIOException，否则用户应该重试。 • 索引列族根据可用性按顺序从以下条件中选择： – 默认索引列族“d”或者如果设置了属性“hindex.default.family.name”的值，则以该值为准。 – 符号#，@，$或% – #0，@ 0，$ 0，%0，#1，@ 1 ...上至#255，@ 255，$ 255，%255 – throw Exception • 可以通过HIndex TableIndexer工具添加索引而无需建立索引数据。
	addIndicesWithData()	将索引添加到有数据的表中。此方法将用户指定的索引添加到表中，并会对已经存在的用户数据创建对应的索引数据，也可先调用该方法生成索引再在存入用户数据的同时生成索引数据。在此操作之后，这些索引立即可用于scan/filter操作。	

操作	接口	描述	注意事项
删除索引	dropIndices()	仅删除索引。该API从表中删除用户指定的索引，但跳过相应的索引数据。在此操作之后，索引不能用于scan/filter操作。集群在major compaction期间会自动删除旧的索引数据。 此API使用场景为表中包含大量索引数据且dropIndicesWithData()不可行。另外，用户也可以通过TableIndexer工具删除索引以及索引数据。	• 在索引的状态为ACTIVE，INACTIVE和DROPPING时，允许禁用索引的操作。 • 对于使用dropIndices()删除索引的操作，用户必须确保在将索引添加到具有相同索引名的表之前，相应的索引数据已被删除（即major compaction已完成）。 • 用户删除相应的索引会删除： – 一个带有索引的列族。 – 组合索引所有列族中的任一个列族。 • 索引可以通过HIndex TableIndexer工具与索引数据一起删除。
	dropIndicesWithData()	删除索引数据。此API删除用户指定的索引，并删除用户表中与这些索引对应的所有索引数据。在此操作之后，删除的索引完全从表中删除，不再可用于scan/filter操作。	

操作	接口	描述	注意事项
启用/禁用索引	disableIndices()	该API禁用所有用户指定的索引，使其不再可用于scan/filter操作。	• 在索引的状态为ACTIVE，INACTIVE和BUILDING时允许启用索引的操作。 • 在索引的状态为ACTIVE和INACTIVE时允许禁用索引操作。 • 在禁用索引之前，用户必须确保索引数据与用户数据一致。如果在索引处于禁用状态期间没有在表中添加新的数据，索引数据与用户数据将保持一致。 • 启用索引时，可以通过使用TableIndexer工具构建索引来保证数据一致性。
	enableIndices()	该API启用所有用户指定的索引，使其可用于scan/filter操作。	
查看已创建的索引	listIndices()	该API可用于列出给定表中的所有索引。	无

基于索引查询数据

在具有索引的用户表中，可以使用Filter来查询数据。对于创建单索引和组合索引的用户表，使用过滤器查询的结果与没有使用索引的表相同，但数据查询性能高于没有使用索引的表。

索引的使用规则如下：

- 对于为一个或多个列创建单个索引的情况：
  - 当将此列用于AND或OR查询筛选时，使用索引可以提高查询性能。

    例如，Filter_Condition（IndexCol1）AND / OR Filter_Condition（IndexCol2）。
  - 当在查询中使用“索引列和非索引列”进行过滤时，此索引可以提高查询性能。

    例如，Filter_Condition（IndexCol1）AND Filter_Condition（IndexCol2）AND Filter_Condition（NonIndexCol1）。
  - 当在查询中使用“索引列或非索引列”进行筛选时，但不使用索引，查询性能不会提高。

    例如，Filter_Condition（IndexCol1）AND / OR Filter_Condition（IndexCol2） OR Filter_Condition（NonIndexCol1）。

- 对于为多个列创建组合索引的情况：
  - 当用于查询的列是组合索引的全部或部分列并且与组合索引具有相同的顺序时，使用索引会提高查询性能。

    例如，为C1，C2和C3创建组合索引。

    - 该索引在以下情况下生效：

      Filter_Condition（IndexCol1）AND Filter_Condition（IndexCol2）AND Filter_Condition（IndexCol3）

      Filter_Condition（IndexCol1）AND Filter_Condition（IndexCol2）

      FILTER_CONDITION（IndexCol1）

    - 该索引在下列情况下不生效：

      Filter_Condition（IndexCol2）AND Filter_Condition（IndexCol3）

      Filter_Condition（IndexCol1）AND Filter_Condition（IndexCol3）

      FILTER_CONDITION（IndexCol2）

      FILTER_CONDITION（IndexCol3）
  - 当在查询中使用“索引列和非索引列”进行过滤时，使用索引可提高查询性能。

    例如：

    Filter_Condition（IndexCol1）AND Filter_Condition（NonIndexCol1）

    Filter_Condition（IndexCol1）AND Filter_Condition（IndexCol2）AND Filter_Condition（NonIndexCol1）
  - 当在查询中使用“索引列或非索引列”进行筛选时，但不使用索引，查询性能不会提高。

    例如：

    Filter_Condition（IndexCol1）OR Filter_Condition（NonIndexCol1）

    （Filter_Condition（IndexCol1）AND Filter_Condition（IndexCol2））OR（Filter_Condition（NonIndexCol1））
  - 当多个列用于查询时，只能为组合索引中的最后一列指定值范围，而其他列只能设置为指定值。

    例如，为C1，C2和C3创建组合索引。在范围查询中，只能为C3设置数值范围，过滤条件为“C1 = XXX，C2 = XXX，C3 = 数值范围”。

查询策略选择

使用SingleColumnValueFilter或SingleColumnRangeFilter，它会在一个在过滤条件中提供确定值column_family:qualifierpair（称该列为col1）。

若col1作为表上的第一个索引列，那么该表上的任何索引都可以成为查询期间使用的候选索引。例如：

如果有col1上的索引，可以将此索引作为候选索引，因为col1是此索引的第一列也是唯一的列；如果在col1和col2上有另一个索引，可以将此索引视为候选索引，因为col1是索引列列表中的第一列。另一方面，如果在col2和col1上有一个索引，则不能将此索引作为候选索引，因为索引列列表中的第一列不是col1。

现在最适合使用索引的方法是，当有多个候选索引时，需要从可能的候选索引中选择最适合scan数据的索引。

可借助以下方案来了解如何选择索引策略：

- 需要可以完全匹配。
  场景：有两个索引可用，一个用于col1＆col2，另一个单独用于col1。
  在上面的场景中，第二个索引会比第一个索引更好，因为它会使scan的较少索引数据。
- 如果有多个候选多列索引，则选择具有较少索引列的索引。
  场景：有两个索引可用，一个用于col1＆col2，另一个用于col1＆col2＆col3。
  在这种情况下，可以使用col1和col2上的索引，因为它会使scan的较少索引数据。

说明

- 基于索引查询时索引的状态必须为ACTIVE（可通过调用listIndices() API查看索引的状态）。
- 为了保证基于索引查询数据的正确性，用户应该确保索引数据与用户数据的一致性。
- 使用以下命令可通过HBase shell客户端执行复杂查询（假定此时 已为指定列建立索引）。
  scan 'tablename', {FILTER => "SingleColumnValueFilter(family, qualifier, compareOp, comparator, filterIfMissing, latestVersionOnly)"}
  例如：**scan 'test', {FILTER => "SingleColumnValueFilter('info', 'age', =, 'binary:26', true, true)"}**
  （在以上场景中，用户希望在结果中保存没有查询到的列所在行时，不应该在任何这样的列上创建任何索引，因为如果查询的列不存在于其中时，使用SCVF扫描索引列会过滤出一行。而使用filterIfMissingset为false（这是默认值）的SCVF扫描非索引列时，也将会在结果中返回没有查询列的行。因此，为避免查询结果不一致，建议在为索引列创建SCVF后将filterIfMissing设置为true。）
- 在hbase shell中可以通过以下命令查看为用户数据建立的索引数据。
  scan 'tablename', {ATTRIBUTES => {'FETCH_INDEX_DATA' => 'true'}}

12.8.8.2 批量加载索引数据

场景介绍

HBase本身提供了ImportTsv＆LoadIncremental工具来批量加载用户数据。当前提供了HIndexImportTsv来支持加载用户数据的同时可以完成对索引数据的批量加载。HIndexImportTsv继承了HBase批量加载数据工具ImportTsv的所有功能。此外，若在执行HIndexImportTsv工具之前未建表，直接运行该工具，将会在创建表时创建索引，并在生成用户数据的同时生成索引数据。

操作步骤

1. 将数据导入到HDFS中。
   hdfs dfs -mkdir <*inputdir*>
   hdfs dfs -put <local_data_file> <*inputdir*>
   例如定义数据文件“data.txt”，内容如下：

```
12005000201,Zhang San,Male,19,A City, A Province
12005000202,Li Wanting,Female,23,B City, B Province
12005000203,Wang Ming,Male,26,C City, C Province
12005000204,Li Gang,Male,18,D City, D Province
12005000205,Zhao Enru,Female,21,E City, E Province
12005000206,Chen Long,Male,32,F City, F Province
12005000207,Zhou Wei,Female,29,G City, G Province
12005000208,Yang Yiwen,Female,30,H City, H Province
12005000209,Xu Bing,Male,26,I City, I Province
12005000210,Xiao Kai,Male,25,J City, J Province
```

执行以下命令：

hdfs dfs -mkdir /datadirImport

hdfs dfs -put data.txt /datadirImport

2. 建表bulkTable，进入hbase shell，执行命令建表，例如：

   create 'bulkTable', {NAME => 'info',COMPRESSION => 'SNAPPY', DATA_BLOCK_ENCODING => 'FAST_DIFF'},{NAME=>'address'}

   执行完成后退出hbase shell。

3. 执行如下命令，生成HFile文件（StoreFiles）：

   hbase org.apache.hadoop.hbase.hindex.mapreduce.HIndexImportTsv -Dimporttsv.separator=*<separator>*

   -Dimporttsv.bulk.output=*</path/for/output>* **-Dindexspecs.to.add=**<indexspecs> **-Dimporttsv.columns=**<columns> tableName *<inputdir>*

   - -Dimport.separator：分隔符，例如，-Dimport.separator=','。
   - -Dimport.bulk.output=</path/for/output>：指的是执行结果输出路径，需指定一个不存在的路径。
   - <columns>：指的是导入数据在表中的对应关系，例如，-Dimporttsv.columns=HBASE_ROW_KEY,info:name,info:gender,info:age,address:city,address:province。
   - <tablename>：指的是要操作的表名。
   - <inputdir>：指的是要批量导入的数据目录。
   - -Dindexspecs.to.add=<indexspecs>：指的是索引名与列的映射，例如-Dindexspecs.to.add='index_bulk=>info:[age->String]'。 其构成可以表示如下：

     indexNameN=>familyN :[columnQualifierN-> columnQualifierDataType], [columnQualifierM-> columnQualifierDataType];familyM: [columnQualifierO-> columnQualifierDataType]# indexNameN=> familyM: [columnQualifierO-> columnQualifierDataType]

     其中，列限定符用逗号（，）分隔

     例如：“index1 => f1：[c1-> String]，[c2-> String]”

     列族由分号（;）分隔

     例如：“index1 => f1：[c1-> String]，[c2-> String]; f2：[c3-> Long]”

     多个索引由#号键（ # ）分隔

     例如：“index1 => f1：[c1-> String]，[c2-> String]; f2：[c3-> Long] # index2 => f2：[c3-> Long]”

     列限定的数据类型：

     可用的数据类型有：STRING，INTEGER，FLOAT，LONG，DOUBLE，SHORT，BYTE，CHAR

     说明

     数据类型也可以用小写传递。

   例如执行以下命令：

   hbase org.apache.hadoop.hbase.hindex.mapreduce.HIndexImportTsv -Dimporttsv.separator=',' -Dimporttsv.bulk.output=/dataOutput -Dindexspecs.to.add='index_bulk=>info:[age->String]' -

Dimporttsv.columns=HBASE_ROW_KEY,info:name,info:gender,info:age,address:city,address:province bulkTable /datadirImport/data.txt

输出：

```
[root@shap000000406 opt]# hbase org.apache.hadoop.hbase.hindex.mapreduce.HIndexImportTsv -Dimporttsv.separator=',' -Dimporttsv.bulk.output=/dataOutput -Dindexspecs.to.add='index_bulk=>info:[age->String]' -Dimporttsv.columns=HBASE_ROW_KEY,info:name,info:gender,info:age,address:city,address:province bulkTable /datadirImport/data.txt
2018-05-08 21:29:16,059 INFO  [main] mapreduce.HFileOutputFormat2: Incremental table bulkTable output configured.
2018-05-08 21:29:16,069 INFO  [main] client.ConnectionManager$HConnectionImplementation: Closing master protocol: MasterService
2018-05-08 21:29:16,069 INFO  [main] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x80007c2cb4fd5b4d
2018-05-08 21:29:16,072 INFO  [main] zookeeper.ZooKeeper: Session: 0x80007c2cb4fd5b4d closed
2018-05-08 21:29:16,072 INFO  [main-EventThread] zookeeper.ClientCnxn: EventThread shut down for session: 0x80007c2cb4fd5b4d
2018-05-08 21:29:16,379 INFO  [main] client.ConfiguredRMFailoverProxyProvider: Failing over to 147
2018-05-08 21:29:17,328 INFO  [main] input.FileInputFormat: Total input files to process : 1
2018-05-08 21:29:17,413 INFO  [main] mapreduce.JobSubmitter: number of splits:1
2018-05-08 21:29:17,430 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-05-08 21:29:17,687 INFO  [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1525338489458_0002
2018-05-08 21:29:18,100 INFO  [main] impl.YarnClientImpl: Submitted application application_1525338489458_0002
2018-05-08 21:29:18,136 INFO  [main] mapreduce.Job: The url to track the job: http://shap000000407:8088/proxy/application_1525338489458_0002/
2018-05-08 21:29:18,136 INFO  [main] mapreduce.Job: Running job: job_1525338489458_0002
2018-05-08 21:29:28,248 INFO  [main] mapreduce.Job: Job job_1525338489458_0002 running in uber mode : false
2018-05-08 21:29:28,249 INFO  [main] mapreduce.Job:  map 0% reduce 0%
2018-05-08 21:29:38,344 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2018-05-08 21:29:51,421 INFO  [main] mapreduce.Job:  map 100% reduce 100%
2018-05-08 21:29:51,428 INFO  [main] mapreduce.Job: Job job_1525338489458_0002 completed successfully
2018-05-08 21:29:51,523 INFO  [main] mapreduce.Job: Counters: 50
```

4. 执行如下命令将生成的HFile导入HBase中：

   hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles *</path/for/output> <tablename>*

   例如执行以下命令：

   hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /dataOutput bulkTable

   输出：

```
[root@shap000000406 opt]# hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /dataOutput bulkTable
2018-05-08 21:30:01,398 WARN  [main] mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://hacluster/dataOutput/_SUCCESS
2018-05-08 21:30:02,006 INFO  [LoadIncrementalHFiles-0] hfile.CacheConfig: Created cacheConfig: CacheConfig:disabled
2018-05-08 21:30:02,006 INFO  [LoadIncrementalHFiles-2] hfile.CacheConfig: Created cacheConfig: CacheConfig:disabled
2018-05-08 21:30:02,006 INFO  [LoadIncrementalHFiles-1] hfile.CacheConfig: Created cacheConfig: CacheConfig:disabled
2018-05-08 21:30:02,085 INFO  [LoadIncrementalHFiles-2] compress.CodecPool: Got brand-new decompressor [.snappy]
2018-05-08 21:30:02,120 INFO  [LoadIncrementalHFiles-0] mapreduce.LoadIncrementalHFiles: Trying to load hfile=hdfs://hacluster/dataOutput/address/042426c252f74e859858c7877b95e510 first=12005000201 last=12005000210
2018-05-08 21:30:02,120 INFO  [LoadIncrementalHFiles-2] mapreduce.LoadIncrementalHFiles: Trying to load hfile=hdfs://hacluster/dataOutput/info/f3995920ae0247a88182f637aa031c49 first=12005000201 last=12005000210
2018-05-08 21:30:02,128 INFO  [LoadIncrementalHFiles-1] mapreduce.LoadIncrementalHFiles: Trying to load hfile=hdfs://hacluster/dataOutput/d/c53b252248af42779f29442ab84f86b8 first=\x00index_bulk
```

```
\x00\x00\x00\x00\x00\x00\x00\x0018\x00\x0012005000204 last=\x00index_bulk
\x00\x00\x00\x00\x00\x00\x00\x0032\x00\x0012005000206
2018-05-08 21:30:02,231 INFO  [main] client.ConnectionManager$HConnectionImplementation:
Closing master protocol: MasterService
2018-05-08 21:30:02,231 INFO  [main] client.ConnectionManager$HConnectionImplementation:
Closing zookeeper sessionid=0x81007c2cf0f55cc5
2018-05-08 21:30:02,235 INFO  [main] zookeeper.ZooKeeper: Session: 0x81007c2cf0f55cc5 closed
2018-05-08 21:30:02,235 INFO  [main-EventThread] zookeeper.ClientCnxn: EventThread shut down
for session: 0x81007c2cf0f55cc5
```

12.8.8.3 使用索引生成工具

场景介绍

为了快速对用户数据创建索引，HBase提供了可通过MapReduce功能创建索引的TableIndexer工具，该工具可实现添加，构建和删除索引。具体使用场景如下：

- 在用户的表中预先存在大量数据的情况下，可能希望在某个列上添加索引。但是，使用addIndicesWithData（ ）API添加索引会生成与相关用户数据对应的索引数据，这将花费大量时间。另一方面，使用addIndices（ ）创建的索引不会构建与用户数据对应的索引数据。因此，为了为这样的用户数据建立索引数据，用户可以使用TableIndexer工具来完成索引的构建。
- 如果索引数据与用户数据不一致，该工具可用于重新构建索引数据。

  如果用户暂时禁用索引并且在此期间，向禁用的索引列执行新的put操作，然后直接将索引从禁用状态启用可能会导致索引数据与用户数据不一致。因此，用户必须注意在再次使用之前重新构建所有索引数据。
- 对于大量现有的索引数据，用户可以使用TableIndexer工具将索引数据从用户表中完全删除。
- 对于未建立索引的用户表，该工具允许用户同时添加和构建索引。

使用方法

- **添加新的索引到用户表**

  命令如下所示：

  hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer -Dtablename.to.index=tablename -Dindexspecs.to.add='idx_0=>cf_0:[q_0->string],[q_1];cf_1:[q_2],[q_3]#*idx_1*=>cf_1:[q_4]'

  它需要以下参数:

  - **tablename.to.index**：表示创建索引的表的名称
  - **indexspecs.to.add**：表示与索引名与对应用户表的列的映射
  - **scan.caching**（可选）：包含一个整数值，表示在扫描数据表时将传递给扫描器的缓存行数

  上述命令中的参数描述如下：

  - **idx_1**：表示索引名称
  - **cf_0**：表示列族名称
  - **q_0**：表示列名称
  - **string**：表示数据类型。它可以是STRING，INTEGER，FLOAT，LONG，DOUBLE，SHORT，BYTE或CHAR

说明

- '#'用于分隔索引，';' 用于分隔列族，','用于分隔列限定符。
- 列名及其数据类型应包含在'[]'中。
- 列名及其数据类型通过' - >'分隔。
- 如果未指定具体列的数据类型，则使用默认数据类型（ string ）。
- 如果未设置可选参数scan.caching，则将采用默认值1000。
- 用户表必须存在。
- 表中指定的索引不能存在.。
- 如果用户表中已经存在名称为'd'的ColumnFamily，则用户必须使用TableIndexer工具构建索引数据。

在执行以上的命令之后，指定的索引将被添加到表中并且将处于INACTIVE状态。该行为与addIndices() API类似。

- **为用户表中的现有索引构建索引数据**

  该命令如下：

  hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer -Dtablename.to.index=tablename -Dindexnames.to.build ='idx_0#idx_1'

  它采用以下参数：

  - **tablename.to.index**：表示创建索引的表的名称
  - **indexspecs.to.build**：表示与索引名称
  - **scan.caching**（可选）：包含一个整数值，表示在扫描数据表时将传递给扫描器的缓存行数

  上述命令中的参数描述如下：

  - **idx_1**：表示索引名称

  说明

  - '#'用于分隔索引名称。
  - 如果未设置可选参数scan.caching，则将采用默认值1000。
  - 用户表必须存在。

  在执行以上的命令之后，指定的索引将被设置为ACTIVE状态。用户扫描数据时可以使用它们。

- **从用户表中删除现有索引及其数据**

  该命令如下：

  hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer -Dtablename.to.index=tablename -Dindexnames.to.drop='idx_0#idx_1'

  它需要以下参数：

  - **tablename.to.index**：表示创建索引的表的名称
  - **indexnames.to.drop**：表示应该和其数据一起删除的索引的名称（必须存在于表中）
  - **scan.caching**（可选）：其中包含一个整数值，指示在扫描数据表时将传递给扫描器的缓存行数

  上述命令中的参数描述如下：

  - **idx_1**：表示索引名称

说明

- '#'用于分隔索引名称。
- 如果未设置可选参数scan.caching，则将采用默认值1000。
- 用户表必须存在。

在执行前面的命令之后，指定的索引将从表中删除。

- **为用户表添加新的索引以及基于现有数据的数据构建**

  该命令如下：

  hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer -Dtablename.to.index=tablename -Dindexspecs.to.add='idx_0 => cf_0:[q_0-> string],[q_1];cf_1:[q_2]，[q_3] # idx_1 => cf_1:[q_4]' -Dindexnames.to.build='idx_0'

  说明

  - 参数与之前的情况相同。
  - 用户表必须存在。
  - indexspecs.to.add中指定的索引不得存在于表中。
  - indexnames.to.build中指定的索引名称必须已经存在于表中，或者应该是indexspecs.to.add的一部分。

  在执行前面的命令之后，indexspecs.to.add中指定的所有索引都将添加到该表中，并且将为通过indexnames.to.build为指定的所有索引构建索引数据。

12.8.8.4 索引数据迁移

操作场景

MRS 1.7及其以后版本中使用的索引与以前MRS版本中HBase使用的二级索引都不兼容。因此，为了将索引数据从以前的版本（MRS 1.5及其以前版本）迁移到MRS 1.7及其以后版本，需要遵循以下步骤。

前提条件

1. 迁移数据时旧版本集群应为MRS1.5及其以前的版本，新版本集群应为MRS1.7及其以后的版本。
2. 迁移数据前用户应该有旧的索引数据。
3. 安全集群需配置跨集群互信和启用集群间拷贝功能，普通集群仅需启用集群间拷贝功能。

操作步骤

把旧集群中的用户数据迁移至新集群中。迁移数据需单表手动同步新旧集群的数据，通过Export、distcp、Import来完成。

例如，当前旧集群有用户表（t1，索引名为idx_t1）及其对应的索引表（t1_idx）。迁移数据的操作步骤如下：

1. 从旧集群导出表中数据。

   ```
   hbase org.apache.hadoop.hbase.mapreduce.Export -Dhbase.mapreduce.include.deleted.rows=true <tableName> <path/for/data>
   ```

   - *<tableName>*：指的是表名。例如，t1。

- *<path/for/data>*：指的是保存源数据的路径，例如“/user/hbase/t1”。

例如，**hbase org.apache.hadoop.hbase.mapreduce.Export -Dhbase.mapreduce.include.deleted.rows=true t1 /user/hbase/t1**

2. 把导出的数据按如下步骤复制到新集群中。

```
hadoop distcp <path/for/data> hdfs://ActiveNameNodeIP:9820/<path/for/newData>
```

- *<path/for/data>*：指的是旧集群保存源数据的路径。例如，/user/hbase/t1。
- <path/for/newData>：指的是新集群保存源数据的路径。例如，/user/hbase/t1。

其中，ActiveNameNodeIP是新集群中主NameNode节点的IP地址。

例如，**hadoop distcp /user/hbase/t1 hdfs://192.168.40.2:9820/user/hbase/t1**

说明

- 可手动把导出的数据复制到新集群HDFS中，如上路径：“/user/hbase/t1”。

3. 使用新集群HBase表用户，在新集群中生成HFiles。

```
hbase org.apache.hadoop.hbase.mapreduce.Import -Dimport.bulk.output=<path/for/hfiles> <tableName><path/for/newData>
```

- *<path/for/hfiles>*：指的是新集群生成HFiles的路径。例如，/user/hbase/output_t1。
- <tableName>：指的是表名。例如，t1。
- <path/for/newData>：指的是新集群保存源数据的路径。例如，/user/hbase/t1。

例如，

hbase org.apache.hadoop.hbase.mapreduce.Import -Dimport.bulk.output=/user/hbase/output_t1 t1 /user/hbase/t1

4. 把生成的HFiles导入新集群相应表中。

命令如下：

```
hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles <path/for/hfiles> <tableName>
```

- *<path/for/hfiles>*：指的是新集群生成HFiles的路径。例如，/user/hbase/output_t1。
- <tableName>：指的是表名。例如，t1。

例如，

hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /user/hbase/output_t1 t1

说明

1. 以上为迁移用户数据的过程，旧集群的索引数据迁移只需按照前三步操作，并更改相应表名为索引表名（如，t1_idx）。
2. 迁移索引数据时无需执行4。

5. 向新集群表中导入索引数据。

a. 在新集群的用户表中添加与之前版本用户表相同的索引（名称为'd'的列族不应该已经存在于用户表中）。

命令如下所示：

```
hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer -Dtablename.to.index=<tableName> -Dindexspecs.to.add=<indexspecs>
```

- -Dtablename.to.index=<tableName>：指的是表名。例如，-Dtablename.to.index=t1。
- -Dindexspecs.to.add=<indexspecs>：指的是索引名与列的映射，例如-Dindexspecs.to.add='idx_t1=>info:[name->String]'。

例如，

hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer -Dtablename.to.index=t1 -Dindexspecs.to.add='idx_t1=>info:[name->String]'

说明

如果用户表中已经存在名称为'd'的ColumnFamily，则用户必须使用TableIndexer工具构建索引数据。

b. 运行LoadIncrementalHFiles工具加载索引数据，将旧集群索引数据加载到新集群表中。

命令如下：

```
hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles </path/for/hfiles> <tableName>
```

- </path/for/hfiles>：指的是索引数据在HDFS上的路径（其为-Dimport.bulk.output中指定的索引生成路径）。例如，/user/hbase/output_t1_idx。
- <tableName>：指的是新集群中表名，例如，t1。

例如，

hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /user/hbase/output_t1_idx t1

12.8.9 配置 HBase 容灾

操作场景

HBase集群容灾作为提高HBase集群系统高可用性的一个关键特性，为HBase提供了实时的异地数据容灾功能。它对外提供了基础的运维工具，包含灾备关系维护，重建，数据校验，数据同步进展查看等功能。为了实现数据的实时容灾，可以把本HBase集群中的数据备份到另一个集群。支持HBase表普通写数据与Bulkload批量写数据场景下的容灾。

说明

本章节适用于MRS 3.x及之后版本。

前提条件

- 主备集群都已经安装并启动成功，且获取集群的管理员权限。
- 必须保证主备集群间的网络畅通和端口的使用。
- 如果主集群部署为安全模式且不由一个FusionInsight Manager管理，主备集群必须已配置跨集群互信。如果主集群部署为普通模式，不需要配置跨集群互信。
- 主备集群必须已配置跨集群拷贝。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。

- 必须在主备集群的所有节点以及主集群客户端所在节点的hosts文件中，配置主备集群所有机器的机器名与业务IP地址的对应关系。
- 主备集群间的网络带宽需要根据业务流量而定，不应少于最大的可能业务流量。
- 主备集群安装的MRS版本需要保持一致。
- 备集群规模不小于主集群规模。

使用约束

- 尽管容灾提供了实时的数据复制功能，但实际的数据同步进展，由多方面的因素决定的，例如，当前主集群业务的繁忙程度，备集群进程的健康状态等。因此，在正常情形下，备集群不应该接管业务。极端情形下是否可以接管业务，可由系统维护人员以及决策人员根据当前的数据同步指标来决定。
- 容灾功能当前仅支持一主一备。
- 通常情况下，不允许对备集群的灾备表进行表级别的操作，例如修改表属性、删除表等，一旦误操作备集群后会造成主集群数据同步失败、备集群对应表的数据丢失。
- 主集群的HBase表已启用容灾功能同步数据，用户每次修改表的结构时，需要手动修改备集群的灾备表结构，保持与主集群表结构一致。

操作步骤

配置主集群普通写数据容灾参数。

步骤1 登录主集群的Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”，进入HBase配置界面。

步骤3 （可选）如表12-168所示，为HBase容灾操作过程中的可选配置项，您可以根据描述来进行参数配置，或者使用缺省提供的值。

表 12-168 可选配置项

配置入口	配置项	缺省值	描述
“HMaster > 性能”	hbase.master.logcleaner.ttl	600000	指定HLog的保存期限。如果配置值为“604800000”（单位：毫秒），表示HLog的保存期限为7天。
	hbase.master.cleaner.interval	60000	HMaster清理过去HLog文件的周期，即超过设置的时间的HLog会被自动删除。建议尽可能配置大的值来保留更多的HLog。
“RegionServer > Replication”	replication.source.size.capacity	16777216	edits最大大小。单位为byte。如果edit大小超过这个值Hlog edits将会发送到备集群。

配置入口	配置项	缺省值	描述
	replication.source.nb.capacity	25000	edits最大数目，这是另一个触发Hlog edits到备集群的条件。当主集群同步数据到备集群中时，主集群会从HLog中读取数据，此时会根据本参数配置的个数读取并发送。与“replication.source.size.capacity”一起配置使用。
	replication.source.maxretriesmultiplier	10	replication出现异常时的最大重试次数。
	replication.source.sleepforretries	1000	每次重试的sleep时间。（单位：毫秒）
	hbase.regionserver.replication.handler.count	6	RegionServer上的replication RPC服务器实例数。

配置主集群Bulkload批量写数据容灾参数。

步骤4 是否启用Bulkload批量写数据容灾功能?

是，执行**步骤5**。

否，执行**步骤8**。

步骤5 选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”，进入HBase配置界面。

步骤6 搜索并修改“hbase.replication.bulkload.enabled”参数，将配置项的值修改为“true”，启用Bulkload批量写数据容灾功能。

步骤7 搜索并修改“hbase.replication.cluster.id”参数，表示标识主集群HBase的ID，用于备集群连接主集群。参数值支持大小写字母、数字和下划线（_），长度不超过30。

重启HBase服务并安装客户端。

步骤8 单击“保存”，保存配置。在弹出的窗口中单击“确定”。重启HBase服务。

步骤9 在主备集群，选择“集群 > *待操作集群的名称* > 服务 > HBase > 更多 > 下载客户端”，下载客户端并安装。

添加主备集群容灾关系。

步骤10 以“hbase”用户进入主集群的HBase shell界面。

步骤11 在HBase shell中执行如下命令，创建主集群HBase与备集群HBase之间的容灾同步关系。

add_peer '*备集群ID***', CLUSTER_KEY => "***备集群ZooKeeper业务ip地址***", CONFIG => {"hbase.regionserver.kerberos.principal" => "***备集群RegionServer principal***", "hbase.master.kerberos.principal" => "***备集群HMaster principal***"}**

- 备集群ID表示主集群识别备集群使用的id，请重新指定id值。可以任意指定，建议使用数字。

- 备集群ZooKeeper地址信息包含ZooKeeper业务IP地址、侦听客户端连接的端口和备集群的HBase在ZooKeeper上的根目录。
- hbase.master.kerberos.principal 、hbase.regionserver.kerberos.principal在备集群HBase hbase-site.xml配置文件中查找。

例如，添加主备集群容灾关系，执行：**add_peer '*备集群ID*', CLUSTER_KEY => "*192.168.40.2,192.168.40.3,192.168.40.4:24002:/hbase*", CONFIG => {"hbase.regionserver.kerberos.principal" => "*hbase/hadoop.hadoop.com@HADOOP.COM*", "hbase.master.kerberos.principal" => "*hbase/hadoop.hadoop.com@HADOOP.COM*"}**

步骤12 （可选）如果启用Bulkload批量写数据容灾功能，主集群HBase客户端配置必须拷贝到备集群。

- 在备集群HDFS创建目录/hbase/replicationConf/*主集群hbase.replication.cluster.id*
- 主机群HBase客户端配置文件，拷贝到备集群HDFS目录/hbase/replicationConf/*主集群hbase.replication.cluster.id*

  例如：**hdfs dfs -put HBase/hbase/conf/core-site.xml HBase/hbase/conf/hdfs-site.xml HBase/hbase/conf/yarn-site.xml hdfs://*NameNode IP*:*25000*/hbase/replicationConf/*source_cluster***

启用HBase容灾功能同步数据。

步骤13 检查备集群的HBase服务实例中，是否已存在一个命名空间，与待启用容灾功能的HBase表所属的命名空间名称相同?

- 是，存在同名的命令空间，执行**步骤14**。
- 否，不存在同名的命令空间，需先在备集群的HBase shell中，创建同名的命名空间，然后执行**步骤14**。

步骤14 在主集群的HBase shell中，以“hbase”用户执行以下命令，启用将主集群表的数据实时容灾功能，确保后续主集群中修改的数据能够实时同步到备集群中。

一次只能针对一个HTable进行数据同步。

enable_table_replication '*表名*'

说明

- 若备集群中不存在与要开启实时同步的表同名的表，则该表会自动创建。
- 若备集群中存在与要开启实时同步的表同名的表，则两个表的结构必须一致。
- 若'表名'设置了加密算法SMS4或AES，则不支持对此HBase表启用将数据从主集群实时同步到备集群的功能。
- 若备集群不在线，或备集群中已存在同名但结构不同的表，启用容灾功能将失败。
- 若主集群中部分Phoenix表启用容灾功能同步数据，则备集群中不能存在与主集群Phoenix表同名的普通HBase表，否则启用容灾功能失败或影响备集群的同名表正常使用。
- 若主集群中Phoenix表启用容灾功能同步数据，还需要对Phoenix表的元数据表启用容灾功能同步数据。需配置的元数据表包含SYSTEM.CATALOG、SYSTEM.FUNCTION、SYSTEM.SEQUENCE和SYSTEM.STATS。
- 若主集群的HBase表启用容灾功能同步数据，用户每次为HBase表增加新的索引，需要手动在备集群的灾备表增加二级索引，保持与主集群二级索引结构一致。
- HBase多实例也支持容灾功能，需要修改主集群对应的HBase服务实例参数，且在备集群对应多实例的客户端执行命令。添加容灾关系时需要选择备集群ZooKeeper保存HBase多实例数据的目录，例如“hbase1”。

步骤15 （可选）如果HBase没有使用Ranger，在主集群的HBase shell中，以“hbase”用户执行以下命令，启用主集群的HBase表权限控制信息数据实时容灾功能。

enable_table_replication 'hbase:acl'

创建用户

步骤16 登录备集群的FusionInsight Manager，选择“系统 > 权限 > 角色 > 添加角色”创建一个角色，并根据主集群HBase源数据表的权限，为角色添加备数据表的相同权限。

步骤17 选择“系统 > 权限 > 用户 > 添加用户”创建一个用户，根据业务需要选择用户类型为“人机”或“机机”，并将用户加入创建的角色。使用新创建的用户，访问备集群的HBase容灾数据。

说明

- 主集群HBase源数据表修改权限时，如果备集群需要正常读取数据，请修改备集群角色的权限。
- 如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加HBase的Ranger访问权限策略**。

同步主集群表数据。

步骤18 检查配置HBase容灾并启用数据同步后，主集群是否已存在表及数据，且历史数据需要同步到备集群?

- 是，存在表且需要同步数据，以HBase表用户登录安装主集群HBase客户端的节点，并执行kinit用户名认证身份。该用户需要拥有表的读写权限，以及“hbase:meta”表的执行权限。然后执行**步骤19**。
- 否，不需要同步数据，任务结束。

步骤19 配置HBase容灾时不支持自动同步表中的历史数据，需要对主集群的历史数据进行备份，然后再手动恢复历史数据到备集群中。

手动恢复即单表的恢复，单表手动恢复通过Export、distcp、Import来完成。

单表手动恢复操作步骤：

1. 从主集群导出表中数据。

   hbase org.apache.hadoop.hbase.mapreduce.Export -Dhbase.mapreduce.include.deleted.rows=true *表名 保存源数据的目录*

   例如，**hbase org.apache.hadoop.hbase.mapreduce.Export -Dhbase.mapreduce.include.deleted.rows=true t1 /user/hbase/t1**

2. 把导出的数据复制到备集群。

   hadoop distcp *主集群保存源数据的目录 hdfs://ActiveNameNodeIP:8020/备集群保存源数据的目录*

   其中，ActiveNameNodeIP是备集群中主NameNode节点的IP地址。

   例如，**hadoop distcp /user/hbase/t1 hdfs://192.168.40.2:8020/user/hbase/t1**

3. 使用备集群HBase表用户，在备集群中导入数据。

   在备集群HBase shell界面，使用“hbase”用户执行以下命令保持写数据状态：

   set_clusterState_active

   界面提示以下信息表示执行成功：

   ```
   hbase(main):001:0> set_clusterState_active
   => true
   ```

hbase org.apache.hadoop.hbase.mapreduce.Import *-Dimport.bulk.output=备集群保存输出的目录 表名 备集群保存源数据的目录*

hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles *备集群保存输出的目录 表名*

例如：

```
hbase(main):001:0> set_clusterState_active
=> true
```

hbase org.apache.hadoop.hbase.mapreduce.Import -Dimport.bulk.output=/user/hbase/output_t1 t1 /user/hbase/t1

hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /user/hbase/output_t1 t1

步骤20 在HBase客户端执行以下命令，校验主备集群同步的数据。启用容灾功能同步功能后，也可以执行该命令检验新的同步数据是否一致。

hbase org.apache.hadoop.hbase.mapreduce.replication.VerifyReplication --starttime=*开始时间* **--endtime**=*结束时间 列族名称 备集群ID 表名*

说明

- 开始时间必须早于结束时间
- 开始时间和结束时间需要填写时间戳的格式，例如执行date -d "2015-09-30 00:00:00" +%s将普通时间转化为时间戳格式。

指定主备集群写数据状态。

步骤21 在主集群HBase shell界面，使用“hbase”用户执行以下命令保持写数据状态。

set_clusterState_active

界面提示以下信息表示执行成功：

```
hbase(main):001:0> set_clusterState_active
=> true
```

步骤22 在备集群HBase shell界面，使用“hbase”用户执行以下命令保持只读数据状态。

set_clusterState_standby

界面提示以下信息表示执行成功：

```
hbase(main):001:0> set_clusterState_standby
=> true
```

----结束

相关命令

表 12-169 HBase 容灾

操作	命令	描述
建立灾备关系	add_peer '*备集群ID*', CLUSTER_KEY => "*备集群ZooKeeper业务ip地址*", CONFIG => {"hbase.regionserver.kerberos.principal" => "*备集群RegionServer principal*", "hbase.master.kerberos.principal" => "*备集群HMaster principal*"} **add_peer '1','zk1,zk2,zk3:2181:/hbase1'** **2181**表示集群中ZooKeeper的端口号。	建立主集群与备集群的关系，让其互相对应。 如果启用Bulkload批量写数据容灾： • 在备集群HDFS创建目录/hbase/replicationConf/*主集群hbase.replication.cluster.id* • 主集群HBase客户端配置文件，拷贝到备集群HDFS目录/hbase/replicationConf/*主集群hbase.replication.cluster.id*
移除灾备关系	**remove_peer** *'备集群ID'* 示例： **remove_peer '1'**	在主集群中移除备集群的信息。
查询灾备关系	**list_peers**	在主集群中查询已经设置的备集群的信息，主要为Zookeeper信息。
启用用户表实时同步	**enable_table_replication** *'表名'* 示例： **enable_table_replication 't1'**	在主集群中，设置已存在的表同步到备集群。
禁用用户表实时同步	**disable_table_replication** *'表名'* 示例： **disable_table_replication 't1'**	在主集群中，设置已存在的表不同步到备集群。
主备集群数据校验	**bin/hbase org.apache.hadoop.hbase.mapreduce.replication.VerifyReplication --starttime**=*开始时间* **--endtime=***结束时间 列族名称 备集群ID 表名*	检查指定的表在主备集群间的数据是否一致。 命令行中参数说明如下： • 开始时间：如果未设置，则取默认的开始时间为0。 • 结束时间：如果未设置，则取默认的结束时间为当前操作提交的时间。 • 表名：如果未输入表名，则默认校验所有的启用了实时同步的用户表。
切换数据写入状态	**set_clusterState_active** **set_clusterState_standby**	设置集群HBase表是否可写入数据。

操作	命令	描述
新增或更新已经在对端集群保存的主集群中HDFS配置	**hdfs dfs -put -f HBase/hbase/conf/core-site.xml HBase/hbase/conf/hdfs-site.xml HBase/hbase/conf/yarn-site.xml hdfs://**备集群***NameNode IP:PORT***/hbase/replicationConf/***主集群hbase.replication.cluster.id***	启用包含Bulkload数据的容灾，在主集群修改HDFS参数时，新的参数值默认不会从主集群自动同步到备集群，需要手动执行命令同步。受影响的参数如下： ● “fs.defaultFS” ● “dfs.client.failover.proxy.provider.hacluster” ● “dfs.client.failover.connection.retries.on.timeouts” ● “dfs.client.failover.connection.retries” 例如，“fs.defaultFS”修改为“hdfs://hacluster_sale”， 主集群HBase客户端配置文件，重新拷贝到备集群HDFS目录/hbase/replicationConf/*主集群hbase.replication.cluster.id*

12.8.10 配置 HBase 数据压缩和编码

操作场景

HBase可以通过对HFile中的data block编码，减少keyvalue中key的重复部分，从而减少空间的使用。目前对data block的编码方式有：NONE、PREFIX、DIFF、FAST_DIFF和ROW_INDEX_V1，其中NONE表示不使用编码。另外，HBase还支持使用压缩算法对HFile文件进行压缩，默认支持的压缩算法有：NONE、GZ、SNAPPY和ZSTD，其中NONE表示HFile不压缩。

这两种方式都是作用在HBase的列簇上，可以同时使用，也可以单独使用。

前提条件

- 已安装HBase客户端。例如，客户端安装目录为“opt/client”。
- 如果HBase已经开启了鉴权，操作的用户还需要具备对应的操作权限。即创建表时需要具备对应的namespace或更高级别的创建(C)或者管理(A)权限，修改表时需要具备已创建的表或者更高级别的创建(C)或者管理(A)权限。具体的授权操作请参考**创建HBase角色**章节。

操作步骤

创建时设置data block encoding和压缩算法。

- **方法一：使用hbase shell。**
  a. 以客户端安装用户，登录安装客户端的节点。

b. 执行以下命令切换到客户端目录。

cd /opt/client

c. 执行以下命令配置环境变量。

source bigdata_env

d. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

e. 直接执行HBase组件的客户端命令。

hbase shell

f. 创建表。

create '*t1*', **{NAME =>** '*f1*', **COMPRESSION =>** '*SNAPPY*', **DATA_BLOCK_ENCODING =>** '*FAST_DIFF*'**}**

说明

- t1：表名。
- f1：列簇名。
- SNAPPY：该列簇使用的压缩算法为SNAPPY。
- FAST_DIFF：使用的编码方式为FAST_DIFF。
- {}内的参数为指定列簇的参数，多个列簇可以用多个{}，然后用逗号隔开。关于建表语句的更多使用说明可以在**hbase shell**中执行**help 'create'** 进行查看。

- **方法二：使用Java API**。

以下代码片段仅展示如何在建表时设置列簇的编码和压缩方式，完整的建表代码以及如何通过代码建表请参考中“HBase开发指南 > 修改表”章节。

```
TableDescriptorBuilder htd = TableDescriptorBuilder.newBuilder(TableName.valueOf("t1"));// 创建t1表的descriptor.
ColumnFamilyDescriptorBuilder hcd = ColumnFamilyDescriptorBuilder.newBuilder(Bytes.toBytes("f1"));//  创建列簇f1的builder.
hcd.setDataBlockEncoding(DataBlockEncoding.FAST_DIFF);// 设置列簇f1的编码方式为FAST_DIFF.
hcd.setCompressionType(Compression.Algorithm.SNAPPY);// 设置列簇f1的压缩算法为SNAPPY
htd.setColumnFamily(hcd.build())// 将列簇f1添加到t1表的descriptor.
```

对已存在的表设置或修改data block encoding和压缩算法

- **方法一：使用hbase shell**。

a. 以客户端安装用户，登录安装客户端的节点。

b. 执行以下命令切换到客户端目录。

cd /opt/client

c. 执行以下命令配置环境变量。

source bigdata_env

d. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

e. 直接执行HBase组件的客户端命令。

hbase shell

f. 执行修改表的命令。

alter '*t1*', {NAME => '*f1*', COMPRESSION => '*SNAPPY*', DATA_BLOCK_ENCODING => '*FAST_DIFF*'}

- **方法二：使用Java API**。

  以下代码片段仅展示如何修改指定表的已有列簇的编码和压缩方式，完整的修改表代码以及如何通过代码修改表请参考HBase应用开发指南：

```
TableDescriptor htd = admin.getDescriptor(TableName.valueOf("t1"));// 获取表t1的descriptor
ColumnFamilyDescriptor originCF = htd.getColumnFamily(Bytes.toBytes("f1"));// 获取列簇f1的descriptor
builder.ColumnFamilyDescriptorBuilder hcd = ColumnFamilyDescriptorBuilder.newBuilder(originCF);// 通过已有的列簇属性构造一个builder
hcd.setDataBlockEncoding(DataBlockEncoding.FAST_DIFF);// 重新设置列簇的编码方式为FAST_DIFF
hcd.setCompressionType(Compression.Algorithm.SNAPPY);// 重新设置列簇的压缩算法为SNAPPY
admin.modifyColumnFamily(TableName.valueOf("t1"), hcd.build());// 提交到服务端修改列簇f1的属性
```

  修改后完成后，已有的HFile的编码和压缩方式需要在下次做完compaction后才会生效。

12.8.11 HBase 容灾业务切换

操作场景

系统管理员可配置HBase集群容灾功能，以提高系统可用性。容灾环境中的主集群完全故障影响HBase上层应用连接时，需要为HBase上层应用配置备集群信息，才可以使得该应用在备集群上运行。

说明

本章节适用于MRS 3.x及之后版本。

对系统的影响

切换业务后，写入备集群的数据默认不会同步到主集群。主集群故障修复后，备集群新增的数据需要通过备份恢复的方式同步到主集群。如果需要自动同步数据，需要切换HBase容灾主备集群。

操作步骤

步骤1 登录备集群FusionInsight Manager。

步骤2 下载并安装HBase客户端。

步骤3 在备集群HBase客户端，以**hbase**用户执行以下命令指定备集群写数据状态启用。

kinit hbase

hbase shell

set_clusterState_active

界面提示以下信息表示执行成功：

```
hbase(main):001:0> set_clusterState_active
=> true
```

步骤4 确认HBase上层应用中原有的配置文件“hbase-site.xml”、“core-site.xml”和“hdfs-site.xml”是否为适配应用运行修改或新增过配置内容。

- 是，将相关内容同步更新到新的配置文件中，并替换旧的配置文件。

- 否，使用新的配置文件替换HBase上层应用中原有的配置文件。

步骤5 配置HBase上层应用所在主机与备集群的网络连接。

说明

当客户端所在主机不是集群中的节点时，配置客户端网络连接，可避免执行客户端命令时出现错误。

1. 确保客户端所在主机能与客户端安装包文件解压目录下的“hosts”文件中所列出的集群各主机在网络上互通。
2. 当客户端所在主机不是集群中的节点时，需要在客户端所在节点的“/etc/hosts”文件中设置主机名和IP地址（业务平面）映射。主机名和IP地址请保持一一对应。

步骤6 配置HBase上层应用所在主机的时间与备集群的时间保持一致，时间差要小于5分钟。

步骤7 检查主集群的认证模式。

- 若为安全模式，执行**步骤8**。
- 若为普通模式，任务结束。

步骤8 获取HBase上层应用用户的keytab文件和krb5.conf配置文件。

1. 在备集群FusionInsight Manager界面，选择“系统 > 权限 > 用户”。
2. 在用户所在行的“操作”列单击“更多 > 下载认证凭据”，下载keytab文件到本地。
3. 解压得到“user.keytab”和“krb5.conf”。

步骤9 使用“user.keytab”和“krb5.conf”两个文件替换HBase上层应用中原有的文件。

步骤10 停止上层业务。

步骤11 是否需要切换HBase主备集群，即主变成备，备变成主。如果不切换，数据将不再同步。

- 是，先执行HBase容灾主备集群倒换，具体请参考**HBase容灾主备集群倒换**，然后再执行**步骤12**。
- 否，直接执行**步骤12**。

步骤12 启动上层业务。

----结束

12.8.12 HBase 容灾主备集群倒换

操作场景

当前环境HBase已经是容灾集群，因为某些原因，需要将主备集群互换，即备集群变成主集群，主集群变成备集群。

说明

本章节适用于MRS 3.x及之后版本。

对系统的影响

主备集群互换后，原先主集群将不能再写入数据，原先备集群将变成主集群，接管上层业务。

操作步骤

确保上层业务已经停止

步骤1 确保上层业务已经停止，如果没有停止，先执行 参考**HBase容灾业务切换**。

关闭主集群写功能

步骤2 下载并安装HBase客户端。

步骤3 在备集群HBase客户端，以**hbase**用户执行以下命令指定备集群写数据状态关闭。

kinit hbase

hbase shell

set_clusterState_standby

界面提示以下信息表示执行成功：

```
hbase(main):001:0> set_clusterState_standby
=> true
```

检查当前主备同步是否完成

步骤4 执行以下命令，确保当前数据已经同步，要求SizeOfLogQueue=0，SizeOfLogToReplicate=0，如果不为零，等待，重复执行以下命令，直到等于0。

status 'replication'

关闭主备集群同步

步骤5 查询所有的同步集群，获取PEER_ID。

list_peers

步骤6 删除所有同步集群。

remove_peer '备集群ID'

示例：

remove_peer *'1'*

步骤7 查询所有同步的table。

list_replicated_tables

步骤8 分别disable上面查询到的所有同步的table。

disable_table_replication *'表名'*

示例：

disable_table_replication *'t1'*

切换主备

步骤9 重新配置HBase容灾，参考**配置HBase容灾**。

----结束

12.8.13 社区 BulkLoad Tool

Apache HBase官方网站提供了批量导入数据的功能，详细操作请参见官网对“Import”和“ImportTsv”工具的描述：**http://hbase.apache.org/2.2/book.html#tools**。

12.8.14 配置 MOB

配置场景

在实际应用中，需要存储大大小小的数据，比如图像数据、文档。小于10MB的数据一般都可以存储在HBase上，对于小于100KB的数据，HBase的读写性能是更优的。如果存放在HBase的数据大于100KB甚至到10MB大小时，插入同样个数的数据文件，但是总的数据量会很大，会导致频繁的compaction和split，占用很多CPU，磁盘IO频率很高，性能严重下降。

通过将MOB（Medium-sized Objects）数据（即100KB到10MB大小的数据）直接以HFile的格式存储在文件系统上（例如HDFS文件系统），通过expiredMobFileCleaner和Sweeper工具集中管理这些文件，然后把这些文件的地址信息及大小信息作为value存储在普通HBase的store上。这样就可以大大降低HBase的compaction和split频率，提升性能。

HBase当前默认开启MOB功能，相关配置项如**表12-170**所示。如果需要使用MOB功能，用户需要在创建表或者修改表属性时在指定的列族上指定使用mob方式存储数据。

说明

本章节适用于MRS 3.x及之后版本。

配置描述

为了开启HBase MOB功能，用户需要在创建表或者修改表属性时在指定的列族上指定使用mob方式存储数据。

使用代码声明使用mob存储的方式：

```
HColumnDescriptor hcd = new HColumnDescriptor("f");
hcd.setMobEnabled(true);
```

使用shell声明使用mob的方式，MOB_THRESHOLD单位是字节：

```
hbase(main):009:0> create 't3',{NAME => 'd', MOB_THRESHOLD => '102400', IS_MOB => 'true'}

0 row(s) in 0.3450 seconds

=> Hbase::Table - t3
hbase(main):010:0> describe 't3'
Table t3 is ENABLED


t3


COLUMN FAMILIES DESCRIPTION


{NAME => 'd', MOB_THRESHOLD => '102400', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE',
DATA_BLOCK_ENCODING => 'NONE',
TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
```

```
IN_MEMORY => 'false', IS_MOB => 'true', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536'}

1 row(s) in 0.0170 seconds
```

参数入口：

在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”。在搜索框中输入参数名称。

表 12-170 参数描述

参数	描述	默认值
hbase.mob.file.cache.size	已经打开的文件句柄的缓存区大小。如果该值设置的比较大，cache可以缓存更多的文件句柄，从而降低打开关闭文件的频率。但是如果该值设置过大会导致打开的文件句柄数过多。默认值是：“1000”。此参数在服务端ResionServer上配置。	1000
hbase.mob.cache.evict.period	缓存mob文件在mob缓存中的超期时间，单位为秒。	3600
hbase.mob.cache.evict.remain.ratio	mob cache回收之后保留的文件个数占cache容量个数的比例。hbase.mob.cache.evict.remain.ratio是一个算法因子，当缓存mob文件数达到hbase.mob.file.cache.size*hbase.mob.cache.evict.remain.ratio的大小后触发缓存回收。	0.5
hbase.master.mob.ttl.cleaner.period	过期文件清理任务的运行周期，以秒为单位。默认值是一天(86400秒)。 **说明** 如果生存时间值过期了，即文件从创建起已经超过了24小时，则MOB文件将会被过期mob文件清理工具删除。	86400

12.8.15 配置安全的 HBase Replication

配置场景

安全模式下，在交叉域设置Kerberos时，配置安全的HBase replication的过程。

前提条件

- 在Kerberos配置文件中必须定义所有FQDN映射到它的域。
- ONE.COM和TWO.COM的密码和keytab必须要一样。

操作步骤

步骤1 为两个域创建krbtgt帐户名。

比如，有ONE.COM和TWO.COM两个域，需要添加如下帐户名：krbtgt/ONE.COM@TWO.COM及krbtgt/TWO.COM@ONE.COM。

在两个域中均添加这两个帐户名。

```
kadmin: addprinc -e "<enc_type_list>" krbtgt/ONE.COM@TWO.COM
kadmin: addprinc -e "<enc_type_list>" krbtgt/TWO.COM@ONE.COM
```

说明

在这两个域之间必须至少有一个共同的keytab模式。

步骤2 在Zookeeper中，为创建短名称添加规则。

Dzookeeper.security.auth_to_local是Zookeeper服务器进程的参数。以下例子说明了如何支持ONE.COM，在帐户名中有两个成员（如service/instance@ONE.COM）。

```
Dzookeeper.security.auth_to_local=RULE:[2:\$1@\$0](.*@\\QONE.COM\\E$)s/@\\QONE.COM\\E$//DEFAULT
```

以上代码案例为在不同的域中支持ONE.COM。因此在replication中，需要为在从属集群域的主集群域添加规则。DEFAULT是已经添加了默认规则。

步骤3 在Hadoop进程中，为创建短名称添加规则。

在从属集群的HBase进程中的“core-site.xml”配置文件的属性hadoop.security.auth_to_local。比如：支持ONE.COM：

```
<property>
<name>hadoop.security.auth_to_local</name>
<value>RULE:[2:$1@$0](.*@\QONE.COM\E$)s/@\QONE.COM\E$//DEFAULT</value>
</property>
```

说明

如果启用bulkload replication功能，那么在主集群HBase进程的配置文件“core-site.xml”中需要添加支持从属域的相同属性。

例如：

```
<property>
<name>hadoop.security.auth_to_local</name>
<value>RULE:[2:$1@$0](.*@\QTWO.COM\E$)s/@\QTWO.COM\E$//DEFAULT</value>
</property>
```

----结束

12.8.16 配置 Region Transition 恢复线程

配置场景

在故障环境中，由于诸如region服务器响应慢，网络不稳定，ZooKeeper节点版本不匹配等各种原因，有可能导致region长时间处于transition下。在region transition下，由于一些region不能对外提供服务，客户端操作可能无法正常执行。

配置描述

在HMaster上设置chore服务，用于识别和恢复长期处于transition的region。

下表是用于启用此功能的配置参数。

表 12-171 参数描述

参数	描述	默认值
hbase.region.assignment.auto.recovery.enabled	配置该参数以启用或禁用region分配恢复线程功能。	true

12.8.17 使用二级索引

操作场景

HIndex为HBase提供了按照某些列的值进行索引的能力，缩小搜索范围并缩短时延。

使用约束

- 列族应以“;”分隔。
- 列和数据类型应包含在“[]”中。
- 列数据类型在列名称后使用“->”指定。
- 如果未指定列数据类型，则使用默认数据类型（字符串）。
- “#”用于在两个索引详细信息之间进行分隔。
- 以下是一个可选参数：
  -Dscan.caching：在扫描数据表时的缓存行数。
  如果不设置该参数，则默认值为1000。
- 为单个Region构建索引是为了修复损坏的索引。
  此功能不应用于生成新索引。

操作步骤

步骤1 安装HBase客户端，详情参见**使用HBase客户端**。

步骤2 进入客户端安装路径，例如“/opt/client”

cd /opt/client

步骤3 配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 执行以下命令访问Hindex。

hbase org.apache.hadoop.hbase.hindex.mapreduce.TableIndexer

表 12-172 HIndex 常用命令

说明	命令
增加索引	TableIndexer-Dtablename.to.index=table1-Dindexspecs.to.add='IDX1=>cf1:[q1->datatype],[q2],[q3];cf2:[q1->datatype],[q2->datatype]#IDX2=>cf1:[q5]'
构建索引	TableIndexer -Dtablename.to.index=table1 -Dindexnames.to.build='IDX1#IDX2'

说明	命令
删除索引	TableIndexer -Dtablename.to.index=table1 -Dindexnames.to.drop='IDX1#IDX2'
禁用索引	TableIndexer -Dtablename.to.index=table1 -Dindexnames.to.disable='IDX1#IDX2'
同时添加和构建索引	TableIndexer -Dtablename.to.index=table1 -Dindexspecs.to.add='IDX1=>cf1:[q1->datatype],[q2],[q3];cf2:[q1->datatype],[q2->datatype]#IDX2=>cf1:[q5]' -Dindexnames.to.build='IDX1'
为单个Region构建索引	TableIndexer -Dtablename.to.index=table1 -Dregion.to.index=regionEncodedName -Dindexnames.to.build='IDX1#IDX2'

说明

- **IDX1：**索引名称。
- **cf1**：列族名称。
- **q1**：列名称。
- **datatype**：数据类型，包括String，Integer，Double，Float，Long，Short，Byte，Char。

----结束

12.8.18 HBase 日志介绍

日志描述

日志存储路径：HBase相关日志的默认存储路径为“/var/log/Bigdata/hbase/*角色名*”。

- HMaster：“/var/log/Bigdata/hbase/hm”（运行日志），“/var/log/Bigdata/audit/hbase/hm”（审计日志）。
- RegionServer：“/var/log/Bigdata/hbase/rs”（运行日志），“/var/log/Bigdata/audit/hbase/rs”（审计日志）。
- ThriftServer：“/var/log/Bigdata/hbase/ts2”（运行日志，ts2为具体实例名称），“/var/log/Bigdata/audit/hbase/ts2”（审计日志，ts2为具体实例名称）。

日志归档规则：HBase的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过30MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件，压缩文件保留个数可以在Manager界面中配置。

表 12-173 HBase 日志列表

日志类型	日志文件名	描述
运行日志	hbase-<SSH_USER>-<process_name>-<hostname>.log	HBase系统日志，主要包括启动时间，启动参数信息以及HBase系统运行时候所产生的大部分日志。
	hbase-<SSH_USER>-<process_name>-<hostname>.out	HBase运行环境信息日志。
	<process_name>-<SSH_USER>-<DATE>-<PID>-gc.log	HBase服务垃圾回收日志。
	checkServiceDetail.log	HBase服务启动是否成功的检查日志。
	hbase.log	HBase服务健康检查脚本以及部分告警检查脚本执行所产生的日志。
	sendAlarm.log	HBase告警检查脚本上报告警信息日志。
	hbase-haCheck.log	HMaster主备状态检测日志。
	stop.log	HBase服务进程启停操作日志。
审计日志	hbase-audit-<process_name>.log	HBase安全审计日志。

日志级别

HBase中提供了如表12-174所示的日志级别。日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-174 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO表示记录系统及各事件正常运行状态信息
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 进入HBase服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后立即生效，不需要重启服务。

----结束

日志格式

HBase的日志格式如下所示：

表 12-175 日志格式

日志类型	组件	格式	示例
运行日志	HMaster	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-01-19 16:04:53,558 | INFO | main | env:HBASE_THRIFT_OPTS= | org.apache.hadoop.hbase.util.ServerCommandLine.logProcessInfo(ServerCommandLine.java:113)
	RegionServer	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-01-19 16:05:18,589 | INFO | regionserver16020-SendThread(linux-k6da:2181) | Client will use GSSAPI as SASL mechanism. | org.apache.zookeeper.client.ZooKeeperSaslClient$1.run(ZooKeeperSaslClient.java:285)
	ThriftServer	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-02-16 09:42:55,371 | INFO | main | loaded properties from hadoop-metrics2.properties | org.apache.hadoop.metrics2.impl.MetricsConfig.loadFirst(MetricsConfig.java:111)

日志类型	组件	格式	示例
审计日志	HMaster	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-02-16 09:42:40,934 | INFO | master:linux-k6da:16000 | Master: [master:linux-k6da:16000] start operation called. | org.apache.hadoop.hbase.master.HMaster.run(HMaster.java:581)
	RegionServer	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-02-16 09:42:51,063 | INFO | main | RegionServer: [regionserver16020] start operation called. | org.apache.hadoop.hbase.regionserver.HRegionServer.startRegionServer(HRegionServer.java:2396)
	ThriftServer	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-02-16 09:42:55,512 | INFO | main | thrift2 server start operation called. | org.apache.hadoop.hbase.thrift2.ThriftServer.main(ThriftServer.java:421)

12.8.19 HBase 性能调优

12.8.19.1 提升 BulkLoad 效率

操作场景

批量加载功能采用了MapReduce jobs直接生成符合HBase内部数据格式的文件，然后把生成的StoreFiles文件加载到正在运行的集群。使用批量加载相比直接使用HBase的API会节约更多的CPU和网络资源。

ImportTSV是一个HBase的表数据加载工具。

说明

本章节适用于MRS 3.x及之后版本。

前提条件

在执行批量加载时需要通过“Dimporttsv.bulk.output”参数指定文件的输出路径。

操作步骤

参数入口：执行批量加载任务时，在BulkLoad命令行中加入如下参数。

表 12-176 增强 BulkLoad 效率的配置项

参数	描述	配置的值
-Dimporttsv.mapper.class	用户自定义mapper通过把键值对的构造从mapper移动到reducer以帮助提高性能。mapper只需要把每一行的原始文本发送给reducer，reducer解析每一行的每一条记录并创建键值对。 **说明** 当该值配置为“org.apache.hadoop.hbase.mapreduce.TsvImporterByteMapper”时，只在执行没有*HBASE_CELL_VISIBILITY OR HBASE_CELL_TTL*选项的批量加载命令时使用。使用“org.apache.hadoop.hbase.mapreduce.TsvImporterByteMapper”时可以得到更好的性能。	org.apache.hadoop.hbase.mapreduce.TsvImporterByteMapper 和 org.apache.hadoop.hbase.mapreduce.TsvImporterTextMapper

12.8.19.2 提升连续 put 场景性能

操作场景

对大批量、连续put的场景，配置下面的两个参数为“false”时能大量提升性能。

- “hbase.regionserver.wal.durable.sync”
- “hbase.regionserver.hfile.durable.sync”

当提升性能时，缺点是对于DataNode（默认是3个）同时故障时，存在小概率数据丢失的现象。对数据可靠性要求高的场景请慎重配置。

说明

本章节适用于MRS 3.x及之后版本。

操作步骤

参数入口：

在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”，单击“全部配置”。在搜索框中输入参数名称，并进行修改。

表 12-177 提升连续 put 场景性能的参数

参数	描述	配置值
hbase.wal.hsync	设置是否启用WAL文件耐久性以将WAL数据持久化到磁盘。若将该参数设置为true，则性能将受到影响，原因是每个WAL的编辑都会被hadoop fsync同步到磁盘上。	false

参数	描述	配置值
hbase.hfile.hsync	设置是否启用Hfile耐久性以将数据持久化到磁盘。若将该参数设置为true，则性能将受到影响，原因是每个Hfile写入时都会被hadoop fsync同步到磁盘上。	false

12.8.19.3 Put 和 Scan 性能综合调优

操作场景

HBase有很多与读写性能相关的配置参数。读写请求负载不同的情况下，配置参数需要进行相应的调整，本章节旨在指导用户通过修改RegionServer配置参数进行读写性能调优。

说明

本章节适用于MRS 3.x及之后版本。

操作步骤

- JVM GC参数

  RegionServer GC_OPTS参数设置建议：

  - -Xms与-Xmx设置相同的值，需要根据实际情况设置，增大内存可以提高读写性能，可以参考参数“hfile.block.cache.size”（见**表12-179**）和参数“hbase.regionserver.global.memstore.size”（见**表12-178**）的介绍进行设置。
  - -XX:NewSize与-XX:MaxNewSize设置相同值，建议低负载场景下设置为“512M”，高负载场景下设置为“2048M”。
  - -XX:CMSInitiatingOccupancyFraction建议设置为“100 * (hfile.block.cache.size + hbase.regionserver.global.memstore.size + 0.05)”，最大值不超过90。
  - -XX:MaxDirectMemorySize表示JVM使用的堆外内存，建议低负载情况下设置为“512M”，高负载情况下设置为“2048M”。

    说明

    GC_OPTS参数中-XX:MaxDirectMemorySize默认没有配置，如需配置，用户可在GC_OPTS参数中自定义添加。

- Put相关参数

  RegionServer处理put请求的数据，会将数据写入memstore和hlog，

  - 当memstore大小达到设置的“hbase.hregion.memstore.flush.size”参数值大小时，memstore就会刷新到HDFS生成HFile。
  - 当当前region的列簇的HFile数量达到“hbase.hstore.compaction.min”参数值时会触发compaction。
  - 当当前region的列簇HFile数达到“hbase.hstore.blockingStoreFiles”参数值时会阻塞memstore刷新生成HFile的操作，导致put请求阻塞。

表 12-178 Put 相关参数

参数	描述	默认值
hbase.wal.hsync	每一条wal是否持久化到硬盘。 参考**提升连续put场景性能**。	true
hbase.hfile.hsync	hfile写是否立即持久化到硬盘。 参考**提升连续put场景性能**。	true
hbase.hregion.memstore.flush.size	若MemStore的大小（单位：Byte）超过指定值，MemStore将被冲洗至磁盘。该参数值将被运行每个hbase.server.thread.wakefrequency的线程所检验。建议设置为HDFS块大小的整数倍，在内存足够put负载大情况下可以调整增大。	134217728
hbase.regionserver.global.memstore.size	更新被锁定以及强制冲洗发生之前一个RegionServer上支持的所有MemStore的大小。建议设置为“hbase.hregion.memstore.flush.size * 写活跃region数 / RegionServer GC -Xmx”。默认值为“0.4”，表示使用RegionServer GC -Xmx的40%。	0.4
hbase.hstore.flusher.count	memstore的flush线程数，在put高负载场景下可以适当调大。	2
hbase.regionserver.thread.compaction.small	小压缩线程数，在put高负载情况下可以适当调大。	10
hbase.hstore.blockingStoreFiles	若一个Store内的HStoreFile文件数量超过指定值，则针对此HRegion的更新将被锁定直到一个压缩完成或者base.hstore.blockingWaitTime被超过。每冲洗一次MemStore一个StoreFile文件被写入。在put高负载场景下可以适当调大。	15

- Scan相关参数

表 12-179 Scan 相关参数

参数	描述	默认值
hbase.client.scanner.timeout.period	客户端和RegionServer端参数，表示客户端执行scan的租约超时时间。建议设置为60000ms的整数倍，在读高负载情况下可以适当调大。单位：毫秒。	60000
hfile.block.cache.size	数据缓存所占的RegionServer GC -Xmx百分比，在读高负载情况下可以适当调大以增大缓存命中率以提高性能。表示分配给HFile/StoreFile所使用的块缓存的最大heap（ -Xmx setting ）的百分比。	当offheap关闭时，默认值为0.25，当offheap开启时，默认值是0.1。

- Handler相关参数

表 12-180 Handler 相关参数

参数	描述	默认值
hbase.regionserver.handler.count	RegionServer上的RPC侦听器实例数，建议设置为200 ~ 400之间。	200
hbase.regionserver.metahandler.count	RegionServer中处理优先请求的程序实例的数量，建议设置为200 ~ 400之间。	200

12.8.19.4 提升实时写数据效率

操作场景

需要把数据实时写入到HBase中或者对于大批量、连续put的场景。

说明

本章节适用于MRS 3.x及之后版本。

前提条件

调用HBase的put或delete接口，把数据保存到HBase中。

操作步骤

- **写数据服务端调优**

  参数入口：

进入HBase服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**章节。

表 12-181 影响实时写数据配置项

配置参数	描述	默认值
hbase.wal.hsync	控制HLog文件在写入到HDFS时的同步程度。如果为true，HDFS在把数据写入到硬盘后才返回；如果为false，HDFS在把数据写入OS的缓存后就返回。 把该值设置为false比true在写入性能上会更优。	true
hbase.hfile.hsync	控制HFile文件在写入到HDFS时的同步程度。如果为true，HDFS在把数据写入到硬盘后才返回；如果为false，HDFS在把数据写入OS的缓存后就返回。 把该值设置为false比true在写入性能上会更优。	true

配置参数	描述	默认值
GC_OPTS	HBase利用内存完成读写操作。提高HBase内存可以有效提高HBase性能。GC_OPTS主要需要调整HeapSize的大小和NewSize的大小。调整HeapSize大小的时候，建议将Xms和Xmx设置成相同的值，这样可以避免JVM动态调整HeapSize大小的时候影响性能。调整NewSize大小的时候，建议把其设置为HeapSize大小的1/8。 ● HMaster：当HBase集群规模越大、Region数量越多时，可以适当调大HMaster的GC_OPTS参数。 ● RegionServer：RegionServer需要的内存一般比HMaster要大。在内存充足的情况下，HeapSize可以相对设置大一些。 **说明** 主HMaster的HeapSize为4G的时候，HBase集群可以支持100000 region数的规模。根据经验值，集群每增加35000个region，HeapSize增加2G，主HMaster的HeapSize不建议超过32GB。	● HMaster -server -Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=512M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=512M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M

配置参数	描述	默认值
		• Region Server -server -Xms6G -Xmx6G -XX:NewSize=1024M -XX:MaxNewSize=1024M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=512M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M

配置参数	描述	默认值
hbase.regionserver.handler.count	表示在RegionServer上启动的RPC侦听器实例数。如果设置过高会导致激烈线程竞争，如果设置过小，请求将会在RegionServer长时间等待，降低处理能力。根据资源情况，适当增加处理线程数。 建议根据CPU的使用情况，可以选择设置为100至300之间的值。	200
hbase.hregion.max.filesize	HStoreFile的最大大小（单位：Byte）。若任何一个列族HStoreFile超过此参数值，则托管Hregion将会一分为二。	10737418240
hbase.hregion.memstore.flush.size	在RegionServer中，当写操作内存中存在超过memstore.flush.size大小的memstore，则MemStoreFlusher就启动flush操作将该memstore以hfile的形式写入对应的store中。 如果RegionServer的内存充足，而且活跃Region数量也不是很多的时候，可以适当增大该值，可以减少compaction的次数，有助于提升系统性能。 同时，这种flush产生的时候，并不是紧急的flush，flush操作可能会有一定延迟，在延迟期间，写操作还可以进行，Memstore还会继续增大，最大值为“memstore.flush.size” * “hbase.hregion.memstore.block.multiplier”。当超过最大值时，将会阻塞操作。适当增大“hbase.hregion.memstore.block.multiplier”可以减少阻塞，减少性能波动。单位：字节。	134217728

配置参数	描述	默认值
hbase.regionserver.global.memstore.size	更新被锁定以及强制冲洗发生之前一个RegionServer上支持的所有MemStore的大小。RegionServer中，负责flush操作的是MemStoreFlusher线程。该线程定期检查写操作内存，当写操作占用内存总量达到阈值，MemStoreFlusher将启动flush操作，按照从大到小的顺序，flush若干相对较大的memstore，直到所占用内存小于阈值。 阈值 = “hbase.regionserver.global.memstore.size” * “hbase.regionserver.global.memstore.size.lower.limit” * “HBase_HEAPSIZE” **说明** 该配置与“hfile.block.cache.size”的和不能超过0.8，也就是写和读操作的内存不能超过HeapSize的80%，这样可以保证除读和写外其它操作的正常运行。	0.4
hbase.hstore.blockingStoreFiles	在region flush前首先判断file文件个数，是否大于hbase.hstore.blockingStoreFiles。 如果大于需要先compaction并且让flush延时90s（这个值可以通过hbase.hstore.blockingWaitTime进行配置），在延时过程中，将会继续写从而使得Memstore还会继续增大超过最大值 “memstore.flush.size” * “hbase.hregion.memstore.block.multiplier”，导致写操作阻塞。当完成compaction后，可能就会产生大量写入。这样就导致性能激烈震荡。 增加hbase.hstore.blockingStoreFiles，可以减低BLOCK几率。	15
hbase.regionserver.thread.compaction.throttle	大于此参数值的压缩将被大线程池执行，单位：Byte。控制一次Minor Compaction时，进行compaction的文件总大小的阈值。Compaction时的文件总大小会影响这一次compaction的执行时间，如果太大，可能会阻塞其它的compaction或flush操作。	1610612736

配置参数	描述	默认值
hbase.hstore.compaction.min	每次执行minor compaction的HStoreFile的最小数量。当一个Store中文件超过该值时，会进行compact，适当增大该值，可以减少文件被重复执行compaction。但是如果过大，会导致Store中文件数过多而影响读取的性能。	6
hbase.hstore.compaction.max	每次执行minor compaction的HStoreFile的最大数量。与“hbase.hstore.compaction.max.size”的作用基本相同，主要是控制一次compaction操作的时间不要太长。	10
hbase.hstore.compaction.max.size	如果一个HFile文件的大小大于该值，那么在Minor Compaction操作中不会选择这个文件进行compaction操作，除非进行Major Compaction操作。 这个值可以防止较大的HFile参与compaction操作。在禁止Major Compaction后，一个Store中可能存在几个HFile，而不会合并成为一个HFile，这样不会对数据读取造成太大的性能影响。单位：字节。	9223372036854775807
hbase.hregion.majorcompaction	单个区域内所有HStoreFile文件主压缩的时间间隔，单位：毫秒。由于执行Major Compaction会占用较多的系统资源，如果正在处于系统繁忙时期，会影响系统的性能。 如果业务没有较多的更新、删除、回收过期数据空间时，可以把该值设置为0，以禁止Major Compaction。 如果必须要执行Major Compaction，以回收更多的空间，可以适当增加该值，同时配置参数“hbase.offpeak.end.hour”和“hbase.offpeak.start.hour”以控制Major Compaction发生在业务空闲的时期。单位：毫秒。	604800000

配置参数	描述	默认值
• hbase.regionserver.maxlogs • hbase.regionserver.hlog.blocksize	• 表示一个RegionServer上未进行Flush的Hlog的文件数量的阈值，如果大于该值，RegionServer会强制进行flush操作。 • 表示每个HLog文件的最大大小。如果HLog文件大小大于该值，就会滚动出一个新的HLog文件，旧的将被禁用并归档。 这两个参数共同决定了RegionServer中可以存在的未进行Flush的hlog数量。当这个数据量小于MemStore的总大小的时候，会出现由于HLog文件过多而触发的强制flush操作。这个时候可以适当调整这两个参数的大小，以避免出现这种强制flush的情况。单位：字节。	• 32 • 134217728

- **写数据客户端调优**

  写数据时，在场景允许的情况下，需要使用Put List的方式，可以极大的提升写性能。每一次Put的List的长度，需要结合单条Put的大小，以及实际环境的一些参数进行设定。建议在选定之前先做一些基础的测试。

- **写数据表设计调优**

  表 12-182 影响实时写数据相关参数

配置参数	描述	默认值
COMPRESSION	配置数据的压缩算法，这里的压缩是HFile中block级别的压缩。对于可以压缩的数据，配置压缩算法可以有效减少磁盘的IO，从而达到提高性能的目的。 **说明** 并非所有数据都可以进行有效压缩。例如一张图片的数据，因为图片一般已经是压缩后的数据，所以压缩效果有限。常用的压缩算法是SNAPPY，因为它有较好的Encoding/Decoding速度和可以接受的压缩率。	NONE
BLOCKSIZE	配置HFile中block块的大小，不同的block块大小，可以影响HBase读写数据的效率。越大的block块，配合压缩算法，压缩的效率就越好；但是由于HBase的读取数据是以block块为单位的，所以越大的block块，对于随机读的情况，性能可能会比较差。 如果要提升写入的性能，一般扩大到128KB或者256KB，可以提升写数据的效率，也不会影响太大的随机读性能。单位：字节	65536

配置参数	描述	默认值
IN_MEMORY	配置这个表的数据优先缓存在内存中，这样可以有效提升读取的性能。对于一些小表，而且需要频繁进行读取操作的，可以设置此配置项。	false

12.8.19.5 提升实时读数据效率

操作场景

需要读取HBase数据场景。

前提条件

调用HBase的get或scan接口，从HBase中实时读取数据。

操作步骤

- **读数据服务端调优**

  参数入口：

  进入HBase服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**章节。

表 12-183 影响实时读数据配置项

配置参数	描述	默认值
GC_OPTS	HBase利用内存完成读写操作。提高HBase内存可以有效提高HBase性能。 GC_OPTS主要需要调整HeapSize的大小和NewSize的大小。调整HeapSize大小的时候，建议将Xms和Xmx设置成相同的值，这样可以避免JVM动态调整HeapSize大小的时候影响性能。调整NewSize大小的时候，建议把其设置为HeapSize大小的1/8。 ● HMaster：当HBase集群规模越大、Region数量越多时，可以适当调大HMaster的GC_OPTS参数。 ● RegionServer：RegionServer需要的内存一般比HMaster要大。在内存充足的情况下，HeapSize可以相对设置大一些。 **说明** 主HMaster的HeapSize为4G的时候，HBase集群可以支持100000 region数的规模。根据经验值，集群每增加35000个region，HeapSize增加2G，主HMaster的HeapSize不建议超过32GB。	MRS 3.x之前版本： ● HMaster：-server -Xms2G -Xmx2G -XX:NewSize=256M -XX:MaxNewSize=256M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=512M -XX:MaxDirectMemorySize=512M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -

配置参数	描述	默认值
		XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M • RegionServer: -server -Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=512M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=512M -XX:MaxDirectMemorySize=512M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7F

配置参数	描述	默认值
		FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M MRS 3.x及之后版本： ● HMaster -server -Xms4G -Xmx4G -XX:NewSize=512M -XX:MaxNewSize=512M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=512M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dg

配置参数	描述	默认值
		c.client.gcInterval=0x7FFFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M • RegionServer -server -Xms6G -Xmx6G -XX:NewSize=1024M -XX:MaxNewSize=1024M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=512M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccup

配置参数	描述	默认值
		ancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M
hbase.regionserver.handler.count	表示RegionServer在同一时刻能够并发处理多少请求。如果设置过高会导致激烈线程竞争，如果设置过小，请求将会在RegionServer长时间等待，降低处理能力。根据资源情况，适当增加处理线程数。 建议根据CPU的使用情况，可以选择设置为100至300之间的值。	200
hfile.block.cache.size	HBase缓存区大小，主要影响查询性能。根据查询模式以及查询记录分布情况来决定缓存区的大小。如果采用随机查询使得缓存区的命中率较低，可以适当降低缓存区大小。	当offheap关闭时，默认值为0.25。当offheap开启时，默认值是0.1。

说明

如果同时存在读和写的操作，这两种操作的性能会互相影响。如果写入导致的flush和Compaction操作频繁发生，会占用大量的磁盘IO操作，从而影响读取的性能。如果写入导致阻塞较多的Compaction操作，就会出现Region中存在多个HFile的情况，从而影响读取的性能。所以如果读取的性能不理想的时候，也要考虑写入的配置是否合理。

- **读数据客户端调优**

  Scan数据时需要设置caching（一次从服务端读取的记录条数，默认是1），若使用默认值读性能会降到极低。

  当不需要读一条数据所有的列时，需要指定读取的列，以减少网络IO。

  只读取RowKey时，可以为Scan添加一个只读取RowKey的filter（FirstKeyOnlyFilter或KeyOnlyFilter）。

- **读数据表设计调优**

  表 12-184 影响实时读数据相关参数

配置参数	描述	默认值
COMPRESSION	配置数据的压缩算法，这里的压缩是HFile中block级别的压缩。对于可以压缩的数据，配置压缩算法可以有效减少磁盘的IO，从而达到提高性能的目的。 **说明** 并非所有数据都可以进行有效压缩。例如一张图片的数据，因为图片一般已经是压缩后的数据，所以压缩效果有限。常用的压缩算法是SNAPPY，因为它有较好的Encoding/Decoding速度和可以接受的压缩率。	NONE
BLOCKSIZE	配置HFile中block块的大小，不同的block块大小，可以影响HBase读写数据的效率。越大的block块，配合压缩算法，压缩的效率就越好；但是由于HBase的读取数据是以block块为单位的，所以越大的block块，对于随机读的情况，性能可能会比较差。 如果要提升写入的性能，一般扩大到128KB或者256KB，可以提升写数据的效率，也不会影响太大的随机读性能。单位：字节。	65536
DATA_BLOCK_ENCODING	配置HFile中block块的编码方法。当一行数据中存在多列时，一般可以配置为“FAST_DIFF”，可以有效的节省数据存储的空间，从而提供性能。	NONE

12.8.19.6 JVM 参数优化

操作场景

当集群数据量达到一定规模后，JVM的默认配置将无法满足集群的业务需求，轻则集群变慢，重则集群服务不可用。所以需要根据实际的业务情况进行合理的JVM参数配置，提高集群性能。

操作步骤

参数入口：

HBase角色相关的JVM参数需要配置在安装有HBase服务的节点的“${BIGDATA_HOME}/FusionInsight_HD_*/install/FusionInsight-HBase-2.2.3/hbase/conf/”目录下的“hbase-env.sh”文件中。

每个角色都有各自的JVM参数配置变量，如表12-185。

表 12-185 HBase 相关 JVM 参数配置变量

变量名	变量影响的角色
HBASE_OPTS	该变量中设置的参数，将影响HBase的所有角色。
SERVER_GC_OPTS	该变量中设置的参数，将影响HBase Server端的所有角色，例如：Master、RegionServer等。
CLIENT_GC_OPTS	该变量中设置的参数，将影响HBase的Client进程。
HBASE_MASTER_OPTS	该变量中设置的参数，将影响HBase的Master。
HBASE_REGIONSERVER_OPTS	该变量中设置的参数，将影响HBase的RegionServer。
HBASE_THRIFT_OPTS	该变量中设置的参数，将影响HBase的Thrift。

配置方式举例：

```
export HADOOP_NAMENODE_OPTS="-Dhadoop.security.logger=${HADOOP_SECURITY_LOGGER:-INFO,RFAS} -Dhdfs.audit.logger=${HDFS_AUDIT_LOGGER:-INFO,NullAppender} $HADOOP_NAMENODE_OPTS"
```

12.8.20 HBase 常见问题

12.8.20.1 客户端连接服务端时，长时间无法连接成功

问题

在HBase服务端出现问题，无法提供服务，此时HBase客户端进行表操作，会出现该操作挂起，长时间无任何反应。

回答

问题分析

当HBase服务端出现问题，HBase客户端进行表操作的时候，会进行重试，并等待超时。该超时默认值为Integer.MAX_VALUE (2147483647 ms)，所以HBase客户端会在这么长的时间内一直重试，造成挂起表象。

解决方法

HBase客户端提供两个配置项来控制客户端的重试超时方式，如表12-186。

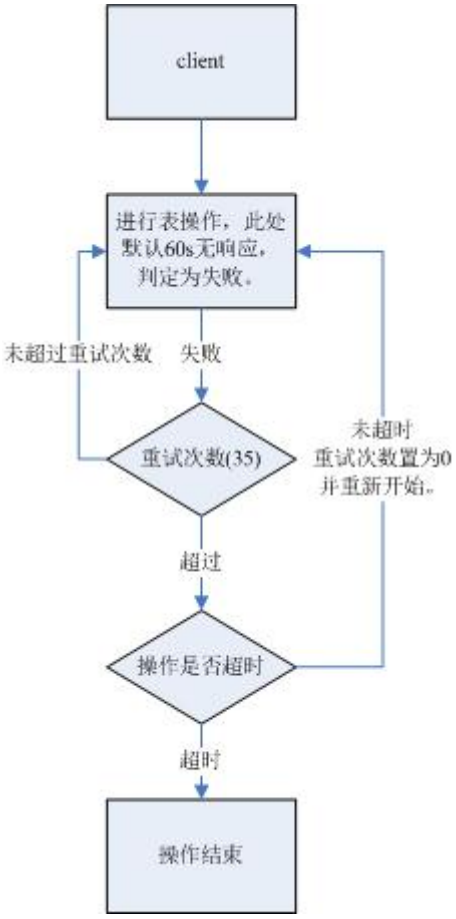
在“*客户端安装路径*/HBase/hbase/conf/hbase-site.xml”配置文件中配置如下参数。

表 12-186 HBase 客户端操作重试超时相关配置

配置参数	描述	默认值
hbase.client.operation.timeout	客户端操作超时时间。需在配置文件中手动添加。	2147483647 ms
hbase.client.retries.number	最大重试次数。用于表示所有可重试操作所支持的最大重试次数。	35

这两个参数的重试超时的配合方式如图12-15所示。

图 12-15 HBase 客户端操作重试超时流程



从该流程可以看出，如果未对这两个配置参数根据具体使用场景进行配置，会造成挂起迹象。建议根据使用场景，配置合适的超时时间，如果是长时间操作，则把超时时间设置长一点；如果是短时间操作，则把超时时间设置短一点。而重试次数可以设置为：“(hbase.client.retries.number)*60*1000(ms)”。刚好大于“hbase.client.operation.timeout”设置的超时时间。

12.8.20.2 结束 BulkLoad 客户端程序，导致作业执行失败

问题

执行BulkLoad程序导入数据时，如果结束客户端程序，为什么有时会导致已提交的作业执行失败?

回答

BulkLoad程序在客户端启动时会生成一个partitioner文件，用于划分Map任务数据输入的范围。此文件在BulkLoad客户端退出时会被自动删除。一般来说当所有Map任务都启动运行以后，退出BulkLoad客户端也不会导致已提交的作业失败。但由于Map任务存在重试机制和推测执行机制；Reduce任务下载一个已运行完成的Map任务的数据失败次数过多时，Map任务也会被重新执行。如果此时BulkLoad客户端已经退出，则重试的Map任务会因为找不到partitioner文件而执行失败，导致作业执行失败。因此，强烈建议BulkLoad程序在数据导入期间不要结束客户端程序。

12.8.20.3 在 HBase 连续对同一个表名做删除创建操作时，可能出现创建表异常

问题

在HBase连续对同一个表名做删除创建操作时，可能出现创建表异常。

回答

执行过程：Disable Table > Drop Table > Create Table > Disable Table > Drop Table >...

1. 在Disable表时，HMaster会发送RPC请求到RegionServer，RegionServer会将相关Region下线。当RegionServer上的Region关闭所需的时间超过HBase的HMaster等待Region处于RIT状态的超时时间，HMaster会默认该Region下线，实际上该Region可能还处在flush memstore阶段。
2. 发送RPC请求关闭Region之后，HMaster会判断该表的所有Region是否下线，上述1的情况下关闭超时也会认为是下线，然后HMaster返回关闭成功。
3. 关闭成功之后，删除表，HBase表对应的数据目录被删掉。
4. 在删除表之后，该数据目录会被还处于flush memstore阶段的Region重新创建。
5. 再创建该表时，将temp目录拷贝到HBase数据目录时，由于HBase数据目录不为空，导致调用HDFS rename接口时，数据目录变为temp目录最后一层追加到HBase的数据目录下，如$rootDir/data/$nameSpace/$tableName/$tableName，那样创建表就会失败。

解决办法：

出现该问题时，请检查该表对应的HBase数据目录是否存在，如果存在请将该目录重命名。

HBase数据目录由$rootDir/data/$nameSpace/$tableName组成，例如“hdfs://hacluster/hbase/data/default/TestTable”，其中$rootDir是HBase的根目录，该值通过在“hbase-site.xml”中配置hbase.rootdir.perms得到，data目录是HBase的固定目录，$nameSpace是nameSpace名字，$tableName是表名。

12.8.20.4 HBase 占用网络端口，连接数过大会导致其他服务不稳定

问题

HBase占用网络端口，连接数过大会导致其他服务不稳定。

回答

使用操作系统命令***lsof***或者***netstat***发现大量TCP连接处于CLOSE_WAIT状态，且连接持有者为HBase RegionServer，可能导致网络端口耗尽或HDFS连接超限，那样可能会导致其他服务不稳定。HBase CLOSE_WAIT现象为HBase机制。

HBase CLOSE_WAIT产生原因：HBase数据以HFile形式存储在HDFS上，这里可以叫StoreFiles，HBase作为HDFS的客户端，HBase在创建StoreFile或启动加载StoreFile时创建了HDFS连接，当创建StoreFile或加载StoreFile完成时，HDFS方面认为任务已完成，将连接关闭权交给HBase，但HBase为了保证实时响应，有请求时就可以连接对应数据文件，需要保持连接，选择不关闭连接，所以连接状态为CLOSE_WAIT（需客户端关闭）。

什么时候会创建StoreFile：当HBase执行Flush时。

什么时候执行Flush：HBase写入数据首先会存在内存memstore，只有内存使用达到阈值或手动执行***flush***命令时会触发flush操作，将数据写入HDFS。

解决方法：

由于HBase连接机制，若想减小HBase端口占用，则需控制StoreFile数量，具体可以通过触发HBase的compaction动作完成，即触发HBase文件合并，方法如下：

方法1：使用HBase shell客户端，在客户端手动执行***major_compact***操作。

方法2：编写HBase客户端代码，调用HBaseAdmin类中的compact方法触发HBase的compaction动作。

如果compact无法解决HBase端口占用现象，说明HBase使用情况已经达到瓶颈，需考虑如下几点：

- table的Region数初始设置是否合适。
- 是否存在无用数据。

若存在无用数据，可删除对应数据以减小HBase存储文件数量，若以上情况都不满足，则需考虑扩容。

12.8.20.5 HBase bulkload 任务（单个表有 26T 数据）有 210000 个 map 和 10000 个 reduce，任务失败

问题

MRS 3.x及之后版本HBase bulkLoad任务（单个表有26T数据）有210000个map和10000个reduce，任务失败。

回答

ZooKeeper IO瓶颈观测手段:

1. 通过Manager的监控页面查看单个节点上ZooKeeper请求监控，判断是否严重超出规格限制。
2. 通过观测ZooKeeper的日志以及HBase的日志，查看是否有大量的IO Exception Timeout或者SocketTimeout Exception异常。

调优建议:

1. 将ZooKeeper实例个数调整为5个及以上，可以通过设置peerType=observer来增加observer的数目。
2. 通过控制单个任务并发的map数或减少每个节点下运行task的内存，降低节点负载。
3. 升级ZooKeeper数据磁盘，如SSD等。

12.8.20.6 如何修复长时间处于 RIT 状态的 Region

问题

在HBase WEBUI界面看到有长时间处于RIT状态的Region，如何修复?

回答

登录HMaster WebUI，在导航栏选择“Procedure & Locks”，查看是否有处于Waiting状态的process id。如果有，需要执行以下命令将procedure lock释放:

hbase hbck -j *客户端安装目录*/**HBase/hbase/tools/hbase-hbck2-*.jar bypass -o** *pid*

查看State是否处于Bypass状态，如果界面上的procedures一直处于RUNNABLE(Bypass)状态，需要进行主备切换。执行**assigns**命令使region重新上线。

hbase hbck -j *客户端安装目录*/**HBase/hbase/tools/hbase-hbck2-*.jar assigns -o** *regionName*

12.8.20.7 HMaster 等待 namespace 表上线时超时退出

问题

为什么在等待namespace表上线时超时HMaster退出?

回答

在HMaster主备倒换或启动期间，HMaster为先前失败/停用的RegionServer执行WAL splitting及region恢复。

在后台运行有多个监控HMaster启动进程的线程:

- TableNamespaceManager

  这是一个帮助类，用于在HMaster主备倒换或启动期间，管理namespace表及监控表region的分配。如果namespace表在规定时间

（hbase.master.namespace.init.timeout，默认为3600000ms）内没有上线，那么它就会异常中断HMaster进程。

- InitializationMonitor

  这是一个主HMaster初始化线程监控类，用于监控主Master的初始化。如果在规定时间（hbase.master.initializationmonitor.timeout，默认为3600000ms）内初始化线程失败，该线程会异常终止HMaster（如果该hbase.master.initializationmonitor.haltontimeout被启动，默认为false）。

在HMaster主备倒换或启动期间，如果WAL hlog文件存在，它会初始化WAL splitting任务。如果WAL hlog splitting任务完成，它将初始化表region分配任务。

HMaster通过ZooKeeper协调log splitting任务和有效的RegionServer，并追踪任务的发展。如果主HMaster在log splitting任务期间退出，新的主HMaster会尝试重发没有完成的任务，RegionServer从头启动log splitting任务。

HMaster初始化工作完成情况会由于很多原因被延迟：

- 间歇性的网络故障。
- 磁盘瓶颈。
- log split任务工作负荷较大，RegionServer运行缓慢。
- RegionServer（region openning）响应缓慢。

在以上场景中，为使HMaster更早完成恢复任务，建议增加以下配置参数，否则Master将退出导致整个恢复进程被更大程度地延迟。

- 增加namespace表在线等待超时周期，保证Master有足够的时间协调RegionServer workers split任务，避免一次次重复相同的任务。

  “hbase.master.namespace.init.timeout”（默认为3600000ms）
- 通过RegionServer worker增加并行split任务执行数，保证RegionServer worker能并行处理split work（RegionServer需要有更多的核心）。在“*客户端安装路径*/HBase/hbase/conf/hbase-site.xml”中添加参数：

  “hbase.regionserver.wal.max.splitters”（默认为2）
- 如果所有的恢复过程都需要时间，增加初始化监控线程超时时间。

  “hbase.master.initializationmonitor.timeout”（默认为3600000ms）

12.8.20.8 客户端查询 HBase 出现 SocketTimeoutException 异常

问题

使用HBase客户端操作表数据的时候客户端出现类似如下异常：

```
2015-12-15 02:41:14,054 | WARN  | [task-result-getter-2] | Lost task 2.0 in stage 58.0 (TID 3288, linux-175): 
org.apache.hadoop.hbase.client.RetriesExhaustedException: Failed after attempts=36, exceptions:
Tue Dec 15 02:41:14 CST 2015, null, java.net.SocketTimeoutException: callTimeout=60000, 
callDuration=60303:
row 'xxxxxx' on table 'xxxxxx' at region=xxxxxx,\x05\x1E
\x80\x00\x00\x00\x80\x00\x00\x00\x00\x00\x00\x00\x80\x00\x00\x00\x00\x00\x00\x000\x00\x80\x00\x00\x0
0\x80\x00\x00\x00\x80\x00\x00, 
1449912620868.6a6b7d0c272803d8186930a3bfdb10a9., hostname=xxxxxx,16020,1449941841479, 
seqNum=5
at 
org.apache.hadoop.hbase.client.RpcRetryingCallerWithReadReplicas.throwEnrichedException(RpcRetryingCall
erWithReadReplicas.java:275)
at org.apache.hadoop.hbase.client.ScannerCallableWithReplicas.call(ScannerCallableWithReplicas.java:223)
at org.apache.hadoop.hbase.client.ScannerCallableWithReplicas.call(ScannerCallableWithReplicas.java:61)
```

```
at org.apache.hadoop.hbase.client.RpcRetryingCaller.callWithoutRetries(RpcRetryingCaller.java:200)
at org.apache.hadoop.hbase.client.ClientScanner.call(ClientScanner.java:323)
```

同时，在RegionServer上出现类似如下日志：

```
2015-12-15 02:45:44,551 | WARN  | PriorityRpcServer.handler=7,queue=1,port=16020 | (responseTooSlow):
{"call":"Scan(org.apache.hadoop.hbase.protobuf.generated.ClientProtos$ScanRequest)
","starttimems":1450118730780,"responsesize":416,"method":"Scan","processingtimems":13770,"client":"10.9
1.8.175:41182","queuetimems":0,"class":"HRegionServer"} |
org.apache.hadoop.hbase.ipc.RpcServer.logResponse(RpcServer.java:2221)
2015-12-15 02:45:57,722 | WARN  | PriorityRpcServer.handler=3,queue=1,port=16020 | (responseTooSlow):
{"call":"Scan(org.apache.hadoop.hbase.protobuf.generated.ClientProtos
$ScanRequest)","starttimems":1450118746297,"responsesize":416,
"method":"Scan","processingtimems":11425,"client":"10.91.8.175:41182","queuetimems":1746,"class":"HRegi
onServer"} | org.apache.hadoop.hbase.ipc.RpcServer.logResponse(RpcServer.java:2221)
2015-12-15 02:47:21,668 | INFO  | LruBlockCacheStatsExecutor | totalSize=7.54 GB, freeSize=369.52 MB,
max=7.90 GB, blockCount=406107,
accesses=35400006, hits=16803205, hitRatio=47.47%, , cachingAccesses=31864266, cachingHits=14806045,
cachingHitsRatio=46.47%,
evictions=17654, evicted=16642283, evictedPerRun=942.69189453125 |
org.apache.hadoop.hbase.io.hfile.LruBlockCache.logStats(LruBlockCache.java:858)
2015-12-15 02:52:21,668 | INFO  | LruBlockCacheStatsExecutor | totalSize=7.51 GB, freeSize=395.34 MB,
max=7.90 GB, blockCount=403080,
accesses=35685793, hits=16933684, hitRatio=47.45%, , cachingAccesses=32150053, cachingHits=14936524,
cachingHitsRatio=46.46%,
evictions=17684, evicted=16800617, evictedPerRun=950.046142578125 |
org.apache.hadoop.hbase.io.hfile.LruBlockCache.logStats(LruBlockCache.java:858)
```

回答

出现该问题的主要原因为RegionServer分配的内存过小、Region数量过大导致在运行过程中内存不足，服务端对客户端的响应过慢。在RegionServer的配置文件“hbase-site.xml”中需要调整如下对应的内存分配参数。

表 12-187 RegionServer 内存调整参数

参数	描述	默认值
GC_OPTS	在启动参数中给RegionServer分配的初始内存和最大内存。	-Xms8G -Xmx8G
hfile.block.cache.size	分配给HFile/StoreFile所使用的块缓存的最大heap（-Xmx setting）的百分比。	当offheap关闭时，默认值为0.25。当offheap开启时，默认值是0.1。

12.8.20.9 使用 scan 命令仍然可以查询到已修改和已删除的数据

问题

为什么使用如下**scan**命令仍然可以查询到已修改和已删除的数据?

```
scan '<table_name>',{FILTER=>"SingleColumnValueFilter('<column_family>','column',=,'binary:<value>')"}
```

回答

由于HBase的可扩展性，在查询表的时候，默认情况下会匹配被查询列的所有版本的值，即使被删除或被修改的值也可以查询出来。对于命中列失败的行（即在某一行中不存在该列），HBase会将该行查询出来。

如果用户仅需查询该表的最新值和命中列成功的行，可使用如下查询语句：

```
scan '<table_name>',
{FILTER=>"SingleColumnValueFilter('<column_family>','column',=,'binary:<value>',true,true)"}
```

使用该命令，不但可以过滤掉命中列失败的行，而且查询的是表的当前数据的最新版本的值，即不查询被修改之前的值和被删除的值。

说明

过滤器SingleColumnValueFilter的相关参数说明如下：

SingleColumnValueFilter(final byte[] family, final byte[] qualifier, final CompareOp compareOp, ByteArrayComparable comparator, final boolean filterIfMissing, final boolean latestVersionOnly)

参数说明：

- family：需要查询的列所在的列族；
- qualifier：需要查询的列；
- compareOp：比较符，如“=”、“>”等等；
- comparator：需要查找的目标值；
- filterIfMissing：如果某一行不存在该列，是否过滤，默认值为false；
- latestVersionOnly：是否仅查询最新版本的值，默认值为false。

12.8.20.10 在启动 HBase shell 时，为什么会抛出“java.lang.UnsatisfiedLinkError: Permission denied”异常

问题

在启动HBase shell时，为什么会抛出“java.lang.UnsatisfiedLinkError: Permission denied”异常?

回答

在执行HBase shell期间，JRuby会在“java.io.tmpdir”路径下创建一个临时文件，该路径的默认值为“/tmp”。如果为“/tmp”目录设置NOEXEC权限，然后HBase shell会启动失败并抛出“java.lang.UnsatisfiedLinkError: Permission denied”异常。

因此，如果为“/tmp”目录设置了NOEXEC权限，那么“java.io.tmpdir”必须设置为HBASE_OPTS/CLIENT_GC_OPTS中不同的路径。

12.8.20.11 在 HMaster Web UI 中显示处于“Dead Region Servers”状态的 RegionServer 什么时候会被清除掉

问题

在HMaster Web UI中显示处于“Dead Region Servers”状态的RegionServer什么时候会被清除掉?

回答

当一个在线的RegionServer突然运行停止，会在HMaster Web UI中显示处于“Dead Region Servers”状态。当停止运行的RegionServer重启并且向HMaster上报成功信息，在HMaster Web UI中会清除掉“Dead Region Servers”信息。

当HMaster主备倒换操作成功执行时，在HMaster Web UI中也会清除掉“Dead Region Servers”信息。

以防掌控有一些region的主用HMaster突然停止响应，备用的HMaster将会成为新的主用HMaster，同时显示先前主用HMaster变成dead RegionServer。当HMaster主备倒换操作成功执行，在HMaster Web UI中也会清除掉“Dead Region Servers”。

12.8.20.12 使用 HBase bulkload 导入数据成功，执行相同的查询时却可能返回不同的结果

问题

在使用HBase bulkload导入数据时，如果导入的数据存在相同的rowkey值，数据可以导入成功，但是执行相同的查询时可能返回不同的结果。

回答

正常情况下，相同rowkey值的数据加载到HBase是有先后顺序的，HBase以最近的时间戳的数据为最新数据，一般的默认查询中，没有指定时间戳的，就会对相同rowkey值的数据仅返回最新数据。

使用bulkload加载数据，由于数据在内存中处理生成HFile，速度是很快的，很可能出现相同rowkey值的数据具有相同时间戳，从而造成查询结果混乱的情况。

建议在建表和数据加载时，设计好rowkey值，尽量避免在同一个数据文件中存在相同rowkey值的情况。

12.8.20.13 如何处理由于 Region 处于 FAILED_OPEN 状态而造成的建表失败异常

问题

如何处理由于Region处于FAILED_OPEN状态而造成的建表失败异常。

回答

建表过程中如果发生网络故障、HDFS故障或者Active HMaster故障等情况时，可能会造成部分Region上线失败而处于FAILED_OPEN状态，导致建表失败。

由于Region上线失败而处于FAILED_OPEN状态造成的建表失败异常不能直接修复，需要删除该表后重新建表。

操作步骤如下：

1. 在集群客户端使用如下命令修复表的状态。
   hbase hbck -fixTableStates
2. 进入HBase shell并执行以下命令完成表的清理。
   truncate *'<table_name>'*
   disable *'<table_name>'*

drop *'<table_name>'*

3. 使用建表命令重新创建该表。

12.8.20.14 如何清理由于建表失败残留在 ZooKeeper 中/hbase/table-lock 目录下的表名

问题

安全模式下，由于建表失败，在ZooKeeper的table-lock节点（默认路径/hbase/table-lock）下残留有新建的表名，请问该如何清理?

回答

操作步骤如下：

1. 在安装好客户端的环境下，使用hbase用户进行kinit认证。
2. 执行***hbase zkcli***命令进入ZooKeeper命令行。
3. 在ZooKeeper命令行中执行***ls /hbase/table***，查看新建的表名是否存在。
   - 是，结束。
   - 否，执行***ls /hbase/table-lock***查看新建的表名是否存在，若存在新建的表名时使用***delete***命令（***delete /hbase/table-lock/<table>***，其中<table>为残留的表名）删除该表名。

12.8.20.15 为什么给 HDFS 上的 HBase 使用的目录设置 quota 会造成 HBase 故障

问题

为什么给HDFS上的HBase使用的目录设置quota会造成HBase故障?

回答

表的flush操作是在HDFS中写memstore数据。

如果HDFS目录没有足够的磁盘空间quota，flush操作会失败，这样region server将会终止。

```
Caused by: org.apache.hadoop.hdfs.protocol.DSQuotaExceededException: The DiskSpace quota of /hbase/data/<namespace>/<tableName> is exceeded: quota = 1024 B = 1 KB but diskspace consumed = 402655638 B = 384.00 MB
?at org.apache.hadoop.hdfs.server.namenode.DirectoryWithQuotaFeature.verifyStoragespaceQuota(DirectoryWithQuotaFeature.java:211)
?at org.apache.hadoop.hdfs.server.namenode.DirectoryWithQuotaFeature.verifyQuota(DirectoryWithQuotaFeature.java:239)
?at org.apache.hadoop.hdfs.server.namenode.FSDirectory.verifyQuota(FSDirectory.java:882)
?at org.apache.hadoop.hdfs.server.namenode.FSDirectory.updateCount(FSDirectory.java:711)
?at org.apache.hadoop.hdfs.server.namenode.FSDirectory.updateCount(FSDirectory.java:670)
?at org.apache.hadoop.hdfs.server.namenode.FSDirectory.addBlock(FSDirectory.java:495)
```

上述异常中，表“/hbase/data/<namespace>/<tableName>”的磁盘空间quota值为1KB，但是memstore数据为384.00MB，所以flush操作失败并且region server会终止。

在region server终止时，HMaster对终止的region server的WAL文件进行replay操作以恢复数据。由于限制了磁盘空间quota值，导致WAL文件的replay操作失败进而导致HMaster进程异常退出。

```
2016-07-28 19:11:40,352 | FATAL | MASTER_SERVER_OPERATIONS-10-91-9-131:16000-0 | Caught throwable while processing event M_SERVER_SHUTDOWN |
org.apache.hadoop.hbase.master.HMaster.abort(HMaster.java:2474)
java.io.IOException: failed log splitting for 10-91-9-131,16020,1469689987884, will retry
?at org.apache.hadoop.hbase.master.handler.ServerShutdownHandler.resubmit(ServerShutdownHandler.java:365)
?at org.apache.hadoop.hbase.master.handler.ServerShutdownHandler.process(ServerShutdownHandler.java:220)
?at org.apache.hadoop.hbase.executor.EventHandler.run(EventHandler.java:129)
?at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
?at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
?at java.lang.Thread.run(Thread.java:745)
Caused by: java.io.IOException: error or interrupted while splitting logs in [hdfs://hacluster/hbase/WALs/<RS-Hostname>,<RS-Port>,<startcode>-splitting] Task = installed = 6 done = 3 error = 3
?at org.apache.hadoop.hbase.master.SplitLogManager.splitLogDistributed(SplitLogManager.java:290)
?at org.apache.hadoop.hbase.master.MasterFileSystem.splitLog(MasterFileSystem.java:402)
?at org.apache.hadoop.hbase.master.MasterFileSystem.splitLog(MasterFileSystem.java:375)
```

因此，不支持用户对HDFS上的HBase目录进行quota值设置。上述问题可通过下述步骤解决：

步骤1 在客户端命令提示符下运行**kinit** *用户名*命令，使HBase用户获得安全认证。

步骤2 运行**hdfs dfs -count -q** */hbase/data/<namespace>/<tableName>*命令检查分配的磁盘空间quota。

步骤3 使用下列命令取消quota值限制，恢复HBase。

hdfs dfsadmin -clrSpaceQuota */hbase/data/<namespace>/<tableName>*

----结束

12.8.20.16 为什么在使用 OfflineMetaRepair 工具重新构建元数据后，HMaster 启动的时候会等待 namespace 表分配超时，最后启动失败

问题

为什么在使用OfflineMetaRepair工具重新构建元数据后，HMaster启动的时候会等待namespace表分配超时，最后启动失败?

且HMaster将输出下列FATAL消息表示中止：

```
2017-06-15 15:11:07,582 FATAL [Hostname:16000.activeMasterManager] master.HMaster: Unhandled exception. Starting shutdown.
java.io.IOException: Timedout 120000ms waiting for namespace table to be assigned
    at org.apache.hadoop.hbase.master.TableNamespaceManager.start(TableNamespaceManager.java:98)
    at org.apache.hadoop.hbase.master.HMaster.initNamespace(HMaster.java:1054)
    at org.apache.hadoop.hbase.master.HMaster.finishActiveMasterInitialization(HMaster.java:848)
    at org.apache.hadoop.hbase.master.HMaster.access$600(HMaster.java:199)
    at org.apache.hadoop.hbase.master.HMaster$2.run(HMaster.java:1871)
    at java.lang.Thread.run(Thread.java:745)
```

回答

当通过OfflineMetaRepair工具重建元数据时，HMaster在启动期间等待所有region server的WAL分割，以避免数据不一致问题。一旦WAL分割完成，HMaster将进行用户region的分配。所以当在集群异常的场景下，WAL分割可能需要很长时间，这取决于多个因素，例如太多的WALs，较慢的I/O，region servers不稳定等。

为确保HMaster能够成功完成所有region server WAL分割，请执行以下步骤：

1. 确保集群稳定，不存在其他问题。如有任何问题，请先修复。
2. 为“hbase.master.initializationmonitor.timeout”参数配置一个较大的值，默认值为“3600000”毫秒。
3. 重启HBase服务。

12.8.20.17 为什么 splitWAL 期间 HMaster 日志中频繁打印出 FileNotFoundException 及 no lease 信息

问题

当集群重启后会进行split WAL操作，在splitWAL期间，HMaster出现不能close log，日志中频繁打印出FileNotFoundException及no lease信息。

```
2017-06-10 09:50:27,586 | ERROR | split-log-closeStream-2 | Couldn't close log at hdfs://hacluster/hbase/
data/default/largeT1/2b48346d087275fe751fc049334fda93/recovered.edits/0000000000000000000.temp |
org.apache.hadoop.hbase.wal.WALSplitter$LogRecoveredEditsOutputSink$2.call(WALSplitter.java:1330)
java.io.FileNotFoundException: No lease on /hbase/data/default/
largeT1/2b48346d087275fe751fc049334fda93/recovered.edits/0000000000000000000.temp (inode
1092653): File does not exist. [Lease.  Holder: DFSClient_NONMAPREDUCE_1202985678_1, pendingcreates:
1936]
?at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkLease(FSNamesystem.java:3432)
?at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.analyzeFileState(FSNamesystem.java:3223)
?at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getNewBlockTargets(FSNamesystem.java:3057)
?at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getAdditionalBlock(FSNamesystem.java:3011)
?at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.addBlock(NameNodeRpcServer.java:842)
?at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.addBlock(ClientNameno
deProtocolServerSideTranslatorPB.java:526)
?at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol
$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
?at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:616)
?at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:973)
?at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2260)
?at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2256)
?at java.security.AccessController.doPrivileged(Native Method)
?at javax.security.auth.Subject.doAs(Subject.java:422)
?at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1769)
?at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2254)

?at sun.reflect.GeneratedConstructorAccessor40.newInstance(Unknown Source)
?at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
?at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
?at org.apache.hadoop.ipc.RemoteException.instantiateException(RemoteException.java:106)
?at org.apache.hadoop.ipc.RemoteException.unwrapRemoteException(RemoteException.java:73)
?at org.apache.hadoop.hdfs.DataStreamer.locateFollowingBlock(DataStreamer.java:1842)
?at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1639)
?at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:665)
```

回答

在splitWAL的过程中，参数“hbase.splitlog.manager.timeout”控制splitWAL的超时时间，若该时间内splitWAL无法完成，则会再次提交相同的任务，在一定时间内多次提交了相同的任务，当其中某次任务执行完毕时会删除这个temp文件，所以在后来的任务执行时无法找到这个文件，故出现FileNotFoudException。需做如下调整：

当前“hbase.splitlog.manager.timeout”的默认时间为“600000ms”，集群规格为每个regionserver上有2000~3000个region，在集群正常情况下(HBase无异常，HDFS无大量的读写操作等)，建议此参数依据集群的规格进行调整，若实际规格（实际平均每个regonserver上region的个数）大于默认规格（默认平均每个regionserver上region的个数，即2000），则调整方案为（实际规格 / 默认规格）* 默认时间。

在服务端的“hbase-site.xml”文件中配置splitlog参数，如表12-188所示。

表 12-188 splitlog 参数说明

参数	描述	默认值
hbase.splitlog.manager.timeout	分布式日志分裂管理程序接收worker回应的超时时间	600000

12.8.20.18 当使用与 Region Server 相同的 Linux 用户但不同的 kerberos 用户时，为什么 ImportTsv 工具执行失败报“Permission denied”的异常

问题

当使用与Region Server相同的Linux用户（例如omm用户）但不同的kerberos用户（例如admin用户）时，为什么ImportTsv工具执行失败报“Permission denied”的异常?

```
Exception in thread "main" org.apache.hadoop.security.AccessControlException: Permission denied:
user=admin, access=WRITE, inode="/user/omm-bulkload/hbase-staging/
partitions_cab16de5-87c2-4153-9cca-a6f4ed4278a6":hbase:hadoop:drwx--x--x
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:342)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:315)
        at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:23
1)
        at
com.xxx.hadoop.adapter.hdfs.plugin.HWAccessControlEnforce.checkPermission(HWAccessControlEnforce.java:
69)
        at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:19
0)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1789)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1773)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkAncestorAccess(FSDirectory.java:1756)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.startFileInternal(FSNamesystem.java:2490)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.startFileInt(FSNamesystem.java:2425)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.startFile(FSNamesystem.java:2308)
        at
org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.create(NameNodeRpcServer.java:745)
        at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.create(ClientNamenodeP
rotocolServerSideTranslatorPB.java:434)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol
$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server
$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:616)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:973)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2260)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2256)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1781)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2254)
```

回答

ImportTsv工具在“*客户端安装路径*/HBase/hbase/conf/hbase-site.xml”文件中“hbase.fs.tmp.dir”参数所配置的HBase临时目录中创建partition文件。因此客户端

（kerberos用户）应该在指定的临时目录上具有rwx的权限来执行ImportTsv操作。“hbase.fs.tmp.dir”参数的默认值为“/user/${user.name}/hbase-staging”（例如“/user/omm/hbase-staging”），此处“$ {user.name}”是操作系统用户名（即omm用户），客户端（kerberos用户，例如admin用户）不具备该目录的rwx权限。

上述问题可通过执行以下步骤解决：

1. 在客户端将“hbase.fs.tmp.dir”参数设置为当前kerberos用户的目录（如“/user/admin/hbase-staging”），或者为客户端（kerberos用户）提供已配置的目录所必需的rwx权限。
2. 重试ImportTsv操作。

12.8.20.19 租户访问 Phoenix 提示权限不足

问题

使用租户访问Phoenix提示权限不足。

回答

创建租户的时候需要关联HBase服务和Yarn队列。

租户要操作Phoenix还需要额外操作的权限，即Phoenix系统表的RWX权限。

例如：

创建好的租户为**hbase**，使用**admin**用户登录hbase shell，执行**scan 'hbase:acl'**命令查询租户对应的角色为**hbase_1450761169920**（格式为：租户名_时间戳）。

执行以下命令进行授权（如果还没有生成Phoenix系统表，请用**admin**用户登录Phoenix客户端后再回到hbase shell里授权）：

grant '@hbase_1450761169920','RWX','*SYSTEM.CATALOG*'

grant '@hbase_1450761169920','RWX','*SYSTEM.FUNCTION*'

grant '@hbase_1450761169920','RWX','*SYSTEM.SEQUENCE*'

grant '@hbase_1450761169920','RWX','*SYSTEM.STATS*'

新建用户**phoenix**并绑定租户**hbase**，该用户phoenix就可以用来访问Phoenix客户端。

12.8.20.20 如何解决 HBase 恢复数据任务失败后错误详情中提示：Rollback recovery failed 的回滚失败问题

问题

HBase恢复任务执行失败后系统自动回滚数据，若页面详情中提示“Rollback recovery failed”信息，表示回滚失败。由于回滚失败后就不会处理数据，所以有可能产生垃圾数据，需要如何解决?

回答

在下次执行备份或恢复任务前，需要手动清除这些垃圾数据。

步骤1 安装集群客户端，例如安装目录为“/opt/client”。

步骤2 使用客户端安装用户，执行**source /opt/client/bigdata_env**命令配置环境变量。

步骤3 执行**kinit admin**认证管理员身份。

步骤4 执行**zkCli.sh -server *ZooKeeper节点业务IP地址*:2181**连接ZooKeeper。

步骤5 执行**deleteall /recovering**删除垃圾数据。然后执行**quit**退出ZooKeeper连接。

说明

执行该命令会导致数据丢失，请谨慎操作。

步骤6 执行**hdfs dfs -rm -f -r /user/hbase/backup**删除临时数据。

步骤7 登录FusionInsight Manager界面，选择“运维 > 备份恢复 > 恢复管理”，在任务列表中对应任务的“操作”列，单击“查询历史”，在弹出的窗口中，在指定一次执行记录前单击，即可查看相关的快照名称信息：

```
Snapshot [ snapshot name ] is created successfully before recovery.
```

步骤8 切换到客户端，执行**hbase shell**，然后运行**delete_all_snapshot** '*snapshot name*.*'删除临时快照。

----结束

12.8.20.21 如何修复 Region Overlap

问题

MRS 3.x及之后版本，使用HBck工具检查Region状态，若日志中存在“ERROR: (regions region1 and region2) There is an overlap in the region chain.”或者“ERROR: (region region1) Multiple regions have the same startkey: xxx”信息，表示某些Region存在Overlap的问题，需要如何解决?

回答

修复步骤如下：

步骤1 执行**hbase hbck -repair** *tableName*命令修复存在overlap的表。

步骤2 执行**hbase hbck** *tableName*命令检查修复的表是否还存在overlap。

- 如果不存在overlap，执行**步骤3**。
- 如果存在overlap，执行**步骤1**。

步骤3 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HBase > 更多 > 执行HMaster倒换”，完成HMaster主备倒换。

步骤4 执行**hbase hbck** *tableName*命令检查修复的表是否还存在overlap。

- 如果不存在overlap，修复完成。
- 如果存在overlap，从**步骤1**开始重新执行修复步骤。

----结束

12.8.20.22 HBase RegionServer GC 参数 Xms，Xmx 配置 31G，导致 RegionServer 启动失败

问题

MRS 3.x及之后版本，查看RegionServer启动失败节点的hbase-omm-*.out日志，发现日志中存在“An error report file with more information is saved as: /tmp/hs_err_pid*.log”，查看/tmp/hs_err_pid*.log发现日志存在“#Internal Error (vtableStubs_aarch64.cpp:213), pid=9456, tid=0x0000ffff97fdd200”和“#guarantee(__ pc() <= s->code_end()) failed: overflowed buffer”，表示此问题是由JDK导致，需要如何解决?

回答

修复步骤如下：

步骤1 在RegionServer启动失败的某个节点执行 **su - omm**，切换到**omm**用户。

步骤2 在omm用户下执行**java -XX:+PrintFlagsFinal -version |grep HeapBase**，出现如下类似结果。

```
uintx HeapBaseMinAddress = 2147483648 {pd product}
```

步骤3 修改“GC_OPTS”中“-Xms”和“-Xmx”的值使其不在32G-HeapBaseMinAddress和32G的值之间，不包括32G和32G-HeapBaseMinAddress的值。

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HBase > 实例”，选择失败实例，选择“更多 > 重启实例”来重启失败实例。

----结束

12.8.20.23 使用集群内节点执行批量导入，为什么 LoadIncrementalHFiles 工具执行失败报“Permission denied”的异常

问题

在普通集群中手动创建Linux用户，并使用集群内DataNode节点执行批量导入时，为什么LoadIncrementalHFiles工具执行失败报“Permission denied”的异常?

```
2020-09-20 14:53:53,808 WARN  [main] shortcircuit.DomainSocketFactory: error creating DomainSocket
java.net.ConnectException: connect(2) error: Permission denied when trying to connect to '/var/run/FusionInsight-HDFS/dn_socket'
    at org.apache.hadoop.net.unix.DomainSocket.connect0(Native Method)
    at org.apache.hadoop.net.unix.DomainSocket.connect(DomainSocket.java:256)
    at org.apache.hadoop.hdfs.shortcircuit.DomainSocketFactory.createSocket(DomainSocketFactory.java:168)
    at org.apache.hadoop.hdfs.client.impl.BlockReaderFactory.nextDomainPeer(BlockReaderFactory.java:804)
    at org.apache.hadoop.hdfs.client.impl.BlockReaderFactory.createShortCircuitReplicaInfo(BlockReaderFactory.java:526)
    at org.apache.hadoop.hdfs.shortcircuit.ShortCircuitCache.create(ShortCircuitCache.java:785)
    at org.apache.hadoop.hdfs.shortcircuit.ShortCircuitCache.fetchOrCreate(ShortCircuitCache.java:722)
    at org.apache.hadoop.hdfs.client.impl.BlockReaderFactory.getBlockReaderLocal(BlockReaderFactory.java:483)
    at org.apache.hadoop.hdfs.client.impl.BlockReaderFactory.build(BlockReaderFactory.java:360)
    at org.apache.hadoop.hdfs.DFSInputStream.getBlockReader(DFSInputStream.java:663)
    at org.apache.hadoop.hdfs.DFSInputStream.blockSeekTo(DFSInputStream.java:594)
    at org.apache.hadoop.hdfs.DFSInputStream.readWithStrategy(DFSInputStream.java:776)
    at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:845)
    at java.io.DataInputStream.readFully(DataInputStream.java:195)
    at org.apache.hadoop.hbase.io.hfile.FixedFileTrailer.readFromStream(FixedFileTrailer.java:401)
```

```
    at org.apache.hadoop.hbase.io.hfile.HFile.isHFileFormat(HFile.java:651)
    at org.apache.hadoop.hbase.io.hfile.HFile.isHFileFormat(HFile.java:634)
    at org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.visitBulkHFiles(LoadIncrementalHFiles.java:1090)
    at
org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.discoverLoadQueue(LoadIncrementalHFiles.java:1006)
    at
org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.prepareHFileQueue(LoadIncrementalHFiles.java:257)
    at org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.doBulkLoad(LoadIncrementalHFiles.java:364)
    at org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.run(LoadIncrementalHFiles.java:1263)
    at org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.run(LoadIncrementalHFiles.java:1276)
    at org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.run(LoadIncrementalHFiles.java:1311)
    at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:76)
    at org.apache.hadoop.hbase.tool.LoadIncrementalHFiles.main(LoadIncrementalHFiles.java:1333)
```

回答

如果LoadIncrementalHFiles工具依赖的Client在集群内安装，且和DataNode在相同的节点上，在工具执行过程中HDFS会创建短路读提高性能。短路读依赖“/var/run/FusionInsight-HDFS”目录(“dfs.domain.socket.path”)，该目录默认权限是750。而当前Linux用户没有权限操作该目录。

上述问题可通过执行以下方法解决：

方法一：创建新用户(推荐使用)。

步骤1 通过Manager页面创建新的用户，该用户属组中默认包含ficommon组。

```
[root@xxx-xxx-xxx-xxx ~]# id test
uid=20038(test) gid=9998(ficommon) groups=9998(ficommon)
```

步骤2 重新执行ImportData。

----结束

方法二：修改当前用户的属组。

步骤1 将该用户添加到ficommon组中。

```
[root@xxx-xxx-xxx-xxx ~]# usermod -a -G ficommon test
[root@xxx-xxx-xxx-xxx ~]# id test
uid=2102(test) gid=2102(test) groups=2102(test),9998(ficommon)
```

步骤2 重新执行ImportData。

----结束

12.8.20.24 Phoenix sqlline 脚本使用，报 import argparse 错误

问题

在客户端使用sqlline脚本时，报import argparse错误。

回答

步骤1 以**root**用户登录安装HBase客户端的节点，使用**hbase**用户进行安全认证。

步骤2 进入HBase客户端sqlline脚本所在目录执行**python3 sqlline.py**命令。

----结束

12.8.20.25 Phoenix BulkLoad Tool 限制

问题

当更新索引字段数据时，若用户表已经存在一批数据，则BulkLoad工具不能更新全局和局部可变索引。

回答

问题分析

1. 创建表。

```
CREATE TABLE TEST_TABLE(
DATE varchar not null,
NUM integer not null,
SEQ_NUM integer not null,
ACCOUNT1 varchar not null,
ACCOUNTDES varchar,
FLAG varchar,
SALL double,
CONSTRAINT PK PRIMARY KEY (DATE,NUM,SEQ_NUM,ACCOUNT1)
);
```

2. 创建全局索引

   CREATE INDEX TEST_TABLE_INDEX ON TEST_TABLE(ACCOUNT1,DATE,NUM,ACCOUNTDES,SEQ_NUM);

3. 插入数据

   UPSERT INTO TEST_TABLE (DATE,NUM,SEQ_NUM,ACCOUNT1,ACCOUNTDES,FLAG,SALL) values ('20201001',30201001,13,'367392332','sffa1','','');

4. 执行BulkLoad任务更新数据

   hbase org.apache.phoenix.mapreduce.CsvBulkLoadTool -t TEST_TABLE -i /tmp/test.csv，test.csv内容如下：

20201001	30201001	13	367392332	sffa888	1231243	23

5. 问题现象：无法直接更新之前存在的索引数据，导致存在两条索引数据。

```
+------------+-----------+-----------+---------------+----------------+
| :ACCOUNT1  |  :DATE    |  :NUM     | 0:ACCOUNTDES  | :SEQ_NUM  |
+------------+-----------+-----------+---------------+----------------+
| 367392332  | 20201001  | 30201001  | sffa1         | 13             |
| 367392332  | 20201001  | 30201001  | sffa888       | 13             |
+------------+-----------+-----------+---------------+----------------+
```

解决方法

步骤1 删除旧的索引表。

DROP INDEX TEST_TABLE_INDEX ON TEST_TABLE;

步骤2 异步方式创建新的索引表。

CREATE INDEX TEST_TABLE_INDEX ON TEST_TABLE(ACCOUNT1,DATE,NUM,ACCOUNTDES,SEQ_NUM) ASYNC;

步骤3 索引重建。

hbase org.apache.phoenix.mapreduce.index.IndexTool --data-table TEST_TABLE --index-table TEST_TABLE_INDEX --output-path /user/test_table

----结束

12.8.20.26 CTBase 对接 Ranger 权限插件，提示权限不足

问题

CTBase访问启用Ranger插件的HBase服务时，如果创建聚簇表，提示权限不足。

```
ERROR: Create ClusterTable failed. Error: org.apache.hadoop.hbase.security.AccessDeniedException:
Insufficient permissions for user 'ctbase2@HADOOP.COM' (action=create)
at org.apache.ranger.authorization.hbase.AuthorizationSession.publishResults(AuthorizationSession.java:278)
at
org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor.authorizeAccess(RangerAuthorizatio
nCoprocessor.java:654)
at
org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor.requirePermission(RangerAuthorizati
onCoprocessor.java:772)
at
org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor.preCreateTable(RangerAuthorization
Coprocessor.java:943)
at
org.apache.ranger.authorization.hbase.RangerAuthorizationCoprocessor.preCreateTable(RangerAuthorization
Coprocessor.java:428)
at org.apache.hadoop.hbase.master.MasterCoprocessorHost$12.call(MasterCoprocessorHost.java:351)
at org.apache.hadoop.hbase.master.MasterCoprocessorHost$12.call(MasterCoprocessorHost.java:348)
at org.apache.hadoop.hbase.coprocessor.CoprocessorHost
$ObserverOperationWithoutResult.callObserver(CoprocessorHost.java:581)
at org.apache.hadoop.hbase.coprocessor.CoprocessorHost.execOperation(CoprocessorHost.java:655)
at
org.apache.hadoop.hbase.master.MasterCoprocessorHost.preCreateTable(MasterCoprocessorHost.java:348)
at org.apache.hadoop.hbase.master.HMaster$5.run(HMaster.java:2192)
at
org.apache.hadoop.hbase.master.procedure.MasterProcedureUtil.submitProcedure(MasterProcedureUtil.java:1
34)
at org.apache.hadoop.hbase.master.HMaster.createTable(HMaster.java:2189)
at org.apache.hadoop.hbase.master.MasterRpcServices.createTable(MasterRpcServices.java:711)
at org.apache.hadoop.hbase.shaded.protobuf.generated.MasterProtos$MasterService
$2.callBlockingMethod(MasterProtos.java)
at org.apache.hadoop.hbase.ipc.RpcServer.call(RpcServer.java:458)
at org.apache.hadoop.hbase.ipc.CallRunner.run(CallRunner.java:133)
at org.apache.hadoop.hbase.ipc.RpcExecutor$Handler.run(RpcExecutor.java:338)
at org.apache.hadoop.hbase.ipc.RpcExecutor$Handler.run(RpcExecutor.java:318)
```

回答

CTBase用户在Ranger界面配置权限策略，赋予CTBase元数据表_ctmeta_、聚簇表和索引表RWCAE（READ，WRITE，EXEC，CREATE，ADMIN）权限。

12.9 使用 HDFS

12.9.1 从零开始使用 Hadoop

本章节提供从零开始使用Hadoop提交wordcount作业的操作指导，wordcount是最经典的Hadoop作业，它用来统计海量文本的单词数量。

操作步骤

步骤1 准备wordcount程序。

开源的Hadoop的样例程序包含多个例子，其中包含wordcount。可以从**https://dist.apache.org/repos/dist/release/hadoop/common/**中下载Hadoop的样例程序。

例如，选择hadoop-2.10.x版本，下载“hadoop-2.10.x.tar.gz”，解压后在“hadoop-2.10.x\share\hadoop\mapreduce”路径下获取“hadoop-mapreduce-examples-2.10.x.jar”，即为Hadoop的样例程序。“hadoop-mapreduce-examples-2.10.x.jar”样例程序包含了wordcount程序。

说明

hadoop-2.10.x表示Hadoop的版本号。

步骤2 准备数据文件。

数据文件无格式要求，准备一个或多个txt文件即可，如下内容为txt文件样例：

```
qwsdfhoedfrffrofhuncckgktpmhutopmma
jjpsffjfjorgjgtyiuyjmhombmbogohoyhm
jhheyeombdhuaqqiquyebchdhmamdhdemmj
doeyhjwedcrfvtgbmojiyhhqssddddddfkf
kjhhjkehdeiyrudjhfhfhffooqweopuyyyy
```

步骤3 上传数据至OBS。

1. 登录OBS控制台。
2. 单击“并行文件系统 > 创建并行文件系统”，创建一个名称为wordcount01的文件系统。

   wordcount01仅为示例，文件系统名称必须全局唯一，否则会创建并行文件系统失败。
3. 在OBS文件系统列表中单击文件系统名称wordcount01，选择“文件 > 新建文件夹”，分别创建program、input文件夹。
   - program：存放用户程序
   - input：存放用户数据文件
4. 进入program文件夹，选择“上传文件 > 添加文件”，从本地选择**步骤1**中下载的程序包，然后单击“上传”。
5. 进入input文件夹，将**步骤2**中准备的数据文件上传到input文件夹。

步骤4 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称，该集群需要包含Hadoop组件。

步骤5 提交wordcount作业。

在MRS控制台选择“作业管理”页签，单击“添加”，进入“添加作业”页面。

- 作业类型选择“MapReduce”。
- 作业名称为“mr_01”。
- 执行程序路径配置为OBS上存放程序的地址。例如：obs://wordcount01/program/hadoop-mapreduce-examples-2.10.x.jar。
- 执行程序参数中填写的参数为：wordcount obs://wordcount01/input/ obs://wordcount01/output/。

说明

- 参数“**obs://wordcount01/input/**”中的OBS文件系统名需要替换为实际环境创建的文件系统名。
- 参数“**obs://wordcount01/output/**”中的OBS文件系统名需要替换为实际环境创建的文件系统名，目录output请手动输入一个不存在的目录。

- 服务配置参数无需填写。

只有集群处于“运行中”状态时才能提交作业。

作业提交成功后默认为“已接受”状态，不需要用户手动执行作业。

步骤6 查看作业执行结果。

1. 进入“作业管理”页面，查看作业是否执行完成。

   作业运行需要时间，作业运行结束后，刷新作业列表。

   作业执行成功或失败后都不能再次执行，只能新增或者复制作业，配置作业参数后重新提交作业。

2. 登录OBS控制台，进入OBS路径，查看作业输出信息。

   进入到**步骤5**中创建的output路径查看相关的output文件，需要下载到本地以文本方式打开进行查看。

----结束

12.9.2 配置内存管理

配置场景

在HDFS中，每个文件对象都需要在NameNode中注册相应的信息，并占用一定的存储空间。随着文件数的增加，当原有的内存空间无法存储相应的信息时，需要修改内存大小的设置。

配置描述

参数入口：

请参考**修改集群服务配置参数**，进入HDFS“全部配置”页面。

表 12-189 参数说明

配置参数	说明	默认值
GC_PROFILE	NameNode所占内存主要由FsImage大小决定。FsImage Size = 文件数 * 900 Bytes，根据计算结果可估算hdfs的NameNode应设内存大小。 该参数项的内存大小取值如下： ● high：4G ● medium：2G ● low：256M ● custom：根据实际数据量大小在GC_OPTS中设置内存大小。	custom
GC_OPTS	JVM用于gc的参数。仅当GC_PROFILE设置为custom时该配置才会生效。需确保GC_OPT参数设置正确，否则进程启动会失败。 **须知** 请谨慎修改该项。如果配置不当，将造成服务不可用。	-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=128M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -Djdk.tls.ephemeralDHKeySize=2048

12.9.3 创建 HDFS 角色

操作场景

该任务指导系统管理员在FusionInsight Manager创建并设置HDFS的角色。HDFS角色可设置HDFS目录或文件的读、写和执行权限。

用户在HDFS中对自己创建的目录或文件拥有完整权限，可直接读取、写入以及授权他人访问此HDFS目录与文件。

说明

- 本章节适用于MRS 3.x及后续版本。
- 安全模式支持创建HDFS角色，普通模式不支持创建HDFS角色。
- 如果当前组件使用了Ranger进行权限控制，须基于Ranger配置HDFS相关策略进行权限管理，具体操作可参考**添加HDFS的Ranger访问权限策略**。

前提条件

系统管理员已明确业务需求。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”中输入角色名字与描述。

步骤3 配置资源权限，请参见**表12-190**。

“文件系统”：HDFS中的目录和文件授权。

HDFS常见目录如下：

- “flume”：Flume数据存储目录。
- “hbase”：HBase数据存储目录。
- “mr-history”：MapReduce任务信息存储目录。
- “tmp”：临时数据存储目录。
- “user”：用户数据存储目录。

表 12-190 设置角色

任务场景	角色授权操作
设置HDFS管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS”，勾选“集群管理操作权限”。 **说明** 设置HDFS管理员权限需要重启HDFS服务才可生效。
设置用户执行HDFS检查和HDFS修复的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS > 文件系统”。 2. 定位到指定目录或文件在HDFS中保存的位置。 3. 在指定目录或文件的“权限”列，勾选“读”和“执行”。
设置用户读取其他用户的目录或文件的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS > 文件系统”。 2. 定位到指定目录或文件在HDFS中保存的位置。 3. 在指定目录或文件的“权限”列，勾选“读”和“执行”。

任务场景	角色授权操作
设置用户在其他用户的文件写入数据的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS > 文件系统”。 2. 定位到指定文件在HDFS中保存的位置。 3. 在指定文件的“权限”列，勾选“写”和“执行”。
设置用户在其他用户的目录新建或删除子文件、子目录的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS > 文件系统”。 2. 定位到指定目录在HDFS中保存的位置。 3. 在指定目录的“权限”列，勾选“写”和“执行”。
设置用户在其他用户的目录或文件执行的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS > 文件系统”。 2. 定位到指定目录或文件在HDFS中保存的位置。 3. 在指定目录或文件的“权限”列，勾选“执行”。
设置子目录继承上级目录权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HDFS > 文件系统”。 2. 定位到指定目录或文件在HDFS中保存的位置。 3. 在指定目录或文件的“权限”列，勾选“递归”。

步骤4 单击“确定”完成，返回“角色”。

----结束

12.9.4 使用 HDFS 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用HDFS客户端。

前提条件

- 已安装客户端。

  例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）

使用 HDFS 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行HDFS Shell命令。例如：

hdfs dfs -ls /

----结束

HDFS 客户端常用命令

常用的HDFS客户端命令如下表所示。

更多命令可参考**https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/CommandsManual.html#User_Commands**

表 12-191 HDFS 客户端常用命令

命令	说明	样例
hdfs dfs -mkdir *文件夹名称*	创建文件夹	**hdfs dfs -mkdir /tmp/mydir**
hdfs dfs -ls *文件夹名称*	查看文件夹	**hdfs dfs -ls /tmp**
hdfs dfs -put *客户端节点上本地文件 HDFS指定路径*	上传本地文件到HDFS指定路径	**hdfs dfs -put /opt/test.txt /tmp** 上传客户端节点“/opt/test.txt”文件到HDFS的“/tmp”路径下
hdfs dfs -get *hdfs指定文件 客户端节点上指定路径*	下载HDFS文件到本地指定路径	**hdfs dfs -get /tmp/test.txt /opt/** 下载HDFS的“/tmp/test.txt”文件到客户端节点的“/opt”路径下
hdfs dfs -rm -r -f *hdfs指定文件夹*	删除文件夹	**hdfs dfs -rm -r -f /tmp/mydir**
hdfs dfs -chmod *权限参数 文件目录*	为用户设置HDFS目录权限	**hdfs dfs -chmod 700 /tmp/test**

客户端常见使用问题

1. 当执行HDFS客户端命令时，客户端程序异常退出，报“java.lang.OutOfMemoryError”的错误。

这个问题是由于HDFS客户端运行时的所需的内存超过了HDFS客户端设置的内存上限（默认为128MB）。可以通过修改“*<客户端安装路径>*/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改HDFS客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

```
CLIENT_GC_OPTS="-Xmx1G"
```

在修改完后，使用如下命令刷新客户端配置，使之生效：

source <*客户端安装路径*>/**bigdata_env**

2. 如何设置HDFS客户端运行时的日志级别?

   HDFS客户端运行时的日志是默认输出到Console控制台的，其级别默认是INFO级别。有的时候为了定位问题，需要开启DEBUG级别日志，可以通过导出一个环境变量来设置，命令如下：

   export HADOOP_ROOT_LOGGER=DEBUG,console

   在执行完上面命令后，再执行HDFS Shell命令时，即可打印出DEBUG级别日志。

   如果想恢复INFO级别日志，可执行如下命令：

   export HADOOP_ROOT_LOGGER=INFO,console

3. 如何彻底删除HDFS文件?

   由于HDFS的回收站机制，一般删除HDFS文件后，文件会移动到HDFS的回收站中。如果确认文件不再需要并且需要立马释放存储空间，可以继续清理对应的回收站目录（例如：hdfs://hacluster/user/*xxx*/.Trash/Current/*xxx*）。

12.9.5 使用 distcp 命令

操作场景

distcp是一种在集群间或集群内部拷贝大量数据的工具。它利用MapReduce任务实现大量数据的分布式拷贝。

前提条件

- 已安装Yarn客户端或者包括Yarn的客户端。例如安装目录为“/opt/client”。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）
- 如需在集群间拷贝数据，拷贝数据的集群双方都需要启用集群间拷贝数据功能。

操作步骤

步骤1 登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行distcp命令的用户所属的用户组必须为**supergroup**组，且执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行distcp命令。例如：

hadoop distcp hdfs://hacluster/source hdfs://hacluster/target

----结束

distcp 常见用法

1. 最常见的distcp用法，示例如下：
   ```
   hadoop distcp -numListstatusThreads 40 -update -delete -prbugpaxtq hdfs://cluster1/source hdfs://cluster2/target
   ```

   说明

   在上述命令中：
   - -numListstatusThreads指定了40个构建被拷贝文件的列表的线程数；
   - -update -delete表示将源位置和目标位置的文件同步，删除掉目标位置多余的文件，注意如果需要增量拷贝文件，请将-delete删掉；
   - -prbugpaxtq与-update配合，表示被拷贝文件的状态信息也会被更新；
   - hdfs://cluster1/source、hdfs://cluster2/target分别表示源位置和目标位置。

2. 集群间的数据拷贝，示例如下：
   ```
   hadoop distcp hdfs://cluster1/foo/bar hdfs://cluster2/bar/foo
   ```

   说明

   集群cluster1和集群cluster2之间的网络必须保持互通，且两个集群需要使用相同或兼容的HDFS版本。

3. 多个源目录的数据拷贝，示例如下：
   ```
   hadoop distcp hdfs://cluster1/foo/a \
   hdfs://cluster1/foo/b \
   hdfs://cluster2/bar/foo
   ```

   上面的命令的效果是将集群cluster1的文件夹a、b拷贝到集群cluster2的“/bar/foo”目录下，它的效果等效于下面的命令：

   ```
   hadoop distcp -f hdfs://cluster1/srclist \
   hdfs://cluster2/bar/foo
   ```

   其中srclist里面的内容如下。注意运行distcp命令前，需要将srclist文件上传到HDFS上。

   ```
   hdfs://cluster1/foo/a
   hdfs://cluster1/foo/b
   ```

4. update和overwrite选项的用法，-update用于被拷贝的文件在目标位置中不存在，或者更新目标位置中被拷贝文件的内容；-overwrite用于覆盖在目标位置中已经存在的文件。

   不加选项和加两个选项中任一个选项的区别，示例如下：

   假设，源位置的文件结构如下：

   ```
   hdfs://cluster1/source/first/1
   hdfs://cluster1/source/first/2
   hdfs://cluster1/source/second/10
   hdfs://cluster1/source/second/20
   ```

   不加选项的命令：

   ```
   hadoop distcp hdfs://cluster1/source/first  hdfs://cluster1/source/second  hdfs://cluster2/target
   ```

   上述命令默认会在目标位置创建文件夹first、second，所以拷贝结果如下：

   ```
   hdfs://cluster2/target/first/1
   hdfs://cluster2/target/first/2
   hdfs://cluster2/target/second/10
   hdfs://cluster2/target/second/20
   ```

加两个选项中任一个选项的命令，例如加update选项：

```
hadoop distcp -update hdfs://cluster1/source/first  hdfs://cluster1/source/second  hdfs://cluster2/target
```

上述命令只会将源位置的内容拷贝到目标位置，所以拷贝结果如下：

```
hdfs://cluster2/target/1
hdfs://cluster2/target/2
hdfs://cluster2/target/10
hdfs://cluster2/target/20
```

说明

- 如果多个源位置有相同名称的文件，则distcp命令会失败。
- 在不使用update和overwrite选项的情况下，如果被拷贝文件在目标位置中已经存在，则该文件会跳过。
- 在使用update选项的情况下，如果被拷贝文件在目标位置中已经存在，但文件内容不同，则目标位置的文件内容会被更新。
- 在使用overwrite选项的情况下，如果被拷贝文件在目标位置中已经存在，目标位置的文件依然会被覆盖。

5. 其它命令选项：

表 12-192 其他命令选项

选项	描述
-p[rbugpcaxtq]	当同时使用-update选项时，即使被拷贝文件的内容没有被更新，它的状态信息也会被更新 r：副本数，b：块大小，u：所属用户，g：所属用户组，p：许可，c：校验和类型，a：访问控制，t：时间戳，q：Quota信息
-i	拷贝过程中忽略失败
-log <logdir>	指定日志路径
-v	指定日志中的额外信息
-m <num_maps>	最大的同时运行的执行拷贝的任务数
-numListstatusThreads	构建被拷贝文件的文件列表时所用的线程数，该选项会提高distcp的运行速度
-overwrite	覆盖目标位置的文件
-update	如果源位置和目标位置的文件的大小，校验和不同，则更新目标位置的文件
-append	当同时使用-update选项时，追加源位置的文件内容到目标位置的文件
-f <urilist_uri>	将<urilist_uri>文件的内容作为需要拷贝的文件列表
-filters	指定一个本地文件，其文件内容是多条正则表达式。当被拷贝的文件与某条正则表达式匹配时，则该文件不会被拷贝
-async	异步运行distcp命令

选项	描述
-atomic {-tmp <tmp_dir>}	指定一次原子性的拷贝，可以添加一个临时目录的选项，作为拷贝过程中的暂存目录
-bandwidth	指定每个拷贝任务的传输带宽，单位MB/s
-delete	删除掉目标位置中存在，但源位置不存在的文件。该选项通常会和-update配合使用，表示将源位置和目标位置的文件同步，删除掉目标位置多余的文件
-diff <oldSnapshot> <newSnapshot>	将新旧版本之间的差异内容，拷贝到目标位置的旧版本文件中
-skipcrccheck	是否跳过源文件和目标文件之间的CRC校验
-strategy {dynamic| uniformsize}	指定拷贝任务的拷贝策略，默认策略是uniformsize，即每个拷贝任务复制相同的字节数

distcp 常见使用问题

1. 当使用distcp命令时，如果某些被拷贝的文件内容较大时，建议修改执行拷贝任务的mapreduce的超时时间。可以通过在distcp命令中指定**mapreduce.task.timeout**选项实现。例如，修改超时时间为30分钟，则命令如下：

   ```
   hadoop distcp -Dmapreduce.task.timeout=1800000 hdfs://cluster1/source hdfs://cluster2/target
   ```

   您也可以使用选项filters，不对这种大文件进行拷贝，命令示例如下：

   ```
   hadoop distcp -filters /opt/client/filterfile hdfs://cluster1/source hdfs://cluster2/target
   ```

   其中filterfile是本地文件，它的内容是多条用于匹配不拷贝文件路径的正则表达式，它的内容示例如下：

   ```
   .*excludeFile1.*
   .*excludeFile2.*
   ```

2. 当使用distcp命令时，命令异常退出，报“java.lang.OutOfMemoryError”的错误。

   这个问题的原因是拷贝任务运行时所需的内存超过了客户端设置的内存上限（默认为128MB）。可以通过修改“*<客户端安装路径>*/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   CLIENT_GC_OPTS="-Xmx1G"
   ```

   在修改完后，使用如下命令刷新客户端配置，使之生效：

   source *<客户端安装路径>*/**bigdata_env**

3. 使用dynamic策略执行distcp命令时，命令异常退出，报“Too many chunks created with splitRatio”的错误。

   这个问题的原因是“distcp.dynamic.max.chunks.tolerable”的值（默认值为20000）小于“distcp.dynamic.split.ratio”的值（默认为2）乘以Map数。即一般出现在Map数超过10000的情况。可以通过-m参数降低Map数小于10000：

   ```
   hadoop distcp -strategy dynamic -m 9500 hdfs://cluster1/source hdfs://cluster2/target
   ```

   或通过-D参数指定更大的“distcp.dynamic.max.chunks.tolerable”的值：

   ```
   hadoop distcp -Ddistcp.dynamic.max.chunks.tolerable=30000 -strategy dynamic hdfs://cluster1/source hdfs://cluster2/target
   ```

12.9.6 HDFS 文件系统目录简介

HDFS文件系统中目录结构如下表所示。

表 12-193 HDFS 文件系统目录结构（适用于 MRS 3.x 之前版本）

路径	类型	简略功能	是否可以删除	删除的后果
/tmp/spark/sparkhive-scratch	固定目录	存放Spark JDBCServer中metastore session临时文件	否	任务运行失败
/tmp/sparkhive-scratch	固定目录	存放Spark cli方式运行metastore session临时文件	否	任务运行失败
/tmp/carbon/	固定目录	数据导入过程中，如果存在异常CarbonData数据，则将异常数据放在此目录下	是	错误数据丢失
/tmp/Loader-${作业名}_${MR作业id}	临时目录	存放Loader Hbase bulkload作业的region信息，作业完成后自动删除	否	Loader Hbase Bulkload作业失败
/tmp/logs	固定目录	MR任务日志在HDFS上的聚合路径	是	MR任务日志丢失
/tmp/archived	固定目录	MR任务日志在HDFS上的归档路径	是	MR任务日志丢失
/tmp/hadoop-yarn/staging	固定目录	保存AM运行作业运行日志、作业概要信息和作业配置属性	否	任务运行异常
/tmp/hadoop-yarn/staging/history/done_intermediate	固定目录	所有任务运行完成后，临时存放/tmp/hadoop-yarn/staging目录下文件	否	MR任务日志丢失
/tmp/hadoop-yarn/staging/history/done	固定目录	周期性扫描线程定期将done_intermediate的日志文件转移到done目录	否	MR任务日志丢失

路径	类型	简略功能	是否可以删除	删除的后果
/tmp/mr-history	固定目录	存储预加载历史记录文件的路径	否	MR历史任务日志数据丢失
/tmp/hive	固定目录	存放Hive的临时文件	否	导致Hive任务失败
/tmp/hive-scratch	固定目录	Hive运行时生成的临时数据，如会话信息等	否	当前执行的任务会失败
/user/{user}/.sparkStaging	固定目录	存储SparkJDBCServer应用临时文件	否	executor启动失败
/user/spark/jars	固定目录	存放Spark executor运行依赖包	否	executor启动失败
/user/loader	固定目录	存放loader的作业脏数据以及HBase作业数据的临时存储目录	否	HBase作业失败或者脏数据丢失
/user/loader/etl_dirty_data_dir				
/user/loader/etl_hbase_putlist_tmp				
/user/loader/etl_hbase_tmp				
/user/mapred	固定目录	存放Hadoop相关的文件	否	导致Yarn启动失败
/user/hive	固定目录	Hive相关数据存储的默认路径，包含依赖的spark lib包和用户默认表数据存储位置等	否	用户数据丢失
/user/omm-bulkload	临时目录	HBase批量导入工具临时目录	否	HBase批量导入任务失败

路径	类型	简略功能	是否可以删除	删除的后果
/user/hbase	临时目录	HBase批量导入工具临时目录	否	HBase批量导入任务失败
/sparkJobHistory	固定目录	Spark eventlog数据存储目录	否	HistoryServer服务不可用，任务运行失败
/flume	固定目录	Flume采集到HDFS文件系统中的数据存储目录	否	Flume工作异常
/mr-history/tmp	固定目录	MapReduce作业产生的日志存放位置	是	日志信息丢失
/mr-history/done	固定目录	MR JobHistory Server管理的日志的存放位置	是	日志信息丢失
/tenant	添加租户时创建	配置租户在HDFS中的存储目录，系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“tenant/ta1”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。	否	租户不可用
/apps{1~5}/	固定目录	WebHCat使用到Hive的包的路径	否	执行WebHCat任务会失败
/hbase	固定目录	HBase数据存储目录	否	HBase用户数据丢失
/hbaseFileStream	固定目录	HFS文件存储目录	否	HFS文件丢失，且无法恢复

路径	类型	简略功能	是否可以删除	删除的后果
/ats/active	固定目录	HDFS路径，用于存储活动的应用程序的timeline数据	否	删除后会导致tez任务运行失败
/ats/done	固定目录	HDFS路径，用于存储完成的应用程序的timeline数据	否	删除后会自动创建
/flink	固定目录	存放checkpoint任务数据	否	删除会导致运行任务失败

表 12-194 HDFS 文件系统目录结构（适用于 MRS 3.x 及之后版本）

路径	类型	简略功能	是否可以删除	删除的后果
/tmp/spark2x/sparkhive-scratch	固定目录	存放Spark2x JDBCServer中metastore session临时文件	否	任务运行失败
/tmp/sparkhive-scratch	固定目录	存放Spark2x cli方式运行metastore session临时文件	否	任务运行失败
/tmp/logs/	固定目录	存放container日志文件	是	container日志不可查看
/tmp/carbon/	固定目录	数据导入过程中，如果存在异常CarbonData数据，则将异常数据放在此目录下	是	错误数据丢失
/tmp/Loader-${作业名}_${MR作业id}	临时目录	存放Loader Hbase bulkload作业的region信息，作业完成后自动删除	否	Loader Hbase Bulkload作业失败

路径	类型	简略功能	是否可以删除	删除的后果
/tmp/hadoop-omm/yarn/system/rmstore	固定目录	ResourceManager运行状态信息	是	ResourceManager重启后状态信息丢失
/tmp/archived	固定目录	MR任务日志在HDFS上的归档路径	是	MR任务日志丢失
/tmp/hadoop-yarn/staging	固定目录	保存AM运行作业运行日志、作业概要信息和作业配置属性	否	任务运行异常
/tmp/hadoop-yarn/staging/history/done_intermediate	固定目录	所有任务运行完成后，临时存放/tmp/hadoop-yarn/staging目录下文件	否	MR任务日志丢失
/tmp/hadoop-yarn/staging/history/done	固定目录	周期性扫描线程定期将done_intermediate的日志文件转移到done目录	否	MR任务日志丢失
/tmp/mr-history	固定目录	存储预加载历史记录文件的路径	否	MR历史任务日志数据丢失
/tmp/hive-scratch	固定目录	Hive运行时生成的临时数据，如会话信息等	否	当前执行的任务会失败
/user/{user}/.sparkStaging	固定目录	存储SparkJDBCServer应用临时文件	否	executor启动失败
/user/spark2x/jars	固定目录	存放Spark2x executor运行依赖包	否	executor启动失败
/user/loader	固定目录	存放loader的作业脏数据以及HBase作业数据的临时存储目录	否	HBase作业失败或者脏数据丢失
/user/loader/etl_dirty_data_dir				

路径	类型	简略功能	是否可以删除	删除的后果
/user/loader/etl_hbase_putlist_tmp				
/user/loader/etl_hbase_tmp				
/user/oozie	固定目录	存放oozie运行时需要的依赖库，需用户手动上传	否	oozie调度失败
/user/mapred/hadoop-mapreduce-*3.1.1*.tar.gz	固定文件	MR分布式缓存功能使用的各jar包	否	MR分布式缓存功能无法使用
/user/hive	固定目录	Hive相关数据存储的默认路径，包含依赖的spark lib包和用户默认表数据存储位置等	否	用户数据丢失
/user/omm-bulkload	临时目录	HBase批量导入工具临时目录	否	HBase批量导入任务失败
/user/hbase	临时目录	HBase批量导入工具临时目录	否	HBase批量导入任务失败
/spark2xJobHistory2x	固定目录	Spark2x eventlog数据存储目录	否	HistoryServer服务不可用，任务运行失败
/flume	固定目录	Flume采集到HDFS文件系统中的数据存储目录	否	Flume工作异常
/mr-history/tmp	固定目录	MapReduce作业产生的日志存放位置	是	日志信息丢失
/mr-history/done	固定目录	MR JobHistory Server管理的日志的存放位置	是	日志信息丢失

路径	类型	简略功能	是否可以删除	删除的后果
/tenant	添加租户时创建	配置租户在HDFS中的存储目录，系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“tenant/ta1”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。	否	租户不可用
/apps{1~5}/	固定目录	WebHCat使用到Hive的包的路径	否	执行WebHCat任务会失败
/hbase	固定目录	HBase数据存储目录	否	HBase用户数据丢失
/hbaseFileStream	固定目录	HFS文件存储目录	否	HFS文件丢失，且无法恢复

12.9.7 更改 DataNode 的存储目录

操作场景

说明

本章节适用于MRS 3.x及后续版本。

HDFS DataNode定义的存储目录不正确或HDFS的存储规划变化时，系统管理员需要在FusionInsight Manager中修改DataNode的存储目录，以保证HDFS正常工作。适用于以下场景：

- 更改DataNode角色的存储目录，所有DataNode实例的存储目录将同步修改。
- 更改DataNode单个实例的存储目录，只对单个实例生效，其他节点DataNode实例存储目录不变。

对系统的影响

- 更改DataNode角色的存储目录需要停止并重新启动HDFS服务，集群未完全启动前无法提供服务。
- 更改DataNode单个实例的存储目录需要停止并重新启动实例，该节点DataNode实例未启动前无法提供服务。

- 服务参数配置如果使用旧的存储目录，需要更新为新目录。

前提条件

- 在各个数据节点准备并安装好新磁盘，并格式化磁盘。
- 规划好新的目录路径，用于保存旧目录中的数据。
- 已安装好HDFS客户端。
- 准备好系统管理员用户**hdfs**。
- 更改DataNode单个实例的存储目录时，保持活动的DataNode实例数必须大于“dfs.replication”的值。

操作步骤

检查环境

步骤1 以**root**用户登录安装HDFS客户端的服务器，执行以下命令配置环境变量。

source *HDFS客户端安装目录*/**bigdata_env**

步骤2 如果集群为安全模式，执行以下命令认证用户身份。

kinit hdfs

步骤3 在HDFS客户端执行以下命令，检查HDFS根目录下全部目录和文件是否状态正常。

hdfs fsck /

检查fsck显示结果：

- 显示如下信息，表示无文件丢失或损坏，执行**步骤4**。
  ```
  The filesystem under path '/' is HEALTHY
  ```
- 显示其他信息，表示有文件丢失或损坏，执行**步骤5**。

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务”查看HDFS的状态“运行状态”是否为“良好”。

- 是，执行**步骤6**。
- 否，HDFS状态不健康，执行**步骤5**。

步骤5 修复HDFS异常的具体操作，任务结束。

步骤6 确定修改DataNode的存储目录场景。

- 更改DataNode角色的存储目录，执行**步骤7**。
- 更改DataNode单个实例的存储目录，执行**步骤12**。

更改DataNode角色的存储目录

步骤7 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 停止服务”，停止HDFS服务。

步骤8 以**root**用户登录到安装HDFS服务的各个数据节点中，执行如下操作：

1. 创建目标目录（data1,data2为集群原有目录）。
   例如目标目录为“${BIGDATA_DATA_HOME}/hadoop/data3/dn”：
   执行**mkdir -p ${BIGDATA_DATA_HOME}/hadoop/data3/dn**。
2. 挂载目标目录到新磁盘。例如挂载“${BIGDATA_DATA_HOME}/hadoop/data3”到新磁盘。

3. 修改新目录的权限。

   例如新目录路径为“${BIGDATA_DATA_HOME}/hadoop/data3/dn”：

   执行**chmod 700 ${BIGDATA_DATA_HOME}/hadoop/data3/dn -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/hadoop/data3/dn -R**。

4. 将数据复制到目标目录。

   例如旧目录为“${BIGDATA_DATA_HOME}/hadoop/data1/dn”，目标目录为“${BIGDATA_DATA_HOME}/hadoop/data3/dn”：

   执行**cp -af ${BIGDATA_DATA_HOME}/hadoop/data1/dn/* ${BIGDATA_DATA_HOME}/hadoop/data3/dn**。

步骤9 在FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”，打开HDFS服务配置页面。

将配置项“dfs.datanode.data.dir”从默认值“%{@auto.detect.datapart.dn}”修改为新的目标目录，例如“${BIGDATA_DATA_HOME}/hadoop/data3/dn”。

例如：原有的数据存储目录为“/srv/BigData/hadoop/data1”，“/srv/BigData/hadoop/data2”，如需将data1目录的数据迁移至新建的“/srv/BigData/hadoop/data3”目录，则将服务级别的此参数替换为现有的数据存储目录，如果有多个存储目录，用“，”隔开。则本示例中，为“/srv/BigData/hadoop/data2,/srv/BigData/hadoop/data3”。

步骤10 单击“保存”。然后在“集群 > *待操作集群的名称* > 服务”界面启动集群中各个停止的服务。

步骤11 启动HDFS成功以后，在HDFS客户端执行以下命令，检查HDFS根目录下全部目录和文件是否复制正确。

hdfs fsck /

检查fsck显示结果：

- 显示如下信息，表示无文件丢失或损坏，数据复制成功，操作结束。
  ```
  The filesystem under path '/' is HEALTHY
  ```
- 显示其他信息，表示有文件丢失或损坏，则检查8.4是否正确，并执行**hdfs fsck** *损坏的文件名称* **-delete**。

更改DataNode单个实例的存储目录

步骤12 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，勾选需要修改存储目录的DataNode单个实例，选择“更多 > 停止实例”。

步骤13 以**root**用户登录到这个DataNode节点，执行如下操作。

1. 创建目标目录。

   例如目标目录为“${BIGDATA_DATA_HOME}/hadoop/data3/dn”：

   执行**mkdir -p ${BIGDATA_DATA_HOME}/hadoop/data3/dn**。

2. 挂载目标目录到新磁盘。

   例如挂载“${BIGDATA_DATA_HOME}/hadoop/data3”到新磁盘。

3. 修改新目录的权限。

   例如新目录路径为“${BIGDATA_DATA_HOME}/hadoop/data3/dn”：

   执行**chmod 700 ${BIGDATA_DATA_HOME}/hadoop/data3/dn -R和chown omm:wheel ${BIGDATA_DATA_HOME}/hadoop/data3/dn -R**。

4. 将数据复制到目标目录。

   例如旧目录为“${BIGDATA_DATA_HOME}/hadoop/data1/dn”，目标目录为“${BIGDATA_DATA_HOME}/hadoop/data3/dn”：

   执行**cp -af ${BIGDATA_DATA_HOME}/hadoop/data1/dn/* ${BIGDATA_DATA_HOME}/hadoop/data3/dn**。

步骤14 在FusionInsight Manager管理界面，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，单击指定的DataNode实例并切换到“实例配置”页签。

将配置项“dfs.datanode.data.dir”从默认值“%{@auto.detect.datapart.dn}”修改为新的目标目录，例如“${BIGDATA_DATA_HOME}/hadoop/data3/dn”。

示例：原有的数据存储目录为“/srv/BigData/hadoop/data1,/srv/BigData/hadoop/data2”，此处如需将data1目录的数据迁移至新建的/srv/BigData/hadoop/data3目录，则将该参数修改为“/srv/BigData/hadoop/data2,/srv/BigData/hadoop/data3”。

步骤15 单击“保存”，单击“确定”。

界面提示“操作成功。”，单击“完成”。

步骤16 选择“更多 > 重启实例”，重启DataNode实例。

----结束

12.9.8 配置 HDFS 目录权限

操作场景

默认情况下，某些HDFS的文件目录权限为777或者750，存在安全风险。建议您在安装完成后修改该HDFS目录的权限，增加用户的安全性。

操作步骤

在HDFS客户端中，使用具有HDFS管理员权限的用户，执行如下命令，将“/user”的目录权限进行修改。

此处将权限修改为“1777”，即在权限处增加“1”，表示增加目录的粘性，即只有创建的用户才可以删除此目录。

hdfs dfs -chmod 1777 */user*

为了系统文件的安全，建议用户将非临时目录进行安全加固，例如：

- /user:777
- /mr-history:777
- /mr-history/tmp:777
- /mr-history/done:777
- /user/mapred:755

12.9.9 配置 NFS

操作场景

说明

本章节适用于MRS 3.x及后续版本。

用户在部署集群前，可根据需要规划Network File System（简称NFS）服务器，用于存储NameNode元数据，以提高数据可靠性。

如果您已经部署NFS服务器，并已配置NFS服务，本操作提供集群侧的配置指导，为可选任务。

操作步骤

步骤1 在NFS服务器上检查NFS的共享目录权限，确认服务器可以访问MRS集群的NameNode。

步骤2 以**root**用户登录NameNode主节点。

步骤3 执行如下命令，创建目录并赋予目录写权限。

mkdir ${BIGDATA_DATA_HOME}/namenode-nfs

chown omm:wheel ${BIGDATA_DATA_HOME}/namenode-nfs

chmod 750 ${BIGDATA_DATA_HOME}/namenode-nfs

步骤4 执行如下命令，挂载NFS到NameNode主节点。

mount -t nfs -o rsize=8192,wsize=8192,soft,nolock,timeo=3,intr *NFS服务器IP地址:共享目录* **${BIGDATA_DATA_HOME}/namenode-nfs**

例如，NFS服务器的IP为“192.168.0.11”，共享目录为“/opt/Hadoop/NameNode”，则执行命令：

mount -t nfs -o rsize=8192,wsize=8192,soft,nolock,timeo=3,intr 192.168.0.11:/opt/Hadoop/NameNode ${BIGDATA_DATA_HOME}/namenode-nfs

步骤5 在NameNode备节点上执行**步骤2**～**步骤4**。

说明

主备NameNode节点在NFS服务器上创建的共享目录名称（如“/opt/Hadoop/NameNode”）不能相同。

步骤6 登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置 > 全部配置”。

步骤7 在界面右侧的“搜索”框中输入“dfs.namenode.name.dir”搜索，在其值中增加“${BIGDATA_DATA_HOME}/namenode-nfs”路径，多个路径间使用“,”隔开，然后单击“保存”。

步骤8 单击“确定”。在概览页面选择“更多 > 重启服务”，重启服务。

----结束

12.9.10 规划 HDFS 容量

HDFS DataNode以Block的形式，保存用户的文件和目录，同时在NameNode中生成一个文件对象，对应DataNode中每个文件、目录和Block。

NameNode中文件对象需要占用一定的内存，消耗内存大小随文件对象的生成而线性递增。DataNode实际保存的文件和目录越多，NameNode文件对象总量增加，需要消耗更多的内存，使集群现有硬件可能会难以满足业务需求，且导致集群难以扩展。

规划存储大量文件的HDFS系统容量，就是规划NameNode的容量规格和DataNode的容量规格，并根据容量设置参数。

容量规格

- NameNode容量规格

  在NameNode中，每个文件对象对应DataNode中的一个文件、目录或Block。

  一个文件至少占用一个Block，默认每个Block大小为“134217728”即128MB，对应参数为“dfs.blocksize”。默认情况下一个文件小于128MB时，只占用一个Block；文件大于128MB时，占用Block数为：文件大小/128MB。目录不占用Block。

  根据“dfs.blocksize”，NameNode的文件对象数计算方法如下：

  表 12-195 NameNode 文件对象数计算

单个文件大小	文件对象数
小于128MB	1（对应文件）+1（对应Block）=2
大于128MB（例如128G）	1（对应文件）+1,024（对应128GB/128MB=1024 Block）=1,025

  主备NameNode支持最大文件对象的数量为300,000,000（最多对应150,000,000个小文件）。“dfs.namenode.max.objects”规定当前系统可生成的文件对象数，默认值为“0”表示不限制。

- DataNode容量规格

  在HDFS中，Block以副本的形式存储在DataNode中，默认副本数为“3”，对应参数为“dfs.replication”。

  集群中所有DataNode角色实例保存的Block总数为：HDFS Block * 3。集群中每个DataNode实例平均保存的Blocks= HDFS Block * 3/DataNode节点数。

  表 12-196 DataNode 支持规格

项目	规格
单个DataNode实例支持最大Block数	5,000,000
单个DataNode实例上单个磁盘支持最大Block数	500,000
单个DataNode实例支持最大Block数需要的最小磁盘数	10

表 12-197 DataNode 节点数规划

HDFS Block数	最少DataNode角色实例数
10,000,000	10,000,000 *3/5,000,000 = 6
50,000,000	50,000,000 *3/5,000,000 = 30
100,000,000	100,000,000 *3/5,000,000 = 60

内存参数设置

- NameNode JVM参数配置规则

  NameNode JVM参数“GC_OPTS”默认值为：

  -Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=128M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -Djdk.tls.ephemeralDHKeySize=3072 -Djdk.tls.rejectClientInitiatedRenegotiation=true -Djava.io.tmpdir=${Bigdata_tmp_dir}

  NameNode文件数量和NameNode使用的内存大小成比例关系，文件对象变化时请修改默认值中的“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”。参考值如下表所示。

表 12-198 NameNode JVM 配置

文件对象数量	参考值
10,000,000	“-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M”
20,000,000	“-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G”
50,000,000	“-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G”
100,000,000	“-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G”
200,000,000	“-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G”
300,000,000	“-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G”

- DataNode JVM参数配置规则

  DataNode JVM参数“GC_OPTS”默认值为：

  -Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M -XX:MetaspaceSize=128M -XX:MaxMetaspaceSize=128M -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:CMSInitiatingOccupancyFraction=65 -XX:+PrintGCDetails -Dsun.rmi.dgc.client.gcInterval=0x7FFFFFFFFFFFFE -Dsun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE -XX:-OmitStackTraceInFastThrow -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -Djdk.tls.ephemeralDHKeySize=3072 -Djdk.tls.rejectClientInitiatedRenegotiation=true -Djava.io.tmpdir=${Bigdata_tmp_dir}

  集群中每个DataNode实例平均保存的Blocks= HDFS Block * 3/DataNode节点数，单个DataNode实例平均Block数量变化时请修改默认值中的“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”。参考值如下表所示。

  表 12-199 DataNode JVM 配置

单个DataNode实例平均Block数量	参考值
2,000,000	“-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M”
5,000,000	“-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G”

  Xmx内存值对应DataNode节点块数阈值，每GB对应500000块数，用户可根据需要调整内存值。

查看 HDFS 容量状态

- NameNode信息

  MRS 3.x之前版本：登录MRS控制台，选择“组件管理 > HDFS > NameNode(主)”，单击“Overview”，查看“Summary”显示的当前HDFS中文件对象、文件数量、目录数量和Block数量信息。

  MRS 3.x及后续版本：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > NameNode(主)”，单击“Overview”，查看“Summary”显示的当前HDFS中文件对象、文件数量、目录数量和Block数量信息。

- DataNode信息

  MRS 3.x之前版本：登录MRS控制台，选择“组件管理 > HDFS > NameNode(主)”，单击“Datanodes”，查看所有告警DataNode节点的Block数量信息。

  MRS 3.x及后续版本：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > NameNode(主)”，单击“DataNodes”，查看所有告警DataNode节点的Block数量信息。

- 告警信息
  监控ID为14007、14008、14009的告警是否产生，根据业务需要修改告警阈值。

12.9.11 设置 HBase 和 HDFS 的 ulimit

现象描述

当打开一个HDFS文件时，句柄数限制导出，出现如下错误：

```
IOException (Too many open files)
```

处理步骤

您可以联系管理员增加各用户的句柄数。该配置为操作系统的配置，并非HBase或者HDFS的配置。建议管理员根据HBase和HDFS的业务量及各操作系统用户的权限进行句柄数设置。如果某一个用户需对业务量很大的HDFS进行很频繁且很多的操作，则为此用户设置较大的句柄数，避免出现以上错误。

步骤1 使用**root**用户登录集群所有节点机器或者客户端机器的操作系统，并进入“/etc/security”目录。

步骤2 执行如下命令编辑“limits.conf”文件。

vi limits.conf

新增如下内容：

```
hdfs  -       nofile  32768
hbase -       nofile  32768
```

其中“hdfs”和“hbase”表示业务中用到的操作系统用户名称。

说明

- 只有**root**用户有权限编辑“limits.conf”文件。
- 如果修改的配置不生效，请确认“/etc/security/limits.d”目录下是否有针对操作系统用户的其他nofile值。这样的值可能会覆盖“/etc/security/limits.conf”中配置的值。
- 如果用户需要对HBase进行操作，建议将该用户的句柄数设置为“10000”以上。如果用户需要对HDFS进行操作，建议根据业务量大小设置对应的句柄数，建议不要给太小的值。如果用户需要对HBase和HDFS操作，建议设置较大的值，例如“32768”。

步骤3 您可以使用如下命令查看某一用户的句柄数限制。

su - *user_name*

ulimit -n

界面会返回此用户的句柄数限制值。如下所示：

```
8194
```

----结束

12.9.12 配置 DataNode 容量均衡

操作场景

说明

本章节适用于MRS 3.x及后续版本。

HDFS集群可能出现DataNode节点间磁盘利用率不平衡的情况，比如集群中添加新数据节点的场景。如果HDFS出现数据不平衡的状况，可能导致多种问题，比如MapReduce应用程序无法很好地利用本地计算的优势、数据节点之间无法达到更好的网络带宽使用率或节点磁盘无法利用等等。所以系统管理员需要定期检查并保持DataNode数据平衡。

HDFS提供了一个容量均衡程序Balancer。通过运行这个程序，可以使得HDFS集群达到一个平衡的状态，使各DataNode磁盘使用率与HDFS集群磁盘使用率的偏差不超过阈值。图12-16和图12-17分别是Balance前后DataNode的磁盘使用率变化。

图 12-16 执行均衡操作前 DataNode 的磁盘使用率

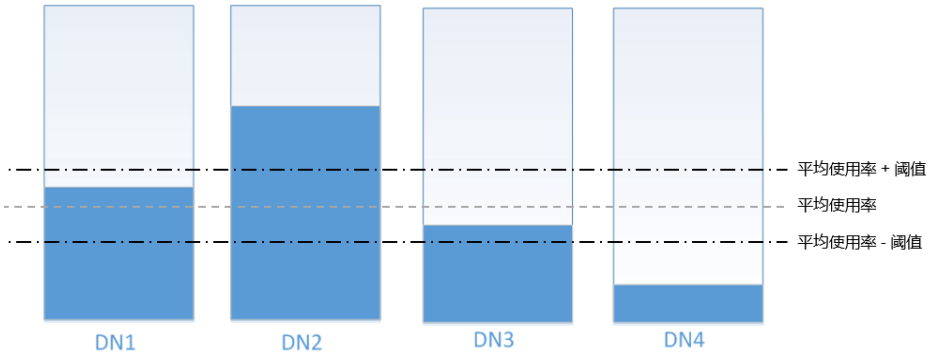
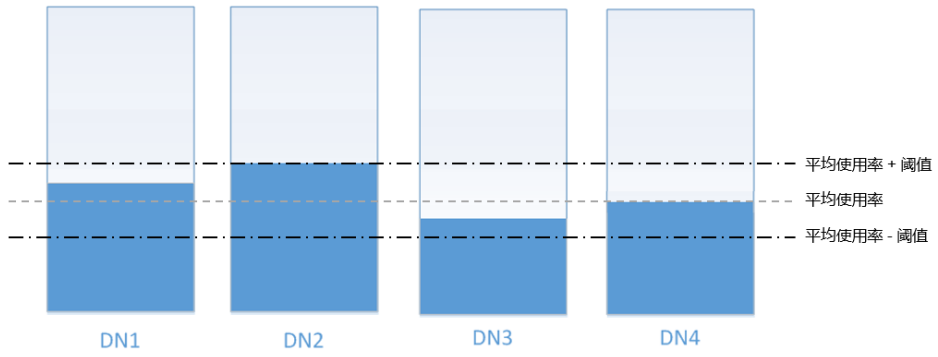


图 12-17 执行均衡操作后 DataNode 的磁盘使用率



均衡操作时间估算受两个因素影响：

1. 需要迁移的总数据量：

   每个DataNode节点的数据量应大于（平均使用率-阈值）*平均数据量，小于（平均使用率+阈值）*平均数据量。若实际数据量小于最小值或大于最大值即存在不平衡，系统选择所有DataNode节点中偏差最多的数据量作为迁移的总数据量。
2. Balancer的迁移是按迭代（iteration）方式串行顺序处理的，每个iteration迁移数据量不超过10GB，每个iteration重新计算使用率的情况。

因此针对集群情况，可以大概估算每个iteration耗费的时间（可以通过执行Balancer的日志观察到每次iteration的时间），并用总数据量除以10GB估算任务执行时间。

由于按iteration处理，Balancer可以随时启动或者停止。

对系统的影响

- 执行Balance操作时会占用DataNode的网络带宽资源，请根据业务需求在维护期间执行任务。
- 默认使用带宽控制为20MB/s，如果重新设置带宽流量或加大数据量，Balance操作可能会对正在运行的业务产生影响。

前提条件

已安装HDFS客户端。

操作步骤

步骤1 使用客户端安装用户登录客户端所在节点。执行命令切换到客户端安装目录，例如“/opt/client”。

cd /opt/client

说明

如果集群为普通模式，需先执行**su - omm**切换为**omm**用户。

步骤2 执行以下命令配置环境变量。

source bigdata_env

步骤3 如果集群为安全模式，执行以下命令认证**hdfs**身份。

kinit hdfs

步骤4 是否调整带宽控制?

- 是，执行**步骤5**。
- 否，执行**步骤6**。

步骤5 执行以下命令，修改Balance的最大带宽，然后执行**步骤6**。

hdfs dfsadmin -setBalancerBandwidth *<bandwidth in bytes per second>*

<bandwidth in bytes per second>表示带宽控制的数值，单位为字节。例如要设置带宽控制为20MB/s，对应值为20971520，完整命令为：

hdfs dfsadmin -setBalancerBandwidth 20971520

说明

- 默认为20MB/s，适用于当前集群使用万兆网络，且有业务正在执行的场景。若没有足够的业务空闲时间窗用于Balance维护，可适当增加该值以缩短Balance时间，如增大到209715200（即200MB/s）。
- 这个参数的调整要看组网情况，如果集群负载较高，可以改为209715200(200MB/s)；如果集群空闲，可以改为1073741824 (1GB/s)。
- 如果DataNode节点的带宽无法达到指定的最大带宽，可以在FusionInsight Manager修改HDFS的参数“dfs.datanode.balance.max.concurrent.moves”，将每个DataNode节点执行均衡的线程数修改为“32”，并重启HDFS服务。

步骤6 执行以下命令，启动Balance任务。

bash /opt/client/HDFS/hadoop/sbin/start-balancer.sh -threshold *<threshold of balancer>*

-threshold表示HDFS数据达到平衡状态时DataNode磁盘使用率偏差值，各个DataNode节点磁盘的使用率和整体HDFS集群的磁盘空间平均使用率偏差小于此阈值时，系统认为HDFS集群已经达到了平衡的状态并结束Balance任务。

例如，需要设置偏差率为5%，则执行：

bash /opt/client/HDFS/hadoop/sbin/start-balancer.sh -threshold 5

说明

- 上述命令会在后台执行该任务，相关日志可以通过客户端安装目录“/opt/client/HDFS/hadoop/logs”下的hadoop-root-balancer-*主机名*.out查看。
- 如果需要停止Balance任务，请执行以下命令：

  bash /opt/client/HDFS/hadoop/sbin/stop-balancer.sh
- 如果只需要对部分节点进行数据均衡，可以在脚本上加上-include参数指定要移动的节点。具体参数使用方法，可通过命令行查看。
- “/opt/client”为客户端安装目录，如果不一致，替换即可。
- 如果该命令执行失败，在日志中看到的错误信息为“Failed to APPEND_FILE /system/balancer.id”，则需要执行如下命令强行删除“/system/balancer.id”，再次执行**start-balancer.sh**脚本即可。

  hdfs dfs -rm -f /system/balancer.id

步骤7 界面提示以下信息表示均衡操作已完成，系统将自动退出任务：

```
Apr 01, 2016 01:01:01 PM  Balancing took 23.3333 minutes
```

用户在执行了**步骤6**的脚本后，会在客户端安装目录“/opt/client/HDFS/hadoop/logs”目录下生成名为hadoop-root-balancer-*主机名*.out日志。打开该日志可以看到如下字段信息：

- Time Stamp：时间戳
- Bytes Already Moved：已经移动的字节数
- Bytes Left To Move：待移动的字节数
- Bytes Being Moved：正在移动的字节数

----结束

相关任务

设置自动执行Balance任务

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，选择“全部配置”，搜索以下参数名并修改参数值。

- “dfs.balancer.auto.enable”表示是否启用自动执行Balance任务，默认值为“false”表示不启用，修改为“true”表示启用。
- “dfs.balancer.auto.cron.expression”表示任务执行的时间，默认值“0 1 * * 6”表示在每周六的1点执行任务。仅在启用自动执行Balance功能时有效。

  修改此参数时，表达式介绍如**表12-200**所示。支持“*”表示连续的时间段。

表 12-200 执行表达式参数解释

列	说明
第1列	分钟，参数值为0～59。
第2列	小时，参数值为0～23。
第3列	日期，参数值为1～31。
第4列	月份，参数值为1～12。
第5列	星期，参数值为0～6，0表示星期日。

- “dfs.balancer.auto.stop.cron.expression”表示任务自动停止的时间，默认值为空，表示不自动停止正在运行的Balancer任务。以“0 5 * * 6”为例，则表示在每周六的5点停止正在运行的Balancer任务。仅在启用自动执行Balance功能时有效。

  修改此参数时，表达式介绍如**表12-200**所示。支持“*”表示连续的时间段。

步骤3 修改自动Balancer的运行参数，如**表12-201**所示：

表 12-201 自动 Balancer 运行参数

参数名	参数介绍	默认值
dfs.balancer.auto.threshold	表示磁盘容量百分比的均衡阈值。仅当dfs.balancer.auto.enable设置为true时才有效。	10
dfs.balancer.auto.exclude.datanodes	不需要执行磁盘自动均衡的DataNode列表，用逗号分隔。仅当dfs.balancer.auto.enable设置为true时才有效。	默认为空
dfs.balancer.auto.bandwidthPerSec	每个DataNode可用于负载均衡的最大带宽量（单位：MB/s）。	20
dfs.balancer.auto.maxIdleIterations	Balancer的最大连续空闲迭代次数。一次空闲迭代为没有Block块被移动的迭代，当连续空闲迭代次数达到最大连续空闲迭代次数时，本次Balancer结束。当取值为-1时，代表无穷大。	5
dfs.balancer.auto.maxDataNodesNum	该参数用来控制进行自动Balancer的DataNode数量。假设该参数值为N，当N大于0， 则选择剩余空间比例最高的N个DataNode和最低的N个DataNode之间进行数据均衡；当N等于0， 则对集群中所有DataNode进行数据均衡。	5

步骤4 单击“保存”使配置生效。无需重启HDFS服务。

任务执行日志保存在主NameNode节点中，请查看“/var/log/Bigdata/hdfs/nn/hadoop-omm-balancer-*主机名*.log”。

----结束

12.9.13 配置 DataNode 节点间容量异构时的副本放置策略

操作场景

默认情况下，NameNode会随机选择DataNode节点写文件。当集群内某些数据节点的磁盘容量不一致（某些节点的磁盘总容量大，某些总容量小），会导致磁盘总容量小的节点先写满。通过修改集群默认的DataNode写数据时的磁盘选择策略为“节点磁盘可用空间块放置策略”，可提高将块数据写到磁盘可用空间较大节点的概率，解决因为数据节点磁盘容量不一致导致的节点使用率不均衡的情况。

对系统的影响

修改磁盘选择策略为“节点磁盘可用空间块放置策略（org.apache.hadoop.hdfs.server.blockmanagement.AvailableSpaceBlockPlacementPolicy）”，经过测试验证，在该测试结果中，修改前后，HDFS写文件性能影响范围在3%以内。

说明

NameNode默认的副本存储策略为：

1. 第一副本：存放到客户端所在节点。
2. 第二副本：远端机架的数据节点。
3. 第三副本：存放到客户端所在节点的相同机架的不同节点。

如还有更多副本，则随机选择其它DataNode。

“节点磁盘可用空间块放置策略”的副本选择机制为：

1. 第一个副本：存放在客户端所在DataNode（和默认的存放策略一样）。
2. 第二个副本：
   - 选择存储节点的时候，先挑选2个满足要求的数据节点。
   - 比较这2个节点磁盘空间使用比例，如果磁盘空间使用率的相差小于5%，随机存放到第一个节点。
   - 如果磁盘空间使用率相差超过5%，即有60%（由dfs.namenode.available-space-block-placement-policy.balanced-space-preference-fraction指定，默认值0.6）的概率写到磁盘空间使用率低的节点。
3. 第三副本等其他后续副本的存储情况，也参考第二个副本的选择方式。

前提条件

集群里DataNode节点的磁盘总容量偏差不能超过100%。

操作步骤

步骤1 请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面。

步骤2 调整HDFS写数据时的依据的磁盘选择策略参数。搜索“dfs.block.replicator.classname”参数，并将参数的值改为

“org.apache.hadoop.hdfs.server.blockmanagement.AvailableSpaceBlockPlacementPolicy”。

步骤3 保存修改的配置。保存完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

12.9.14 配置 HDFS 单目录文件数量

操作场景

通常一个集群上部署了多个服务，且大部分服务的存储都依赖于HDFS文件系统。当集群运行时，不同组件（例如Spark、Yarn）或客户端可能会向同一个HDFS目录不断写入文件。但HDFS系统支持的单目录文件数目是有上限的，因此用户需要提前做好规划，防止单个目录下的文件数目超过阈值，导致任务出错。

HDFS提供了“dfs.namenode.fs-limits.max-directory-items”参数设置单个目录下可以存储的文件数目。

操作步骤

步骤1 请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面。

步骤2 搜索配置项“dfs.namenode.fs-limits.max-directory-items”。

表 12-202 参数说明

参数名称	描述	默认值
dfs.namenode.fs-limits.max-directory-items	定义目录中包含的最大条目数。 取值范围：1～6400000	1048576

步骤3 设置单个HDFS目录下最大可容纳的文件数目。保存修改的配置。保存完成后请重新启动配置过期的服务或实例以使配置生效。

说明

用户尽量将数据做好存储规划，可以按时间、业务类型等分类，不要单个目录下直属的文件过多，建议使用默认值，单个目录下约100万条。

----结束

12.9.15 配置回收站机制

配置场景

在HDFS中，删除的文件将被移动到回收站（trash）中，以便在误操作的情况下恢复被删除的数据。

您可以设置文件保留在回收站中的时间阈值，一旦文件保存时间超过此阈值，将从回收站中永久地删除。如果回收站被清空，回收站中的所有文件将被永久删除。

配置描述

在HDFS中，如果删除HDFS的文件，文件会被保存到trash空间中，不会被立即清除。被删除的文件在超过老化时间后将变为老化文件，会基于系统机制清除或用户手动清除。

参数入口：

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-203 参数说明

参数	描述	默认值
fs.trash.interval	以分钟为单位的垃圾回收时间，垃圾站中数据超过此时间，会被删除。取值范围：1440～259200。	1440
fs.trash.checkpoint.interval	垃圾检查点间的间隔。单位：分钟。应小于等于fs.trash.interval的值。检查点程序每次运行时都会创建一个新的检查点并会移除fs.trash.interval分钟前创建的检查点。例如，系统每10分钟检测是否存在老化文件，如果发现有老化文件，则删除。对于未老化文件，则会存储在checkpoint列表中，等待下一次检查。 如果此参数的值设置为0，则表示系统不会检查老化文件，所有老化文件会被保存在系统中。 取值范围：0～*fs.trash.interval*。 **说明** 不推荐将此参数值设置为0，这样系统的老化文件会一直存储下去，导致集群的磁盘空间不足。	60

12.9.16 配置文件和目录的权限

配置场景

HDFS支持用户进行文件和目录默认权限的修改。HDFS默认用户创建文件和目录的权限的掩码为“022”，如果默认权限满足不了用户的需求，可以通过配置项进行默认权限的修改。

配置描述

参数入口：

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-204 参数说明

参数	描述	默认值
fs.permissions.umask-mode	当客户端在HDFS上创建文件和目录时使用此umask值（用户掩码）。类似于linux上的文件权限掩码。 可以使用八进制数字也可以使用符号，例如：“022”（八进制，等同于以符号表示的u=rwx,g=r-x,o=r-x），或者“u=rwx,g=rwx,o=”（符号法，等同于八进制的“007”）。 **说明** 8进制的掩码，和实际权限设置值正好相反，建议使用符号表示法，描述更清晰。	022

12.9.17 配置 token 的最大存活时间和时间间隔

配置场景

安全模式下，HDFS中用户可以对token的最大存活时间和token renew的时间间隔进行灵活地设置，根据集群的具体需求合理地配置。

配置描述

参数入口：

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-205 参数说明

参数	描述	默认值
dfs.namenode.delegation.token.max-lifetime	该参数为服务器端参数，设置token的最大存活时间，单位为毫秒。取值范围：10000~10000000000000。	604800000
dfs.namenode.delegation.token.renew-interval	该参数为服务器端参数，设置token renew的时间间隔，单位为毫秒。取值范围：10000~10000000000000。	86400000

12.9.18 配置磁盘坏卷

配置场景

在开源版本中，如果为DataNode配置多个数据存放卷，默认情况下其中一个卷损坏，则DataNode将不再提供服务。用户可以通过修改配置项“dfs.datanode.failed.volumes.tolerated”的值，指定失败的个数，小于该个数，DataNode可以继续提供服务。

配置描述

参数入口：

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-206 参数说明

参数	描述	默认值
dfs.datanode.failed.volumes.tolerated	DataNode停止提供服务前允许失败的卷数。默认情况下，必须至少有一个有效卷。值-1表示有效卷的最小值是1。大于等于0的值表示允许失败的卷数。	MRS 3.x之前版本：0 MRS 3.x及之后版本：-1

12.9.19 使用安全加密通道

配置场景

安全加密通道是HDFS中RPC通信的一种加密协议，当用户调用RPC时，用户的login name会通过RPC头部传递给RPC，之后RPC使用Simple Authentication and Security Layer（SASL）确定一个权限协议（支持Kerberos和DIGEST-MD5两种），完成RPC授权。用户在部署安全集群时，需要使用安全加密通道，配置如下参数。安全Hadoop RPC相关信息请参考：**https://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-common/SecureMode.html#Data_Encryption_on_RPC**

配置描述

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-207 参数说明

参数	描述	默认值
hadoop.rpc.protection	**须知** • 设置后需要重启服务生效，且不支持滚动重启。 • 设置后需要重新下载客户端配置，否则HDFS无法提供读写服务。 设置Hadoop中各模块的RPC通道是否加密。通道包括： • 客户端访问HDFS的RPC通道。 • HDFS中各模块间的RPC通道，如DataNode与NameNode间的RPC通道。 • 客户端访问Yarn的RPC通道。 • NodeManager和ResourceManager间的RPC通道。 • Spark访问Yarn，Spark访问HDFS的RPC通道。 • Mapreduce访问Yarn，Mapreduce访问HDFS的RPC通道。 • HBase访问HDFS的RPC通道。 **说明** 用户可在HDFS组件的配置界面中设置该参数的值，设置后全局生效，即Hadoop中各模块的RPC通道的加密属性全部生效。 对RPC的加密方式，有如下三种取值： • “authentication”：普通模式默认值，指数据在鉴权后直接传输，不加密。这种方式能保证性能但存在安全风险。 • “integrity”：指数据直接传输，即不加密也不鉴权。 为保证数据安全，请谨慎使用这种方式。 • “privacy”：安全模式默认值，指数据在鉴权及加密后再传输。这种方式会降低性能。	• 安全模式：privacy • 普通模式：authentication

12.9.20 在网络不稳定的情况下，降低客户端运行异常概率

配置场景

在网络不稳定的情况下，调整如下参数，降低客户端应用运行异常概率。

配置描述

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-208 参数说明

参数	描述	默认值
ha.health-monitor.rpc-timeout.ms	zkfc对namenode健康状态检查的超时时间。增大该参数值，可以防止出现双Active NameNode，降低客户端应用运行异常的概率。 单位：毫秒。取值范围：30000~3600000	180000
ipc.client.connect.max.retries.on.timeouts	客户端与服务端建立Socket连接超时时，客户端的重试次数。 取值范围：1~256	45
ipc.client.connect.timeout	客户端与服务端建立socket连接的超时时间。增大该参数值，可以增加建立连接的超时时间。 单位：毫秒。取值范围：1~3600000	20000

12.9.21 配置 NameNode blacklist

配置场景

说明

本章节适用于MRS 3.x及后续版本。

在现有的缺省DFSclient failover proxy provider中，一旦某进程中的一个NameNode发生故障，在同一进程中的所有HDFS client实例都会尝试再次连接NameNode，导致应用长时间等待超时。

当位于同一JVM进程中的客户端对无法访问的NameNode进行连接时，会对系统造成负担。为了避免这种负担，MRS集群搭载了NameNode blacklist功能。

在新的Blacklisting DFSClient failover provider中，故障的NameNode将被记录至一个列表中。DFSClient会利用这些信息，防止客户端再次连接这些NameNode。该功能被称为NameNode blacklisting。

例如，如下集群配置：

namenode：nn1、nn2

dfs.client.failover.connection.retries：20

单JVM中的进程：10个客户端

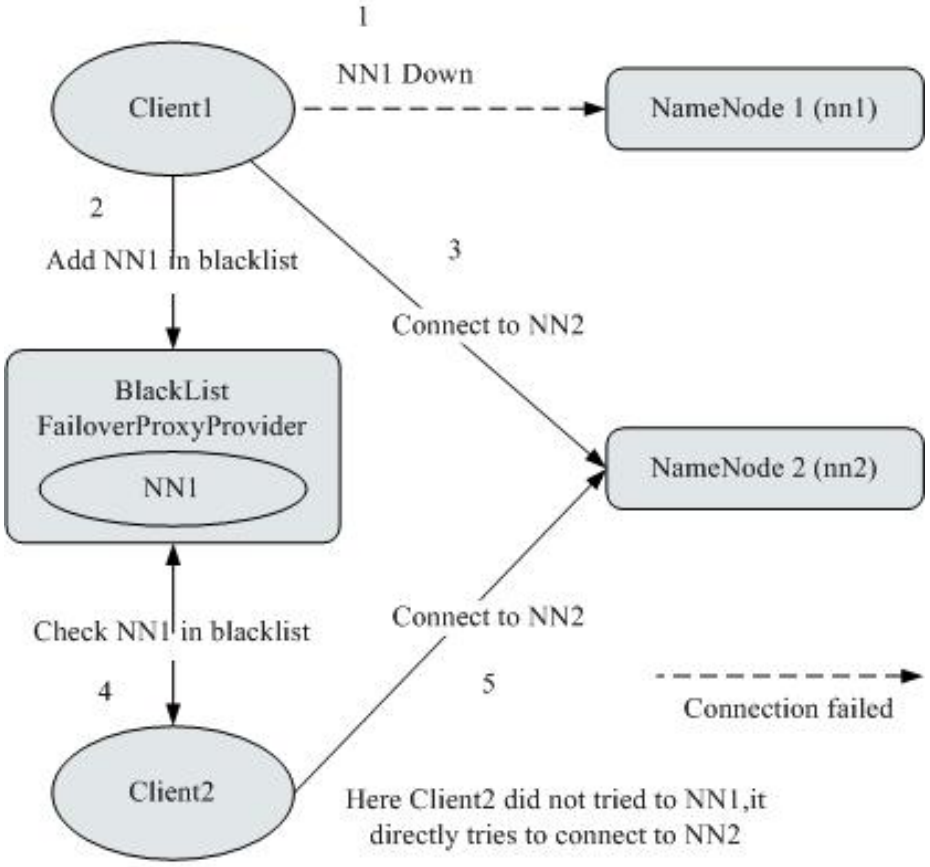
在上述集群中，如果当前处于active状态的nn1无法访问，client1将会对nn1进行20次重新连接，之后发生故障转移，client1将会连接至nn2。与此相同，client2至client10也会在对nn1进行20次重新连接后连接至nn2。这样会延长NameNode的整体故障恢复时间。

针对该情况，当client1试图连接当前处于active状态的nn1，但其已经发生故障时，nn1将会被添加至blacklist。这样其余client就不会连接已被添加至blacklist的nn1，而是会选择连接nn2。

说明

若在任一时刻，所有NameNode都被添加至blacklist，则其内容会被清空，client会按照初始的NameNode list重新尝试连接。若再次出现任何故障，NameNode仍会被添加至blacklist。

图 12-18 NameNode blacklisting 状态图



配置描述

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-209 NameNode blacklisting 的相关参数

参数	描述	默认值
dfs.client.failover.proxy.provider.*[nameservice ID]*	利用已通过的协议创建namenode代理的Client Failover proxy provider类。 将参数值设置为“org.apache.hadoop.hdfs.server.namenode.ha.BlackListingFailoverProxyProvider”， 可使用从NameNode支持读的特性。	org.apache.hadoop.hdfs.server.namenode.ha.AdaptiveFailoverProxyProvider

12.9.22 优化 HDFS NameNode RPC 的服务质量

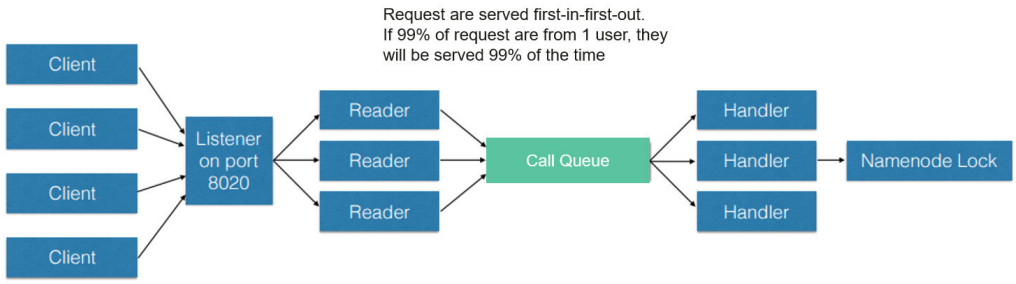
配置场景

> **说明**
>
> 本章节适用于MRS 3.x及后续版本。

数个成品Hadoop集群由于NameNode超负荷运行并失去响应而发生故障。

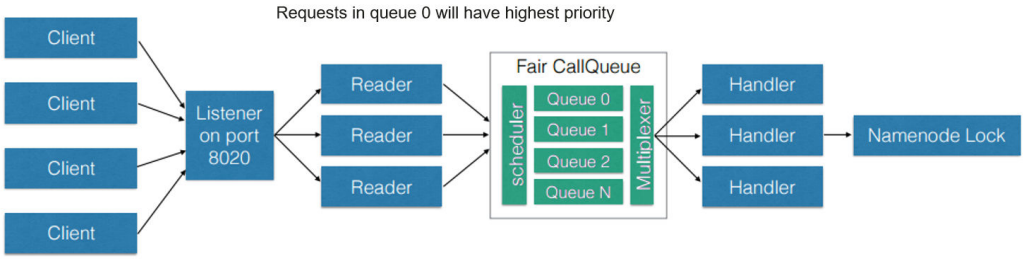
这种阻塞现象是由于Hadoop的初始设计造成的。在Hadoop中，NameNode作为单独的机器，在其namespace内协调HDFS的各种操作。这些操作包括获取数据块位置，列出目录及创建文件。NameNode接受HDFS的操作，将其视作RPC调用并置入FIFO调用队列，供读取线程处理。虽然FIFO在先到先服务的情况下足够公平，但如果用户执行的I/O操作较多，相比I/O操作较少的用户，将获得更多的服务。在这种情况下，FIFO有失公平并且会导致延迟增加。

图 12-19 基于 FIFO 调用队列的 NameNode 请求处理



如果将FIFO队列替换为一种被称作FairCallQueue的新型队列，这种情况就能够得到改善。按照这种方法，FAIR队列会根据调用者的调用规模将传入的RPC调用分配至多个队列中。调度模块会跟踪最新的调用，并为调用量较小的用户分配更高的优先级。

图 12-20 基于 FAIRCallQueue 的 NameNode 请求处理



配置描述

- FairCallQueue通过在内部调整RPC调用的顺序确保服务质量。

  该队列由以下三部分组成：

  a. 调度模块（DecayRpcScheduler）用于提供从0至N的优先值数字（0的优先级最高）。

  b. 多级队列（位于FairCallQueue内部）保持调用在内部按优先级排列。

  c. 多路转换器（提供有WeightedRoundRobinMultiplexer）为队列选择提供逻辑控制。

在对FairCallQueue进行配置后，由控制模块决定将收到的调用分配至哪个子队列。当前调度模块为DecayRpcScheduler。该模块仅持续对各类调用的优先级数字进行追踪，并周期性地对这些数字进行减小处理。

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-210 Fair 调用队列参数

参数	描述	默认值
ipc.*<port>*.callqueue.impl	队列的实现类。用户需要通过“org.apache.hadoop.ipc.FairCallQueue”启用QoS特性。	java.util.concurrent.LinkedBlockingQueue

- RPC BackOff

  Backoff是FairCallQueue的功能之一，要求客户端在一段时间后重试操作（如创建，删除，打开文件等）。当Backoff发生时，RCP服务器将抛出RetriableException异常。FairCallQueue在以下两种情况时进行Backoff。

  - 当队列已满，即队列中有许多客户端调用时。
  - 当队列的响应时间大于配置的阈值（由参数“ipc.<port>.decay-scheduler.backoff.responsetime.thresholds”决定）时。

表 12-211 RPC BackOff 配置

参数	描述	默认值
ipc.*<port>*.backoff.enable	启用Backoff配置参数。当前，如果应用程序中包含较多的用户调用，假设没有达到操作系统的连接限制，则RPC请求将处于阻塞状态。或者，当RPC或NameNode在重负载时，可以基于某些策略将一些明确定义的异常抛回给客户端，客户端将理解这种异常并进行指数回退，以此作为类RetryInvocationHandler的另一个实现。	false
ipc.*<port>*.decay-scheduler.backoff.responsetime.enable	根据队列平均响应时间启用Backoff。	false
ipc.*<port>*.decay-scheduler.backoff.responsetime.thresholds	配置每个队列的响应时间阈值。ResponseTime阈值必须与优先级数目（ipc.<port> .faircallqueue.priority-levels）相匹配。单位：毫秒。	10000,20000,30000,40000

说明

- <port>表示在NameNode上配置的RPC端口。
- 只有在“ipc.<port> .backoff.enable”为“true”时，响应时间backoff功能才会起作用。

12.9.23 优化 HDFS DataNode RPC 的服务质量

配置场景

当客户端写入HDFS的速度大于DataNode的硬盘带宽时，硬盘带宽会被占满，导致DataNode失去响应。客户端只能通过取消或恢复通道进行规避，这会导致写入失败及不必要的通道恢复操作。

说明

本章节适用于MRS 3.x及后续版本。

配置步骤

引入了新的配置参数“dfs.pipeline.ecn”。当该配置启用时，DataNode会在写入通道超出负荷时从其中发出信号。客户端可以基于该阻塞信号进行退避，从而防止系统超出负荷。引入该配置参数的目的是为了使通道更加稳定，并减少不必要的取消或恢复操作。收到信号后，客户端会退避一定的时间（5000ms），然后根据相关过滤器调整退避时间（单次退避最长时间为50000ms）。

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-212 DN ECN 配置

参数	描述	缺省值
dfs.pipeline.ecn	进行该配置后，DataNode能够向客户端发送阻塞通知。	false

12.9.24 配置 DataNode 预留磁盘百分比

配置场景

当YARN本地目录和DataNode目录配置在同一个磁盘时，具有较大容量的磁盘可以运行更多的任务，因此将有更多的中间数据存储在YARN本地目录。

目前DataNode支持通过配置“dfs.datanode.du.reserved”来配置预留磁盘空间大小的绝对值。配置较小的数值不能满足更大的磁盘要求。但对于更小的磁盘配置更大的数值将浪费大量的空间。

为了避免这种情况，添加一个新的参数“dfs.datanode.du.reserved.percentage”来配置预留磁盘空间占总磁盘空间大小的百分比，那样可以基于总的磁盘空间来预留磁盘百分比。

说明

- 如果用户同时配置“dfs.datanode.du.reserved.percentage”和“dfs.datanode.du.reserved”，则采用这两个参数较大的数值作为DataNode的预留空间大小。
- 建议基于磁盘空间设置“dfs.datanode.du.reserved”或者“dfs.datanode.du.reserved.percentage”。

配置描述

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-213 参数描述

参数	描述	默认值
dfs.datanode.du.reserved.percentage	DataNode预留空间占总磁盘空间大小的百分比。DataNode会永久预留由此百分比计算得出的磁盘空间大小。 整数值，取值范围是0～100。	10

12.9.25 配置 HDFS NodeLabel

配置场景

用户需要通过数据特征灵活配置HDFS文件数据块的存储节点。通过设置HDFS目录/文件对应一个标签表达式，同时设置每个Datanode对应一个或多个标签，从而给文件的数据块存储指定了特定范围的Datanode。

当使用基于标签的数据块摆放策略，为指定的文件选择DataNode节点进行存放时，会根据文件的标签表达式选择出Datanode节点范围，然后在这些Datanode节点范围内，选择出合适的存放节点。

说明

本章节适用于MRS 3.x及后续版本。

开启单集群跨AZ高可用后，不支持配置HDFS NodeLabel功能。

- 场景1 DataNodes分区场景。

  场景说明：

  用户需要让不同的应用数据运行在不同的节点，分开管理，就可以通过标签表达式，来实现不同业务的分离，指定业务存放到对应的节点上。

  通过配置NodeLabel特性使得：

  - /HBase下的数据存储在DN1、DN2、DN3、DN4节点上。
  - /Spark下的数据存储在DN5、DN6、DN7、DN8节点上。

图 12-21 DataNode 分区场景

说明

- 通过**hdfs nodelabel -setLabelExpression -expression 'LabelA[fallback=NONE]' -path /Hbase**命令，给Hbase目录设置表达式。从图12-21中可知，“/Hbase”文件的数据块副本会被放置在有LabelA标签的节点上，即DN1、DN2、DN3、DN4。同理，通过**hdfs nodelabel -setLabelExpression -expression 'LabelB[fallback=NONE]' -path /Spark**命令，给Spark目录设置表达式。在“/Spark”目录下文件对应的数据块副本只能放置到LabelB标签上的节点，如DN5、DN6、DN7、DN8。
- 设置数据节点的标签参考**配置描述**。
- 如果同一个集群上存在多个机架，每个标签下需要有多个机架的datanodes，以确保数据块摆放的可靠性。

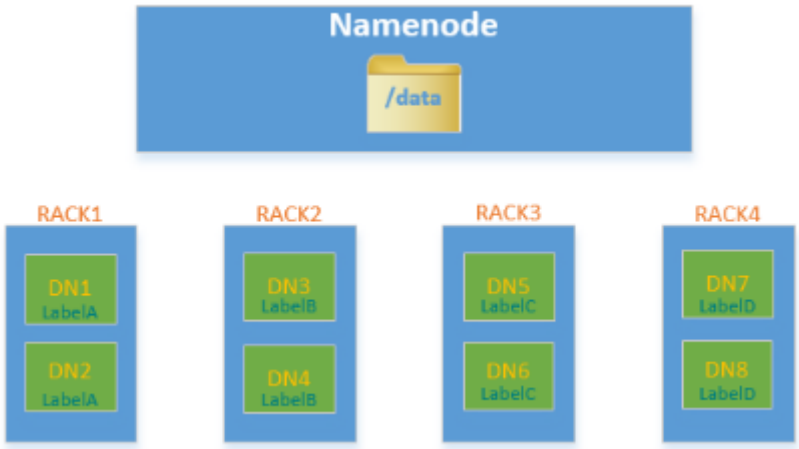
- 场景2 多机架下指定副本位置场景

  场景说明：

  在异构集群中，客户需要分配一些特定的具有高可靠性的节点用以存放重要的商业数据，可以通过标签表达式指定副本位置，指定文件数据块的其中一个副本存放到高可靠性的节点上。

  “/data”目录下的数据块，默认三副本情况下，其中至少有一个副本会被存放到RACK1或RACK2机架的节点上（RACK1和RACK2机架的节点为高可靠性节点），另外两个副本会被分别存放到RACK3和RACK4机架的节点上。

图 12-22 场景样例



说明

通过 **hdfs nodelabel -setLabelExpression -expression 'LabelA|| LabelB[fallback=NONE],LabelC,LabelD' -path /data**命令给“/data”目录设置表达式。

当向“/data”目录下写数据时，至少有一个数据块副本存放在LabelA或者LabelB标签的节点中，剩余的两个数据块副本会被存放在有LabelC和LabelD标签的节点上。

配置描述

- Datanode节点标签配置

  请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

  表 12-214 参数说明

参数	描述	默认值
dfs.block.replicator.classname	配置HDFS的DataNode原则策略。 如果需要开启NodeLabeNodel功能，需要将该值设置为org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyWithNodeLabel。	org.apache.hadoop.hdfs.server.blockmanagement.AvailableSpaceBlockPlacementPolicy
host2tags	配置DataNode主机与标签的对应关系。 主机名称支持配置IP扩展表达式（如192.168.1.[1-128]或者192.168.[2-3].[1-128]，且IP必须为业务IP），或者为前后加上 / 的主机名的正则表达式（如/datanode-[123]/或者/datanode-\d{2}/）。标签配置名称不允许包含 = / \ 字符。【注意】配置IP时必须是业务IP。	-

说明

- host2tags配置项内容详细说明：

  假如有一套集群，有20个Datanode：dn-1到dn-20，对应的IP地址为10.1.120.1到10.1.120.20，host2tags配置文件内容可以使用如下的表示方式。

  主机名正则表达式

  “/dn-\d/ = label-1”表示dn-1到dn-9对应的标签为label-1，即dn-1 = label-1，dn-2 = label-1，...dn-9 = label-1。

  “/dn-((1[0-9]$)|(20$))/ = label-2”表示dn-10到dn-20对应的标签为label-2，即dn-10 = label-2，dn-11 = label-2，...dn-20 = label-2。

  IP地址范围表示方式

  “10.1.120.[1-9] = label-1”表示10.1.120.1到10.1.120.9对应的标签为label-1，即10.1.120.1 = label-1，10.1.120.2 = label-1，...10.1.120.9 = label-1。

  “10.1.120.[10-20] = label-2”表示10.1.120.10到10.1.120.20对应的标签为label-2，即10.1.120.10 = label-2，10.1.120.11 = label-2，...10.1.120.20 = label-2。

- 基于标签的数据块摆放策略支持扩容减容场景：

  当集群中新增加DataNode节点时，如果该DataNode对应的IP匹配**host2tags**配置项中的IP地址范围，或者该DataNode的主机名匹配**host2tags**配置项中的主机名正则表达式，则该DataNode节点会被设置成对应的标签。

  例如“host2tags”配置值为10.1.120.[1-9] = label-1，而当前集群只有10.1.120.1到10.1.120.3三个数据节点。进行扩容后，又添加了10.1.120.4这个数据节点，则该数据节点会被设置成label-1的标签；如果10.1.120.3这个数据节点被删除或者退出服务后，数据块不会再被分配到该节点上。

- 设置目录/文件的标签表达式
  - 在HDFS参数配置页面配置“path2expression”，配置HDFS目录与标签的对应关系。当配置的HDFS目录不存在时，也可以配置成功，新建不存在的同名目录，已设置的标签对应关系将在30分钟之内被继承。设置了标签的目录被删除后，新增一个同名目录，原有的对应关系也将在30分钟之内被继承。
  - 命令行设置方式请参考**hdfs nodelabel -setLabelExpression**命令。
  - Java API设置方式通过NodeLabelFileSystem实例化对象调用setLabelExpression(String src, String labelExpression)方法。*src*为HDFS上的目录或文件路径，“labelExpression”为标签表达式。
- 开启NodeLabel特性后，可以通过命令**hdfs nodelabel -listNodeLabels**查看每个Datanode的标签信息。

块副本位置选择

Nodelabel支持对各个副本的摆放采用不同的策略，如表达式“label-1,label-2,label-3”，表示3个副本分别放到含有label-1、label-2、label-3的DataNode中，不同的副本策略用逗号分隔。

如果label-1，希望放2个副本，可以这样设置表达式：“label-1[replica=2],label-2,label-3”。这种情况下，如果默认副本数是3，则会选择2个带有label-1和一个label-2的节点；如果默认副本数是4，会选择2个带有label-1、一个label-2以及一个label-3的节点。可以注意到，副本数是从左到右依次满足各个副本策略的，但也有副本数超过表达式表述的情况，当默认副本数为5时，多出来的一个副本会放到最后一个节点中，也就是label-3的节点里。

当启用ACLs功能并且用户无权访问表达式中使用的标签时，将不会为副本选择属于该标签的DataNode。

多余块副本删除选择

如果块副本数超过参数“dfs.replication”值（即用户指定的文件副本数），hdfs会删除多余块副本来保证集群资源利用率。

删除规则如下：

- 优先删除不满足任何表达式的副本。

  示例：文件默认副本数为3

  /test标签表达式为“LA[replica=1],LB[replica=1],LC[replica=1]”，

  /test文件副本分布的四个节点（D1～D4）以及对应标签（LA～LD）：

  ```
  D1:LA
  D2:LB
  D3:LC
  D4:LD
  ```

  则选择删除D4节点上的副本块。

- 如果所有副本都满足表达式，删除多于表达式指定的数量的副本。

  示例：文件默认副本数为3

  /test标签表达式为“LA[replica=1],LB[replica=1],LC[replica=1]”，

  /test文件副本分布的四个节点以及对应标签：

  ```
  D1:LA
  D2:LA
  D3:LB
  D4:LC
  ```

  则选择删除D1或者D2上的副本块。

- 如果文件所有者或文件所有者的组不能访问某个标签，则优先删除映射到该标签的DataNode中的副本。

基于标签的数据块摆放策略样例

假如有一套集群，有六个DataNode：dn-1，dn-2，dn-3，dn-4，dn-5以及dn-6，对应的IP为10.1.120.[1-6]。有六个目录需要配置标签表达式，Block默认备份数为3。

- 下面给出3种DataNode标签信息在“host2labels”文件中的表示方式，其作用是一样的。
  - 主机名正则表达式

    ```
    /dn-[1456]/ = label-1,label-2
    /dn-[26]/ = label-1,label-3
    /dn-[3456]/ = label-1,label-4
    /dn-5/ = label-5
    ```

  - IP地址范围表示方式

    ```
    10.1.120.[1-6] = label-1
    10.1.120.1 = label-2
    10.1.120.2 = label-3
    10.1.120.[3-6] = label-4
    10.1.120.[4-6] = label-2
    10.1.120.5 = label-5
    10.1.120.6 = label-3
    ```

  - 普通的主机名表达式

    ```
    /dn-1/ = label-1, label-2
    /dn-2/ = label-1, label-3
    /dn-3/ = label-1, label-4
    /dn-4/ = label-1, label-2, label-4
    /dn-5/ = label-1, label-2, label-4, label-5
    /dn-6/ = label-1, label-2, label-3, label-4
    ```

- 目录的标签表达式设置结果如下：

```
/dir1 = label-1
/dir2 = label-1 && label-3
/dir3 = label-2 || label-4[replica=2]
/dir4 = (label-2 || label-3) && label-4
/dir5 = !label-1
/sdir2.txt = label-1 && label-3[replica=3,fallback=NONE]
/dir6 = label-4[replica=2],label-2
```

  说明

  标签表达式设置方式请参考**hdfs nodelabel -setLabelExpression**命令。

  文件的数据块存放结果如下：

  - “/dir1”目录下文件的数据块可存放在dn-1，dn-2，dn-3，dn-4，dn-5和dn-6六个节点中的任意一个。
  - “/dir2”目录下文件的数据块可存放在dn-2和dn-6节点上。Block默认备份数为3，表达式只匹配了两个DataNode节点，第三个副本会在集群上剩余的节点中选择一个DataNode节点存放。
  - “/dir3”目录下文件的数据块可存放在dn-1，dn-3，dn-4，dn-5和dn-6中的任意三个节点上。
  - “/dir4”目录下文件的数据块可存放在dn-4，dn-5和dn-6。
  - “/dir5”目录下文件的数据块没有匹配到任何一个DataNode，会从整个集群中任意选择三个节点存放（和默认选块策略行为一致）。
  - “/sdir2.txt”文件的数据块，两个副本存放在dn-2和dn-6节点上，虽然还缺失一个备份节点，但由于使用了fallback=NONE参数，所以只存放两个备份。
  - “/dir6”目录下文件的数据块在具备label-4的节点中选择2个节点(dn-3 -- dn-6)，然后在label-2中选择一个节点，如果用户指定“/dir6”下文件副本数大于3，则多出来的副本均在label-2。

使用限制

配置文件中，“key”、“value”是以“=”、“：”及空白字符作为分隔的。因此，“key”对应的主机名中间请勿包含以上字符，否则会被误认为分隔符。

12.9.26 配置 HDFS Mover

配置场景

Mover是一个新的数据迁移工具，工作方式与HDFS的Balancer接口工作方式类似。Mover能够基于设置的数据存储策略，将集群中的数据重新分布。

通过运行Mover，周期性地检测HDFS文件系统中用户指定的HDFS文件或目录，判断该文件或目录是否满足设置的存储策略，如果不满足，则进行数据迁移，使目标目录或文件满足设定的存储策略。

说明

本章节适用于MRS 3.x及后续版本。

配置描述

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-215 参数说明

参数	描述	默认值
dfs.mover.auto.enable	是否开启数据副本迁移功能，该功能支持多种。默认值为“false”，表示关闭该特性。	false
dfs.mover.auto.cron.expression	HDFS执行自动数据迁移的CRON表达式，用于控制数据迁移操作的开始时间。仅当dfs.mover.auto.enable设置为true时才有效。默认值“0 * * * *”表示在每个整点执行任务。表达式的具体含义可参见表12-216。	0 * * * *
dfs.mover.auto.hdfsfiles_or_dirs	指定集群执行自动副本迁移的HDFS文件或目录列表，以空格分隔。仅当dfs.mover.auto.enable设置为true时才有效。	-

表 12-216 Cron 表达式解释

列	说明
第1列	分钟，参数值为0~59。
第2列	小时，参数值为0~23。
第3列	日期，参数值为1~31。
第4列	月份，参数值为1~12。
第5列	星期，参数值为0~6，0表示星期日。

使用限制

若要在HDFS的客户端通过命令行执行mover功能，其命令格式如下：

hdfs mover -p *<HDFS文件全路径或目录路径>*

说明

在客户端执行此命令时，用户需要具备supergroup权限。可以使用HDFS服务的系统用户hdfs。或者在集群上创建一个具有supergroup权限的用户，再在客户端中执行此命令。

12.9.27 使用 HDFS AZ Mover

操作场景

AZ Mover是一个副本迁移工具，用来移动副本以满足目录上设置的新AZ策略。它可以用来从一个AZ策略迁移到另一个AZ策略，AZ Mover通过指示NameNode按照新的AZ策略来移动副本，如果NameNode拒绝删除旧副本就不能保证一定能满足新的策略，例如副本被标记为过时等原因。

使用限制

- 将策略更改为LOCAL_AZ与更改为ONE_AZ相同，因为上传文件写入时无法确定写入期间的客户端位置。
- Mover无法确定AZ的状态，因此可能会导致将副本移动到异常的AZ，并依赖NameNode来进一步处理。
- Mover依赖于每个AZ的DataNode节点数达到最小要求，如果在一个DataNode节点数很少的AZ执行，可能会导致与预期不同的结果。
- Mover只满足AZ级别的策略，并不保证满足基本BPP。
- Mover不支持更改复制因子，新旧AZ策略之间的副本计数差异会导致异常结果。

操作步骤

步骤1 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤2 执行以下命令配置环境变量。

source bigdata_env

步骤3 如果集群为安全模式，执行的用户需要源目录或文件读权限，目的目录有写权限，且执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤4 创建目录并设置AZ策略。

执行以下命令创建目录：

hdfs dfs -mkdir <*path*>

执行以下命令设置AZ策略，azexpression代表AZ策略：

hdfs dfsadmin -setAZExpression *<path> <azexpression>*

执行以下命令查看AZ策略：

hdfs dfsadmin -getAZExpression *<path>*

步骤5 在目录中上传文件。

hdfs dfs -put <*localfile***> <***hdfs-path***>**

步骤6 删除目录上的旧策略，再设置一个新的策略。

执行以下命令清楚旧策略：

hdfs dfsadmin -clearAZExpression <*path***>**

执行以下命令设置新策略：

hdfs dfsadmin -setAZExpression <*path***> <***azexpression***>**

步骤7 执行**azmover**命令，使副本分布满足新的AZ 策略。

hdfs azmover -p /targetDirecotry

----结束

12.9.28 配置 HDFS DiskBalancer

配置场景

DiskBalancer是一个在线磁盘均衡器，旨在根据各种指标重新平衡正在运行的DataNode上的磁盘数据。工作方式与HDFS的Balancer工具类似。不同的是，HDFS Balancer工具用于DataNode节点间的数据均衡，而HDFS DiskBalancer用于单个DataNode节点上各磁盘之间的数据均衡。

长时间运行的集群会因为曾经删除过大量的文件，或者集群中的节点做磁盘扩容等操作导致节点上出现磁盘间数据不均衡的现象。磁盘间数据不均衡会引起HDFS整体并发读写性能的下降或者因为不恰当的HDFS写策略导致业务故障。此时需要平衡节点磁盘间的数据密度，防止异构的小磁盘成为该节点的性能瓶颈。

说明

本章节适用于MRS 3.x及后续版本。

配置描述

请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面，在搜索框中输入参数名称。

表 12-217 参数说明

参数	描述	默认值
dfs.disk.balancer.auto.enabled	是否开启自动执行HDFS diskbalancer特性。默认值为“false”，表示关闭该特性。	false
dfs.disk.balancer.auto.cron.expression	HDFS 磁盘均衡操作的CRON表达式，用于控制均衡操作的开始时间。仅当dfs.disk.balancer.auto.enabled设置为true时才有效。默认值“0 1 * * 6”表示在每周六的1点执行任务。表达式的具体含义可参见**表12-218**。默认值表示每周六一点执行。	0 1 * * 6
dfs.disk.balancer.max.disk.throughputInMBperSec	执行磁盘数据均衡时可使用的最大磁盘带宽。单位为MB/s，默认值为10，可依据集群的实际磁盘条件设置。	10
dfs.disk.balancer.max.disk.errors	设置能够容忍的在指定的移动过程中出现的最大错误次数，超过此阈值则移动失败。	5
dfs.disk.balancer.block.tolerance.percent	设置磁盘之间进行数据均衡操作时，各个磁盘的数据存储量与完美状态之间的差异阈值。例如，各个磁盘的理想数据存储量为1TB，此参数设置为10。那么，当目标磁盘的数据存储量达到900GB时，就认为该磁盘的存储状态就已经足够好了。取值范围[1-100]。	10

参数	描述	默认值
dfs.disk.balancer.plan.threshold.percent	设置在磁盘数据均衡中可容忍的两磁盘之间的数据密度域值差。如果任意两个磁盘数据密度差值的绝对值超过了此阈值，意味着对应的磁盘应该进行数据均衡。取值范围[1-100]。	10
dfs.disk.balancer.top.nodes.number	该参数用来指定集群中需要执行磁盘数据均衡的Top N 节点。	5

使用此功能时，需要先将参数dfs.disk.balancer.auto.enabled设置为true，并配置合理的CRON表达式。其它参数依据集群状况设置。

表 12-218 CRON 表达式解释

列	说明
第1列	分钟，参数值为0~59。
第2列	小时，参数值为0~23。
第3列	日期，参数值为1~31。
第4列	月份，参数值为1~12。
第5列	星期，参数值为0~6，0表示星期日。

使用限制

1. 只支持同类型磁盘之间的数据移动，例如SSD->SSD，DISK->DISK等。
2. 执行该特性会占用涉及节点的磁盘IO资源、网络带宽资源，请尽量在业务不繁忙的时候使用。
3. 参数dfs.disk.balancer.top.nodes.number指定Top N 节点返回的DataNode列表是不断重新计算的，因此不必设置的过大。
4. 如果要在HDFS客户端通过命令行使用DiskBalancer功能，其接口如下：

   表 12-219 DiskBalancer 功能的接口说明

命令格式	说明
hdfs diskbalancer -report -top <N>	N 可以指定为大于0的整数，先利用此条命令查询集群中最需要执行磁盘数据均衡的Top N节点。
hdfs diskbalancer -plan <Hostname|IP Address>	此条命令可以根据传入的DN 生成一个Json文件，该文件包含了数据移动的源磁盘、目标磁盘、待移动的块等信息。同时，该命令还支持指定一些其他网络带宽参数等。

命令格式	说明
hdfs diskbalancer -query <Hostname:$dfs.datanode.ipc.port>	集群默认的port值为9867。此条命令可以查询当前节点上运行的DiskBalancer 任务的运行状态。
hdfs diskbalancer -execute <planfile>	此命令中的planfile指的是第二条命令中生成的Json文件，请使用绝对路径。
hdfs diskbalancer -cancel <planfile>	取消正在运行的planfile，同样需要使用绝对路径。

说明

- 在客户端执行此命令时，用户需要具备supergroup权限。可以使用HDFS服务的系统用户hdfs。或者在集群上创建一个具有supergroup权限的用户，再在客户端中执行此命令。
- 表12-219只说明了命令接口的含义及使用方法，实际每个接口提供了更多的配置参数。具体信息可通过"hdfs diskbalancer -help <command>"命令查看。
- 在集群运维过程中，排查性能类问题时。可查看集群的事件信息中是否有HDFS磁盘均衡任务事件发生，如果有的话。可以排查集群中是否开启了DiskBalancer。
- 自动执行磁盘均衡的特性开启以后，会在此次数据均衡执行完成之后才会退出。无法在执行均衡中途取消本次执行任务。
- 如果想要灵活选择某些指定节点进行数据均衡，可以在客户端手动指定执行。

12.9.29 配置从 NameNode 支持读

配置场景

在配置了HA的HDFS集群中，存在一个主NameNode和一个备NameNode。主NameNode处理所有的客户端请求，备NameNode保持最新的元数据信息和块位置信息。但是在这种架构存在一个缺点：主NameNode会成为客户端请求处理的瓶颈，在请求繁忙的集群中表现更为明显。

为了解决主NameNode的瓶颈问题，引入了一个新状态的NameNode：从NameNode。从NameNode类似于备NameNode，也保持着最新的元数据信息和块位置信息。除此之外，从NameNode也可以像主NameNode一样处理客户端的读请求。由于在典型的HDFS集群中，读请求占大多数，因此从NameNode支持读可以降低主NameNode的负载，提高集群处理能力。

说明

本章节适用于MRS 3.x及后续版本。

对系统的影响

- 配置从NameNode支持读可以降低主NameNode的负载，提高HDFS集群的处理能力，尤其是在大集群下效果明显。
- 配置从NameNode支持读需要更新客户端应用配置。

前提条件

- 已安装HDFS集群，主备NameNode正常，HDFS服务正常。
- 规划安装从NameNode的节点已经创建“${BIGDATA_DATA_HOME}/namenode”分区。

操作步骤

以配置hacluster的从NameNode支持读为例来说明，如果集群中有多对NameService，且都在使用，可参考如下步骤为每对NameService配置从NameNode支持读。

步骤1 登录FusionInsight Manager页面。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 管理NameService”。

步骤3 单击hacluster后的“添加”按钮。

步骤4 在添加NameNode页面，“NameNode类型”选择“从”，单击“下一步”。

步骤5 在分配角色页面，选择已规划的主机，添加从NameNode，单击“下一步”。

说明

每对NameService最多可添加5个从NameNode。

步骤6 在配置页面，按照规划配置NameNode的存储目录、端口等信息，单击“下一步”。

步骤7 确认信息无误，单击“提交”，等待从NameNode安装完成。

步骤8 重启依赖HDFS的上层组件，更新客户端应用配置，重启客户端应用。

----结束

12.9.30 使用 HDFS 文件并发操作命令

操作场景

集群内并发修改文件和目录的权限及访问控制的工具。

说明

本章节适用于MRS 3.x及后续版本。

对系统的影响

因为集群内使用文件并发修改命令会对集群性能造成较大负担，所以在集群空闲时使用文件并发操作命令。

前提条件

- 已安装HDFS客户端或者包括HDFS的客户端。例如安装目录为“/opt/client”。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码（普通模式不涉及）。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行的用户所属的用户组必须为**supergroup**组，且执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 增大客户端的JVM大小，防止OOM，方法如下。（1亿文件建议**32G**）

说明

若执行HDFS客户端命令时，客户端程序异常退出，并且报“java.lang.OutOfMomeryError”错误。

这个问题是由于HDFS客户端运行时的所需的内存超过了HDFS客户端设置的内存上限（默认128M）。可通过修改“*<客户端安装路径>*/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改HDFS客户端的内存上限。例如，需要设置内存上限为1GB，则设置：

CLIENT_GC_OPTS="-Xmx1G"

在修改完后，使用如下命令刷新客户端配置，使之生效;

source *<客户端安装路径>***/bigdata_env**

步骤6 直接执行并发命令，命令详情如下表。

命令	参数及说明	命令作用
hdfs quickcmds [-t threadsNumber] [-p principal] [-k keytab] -setrep <rep> <path> ...	threadsNumber：并发线程数，默认为本机CPU核数 principal：Kerberos用户 keytab：Keytab文件 rep：副本数 path：HDFS目录	多并发设置目录中所有文件的副本数
hdfs quickcmds [-t threadsNumber] [-p principal] [-k keytab] -chown [owner][:[group]] <path> ...	threadsNumber：并发线程数，默认为本机CPU核数 principal：Kerberos用户 keytab：Keytab文件 owner：所属用户 group：所属组 path：HDFS目录	多并发设置目录中所有文件的属组

命令	参数及说明	命令作用
hdfs quickcmds [-t threadsNumber] [-p principal] [-k keytab] -chmod <mode> <path> ...	threadsNumber：并发线程数，默认为本机CPU核数 principal：Kerberos用户 keytab：Keytab文件 mode：权限（如754） path：HDFS目录	多并发设置目录中所有文件的权限
hdfs quickcmds [-t threadsNumber] [-p principal] [-k keytab] -setfacl [{-b|-k} {-m|-x <acl_spec>} <path> ...]|[--set <acl_spec> <path> ...]	threadsNumber：并发线程数，默认为本机CPU核数 principal：Kerberos用户 keytab：Keytab文件 acl_spec：逗号分隔的ACL列表 path：HDFS目录	多并发设置目录中所有文件的ACL信息

----结束

12.9.31 HDFS 日志介绍

日志描述

日志存储路径：HDFS相关日志的默认存储路径为“/var/log/Bigdata/hdfs/角色名”

- NameNode：“/var/log/Bigdata/hdfs/nn”（运行日志），“/var/log/Bigdata/audit/hdfs/nn”（审计日志）。
- DataNode：“/var/log/Bigdata/hdfs/dn”（运行日志），“/var/log/Bigdata/audit/hdfs/dn”（审计日志）。
- ZKFC：“/var/log/Bigdata/hdfs/zkfc”（运行日志），“/var/log/Bigdata/audit/hdfs/zkfc”（审计日志）。
- JournalNode：“/var/log/Bigdata/hdfs/jn”（运行日志），“/var/log/Bigdata/audit/hdfs/jn”（审计日志）。
- Router：“/var/log/Bigdata/hdfs/router”（运行日志），“/var/log/Bigdata/audit/hdfs/router”（审计日志）。
- HttpFS：“/var/log/Bigdata/hdfs/httpfs”（运行日志），“/var/log/Bigdata/audit/hdfs/httpfs”（审计日志）。

日志归档规则：HDFS的日志启动了自动压缩归档功能，默认情况下，当日志大小超过100MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的100个压缩文件，压缩文件保留个数可以在Manager界面中配置。

表 12-220 HDFS 日志列表

日志类型	日志文件名	描述
运行日志	hadoop-<SSH_USER>-<process_name>-<hostname>.log	HDFS系统日志，记录HDFS系统运行时候所产生的大部分日志。
	hadoop-<SSH_USER>-<process_name>-<hostname>.out	HDFS运行环境信息日志。
	hadoop.log	Hadoop客户端操作日志。
	hdfs-period-check.log	周期运行的脚本的日志记录。包括：自动均衡、数据迁移、journalnode数据同步检测等。
	<process_name>-<SSH_USER>-<DATE>-<PID>-gc.log	垃圾回收日志。
	postinstallDetail.log	HDFS服务安装后启动前工作日志。
	hdfs-service-check.log	HDFS服务启动是否成功的检查日志。
	hdfs-set-storage-policy.log	HDFS数据存储策略日志。
	cleanupDetail.log	HDFS服务卸载时候的清理日志。
	prestartDetail.log	HDFS服务启动前集群操作的记录日志。
	hdfs-recover-fsimage.log	NameNode元数据恢复日志。
	datanode-disk-check.log	集群安装过程和使用过程中磁盘状态检测的记录日志。
	hdfs-availability-check.log	HDFS服务是否可用日志。
	hdfs-backup-fsimage.log	NameNode元数据备份日志。
	startDetail.log	hdfs服务启动的详细日志。
	hdfs-blockplacement.log	HDFS块放置策略记录日志。
	upgradeDetail.log	升级日志。

日志类型	日志文件名	描述
	hdfs-clean-acls-java.log	HDFS清除已删除角色的ACL信息的日志。
	hdfs-haCheck.log	NameNode主备状态获取脚本运行日志。
	<process_name>-jvmpause.log	进程运行中，记录JVM停顿的日志。
	hadoop-<SSH_USER>-balancer-<hostname>.log	HDFS自动均衡的运行日志。
	hadoop-<SSH_USER>-balancer-<hostname>.out	HDFS运行自动均衡的环境信息日志。
	hdfs-switch-namenode.log	HDFS主备倒换运行日志
	hdfs-router-admin.log	管理挂载表操作的运行日志
Tomcat日志	hadoop-omm-host1.out，httpfs-catalina.<DATE>.log，httpfs-host-manager.<DATE>.log，httpfs-localhost.<DATE>.log，httpfs-manager.<DATE>.log，localhost_access_web_log.log	tomcat运行日志
审计日志	hdfs-audit-<process_name>.log ranger-plugin-audit.log	HDFS操作审计日志（例如：文件增删改查）。
	SecurityAuth.audit	HDFS安全审计日志。

日志级别

HDFS中提供了如表12-221所示的日志级别，日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-221 日志级别

级别	描述
FATAL	FATAL表示系统运行的致命错误信息。
ERROR	ERROR表示系统运行的错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示系统及各事件正常运行状态信息。

级别	描述
DEBUG	DEBUG表示系统及系统调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 请参考**修改集群服务配置参数**，进入HDFS的“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后立即生效，不需要重启服务。

----结束

日志格式

HDFS的日志格式如下所示：

表 12-222 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2015-01-26 18:43:42,840 | INFO | IPC Server handler 40 on 8020 | Rolling edit logs | org.apache.hadoop.hdfs.server.namenode.FSEditLog.rollEditLog(FSEditLog.java:1096)

日志类型	格式	示例
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2015-01-26 18:44:42,607 | INFO | IPC Server handler 32 on 8020 | allowed=true ugi=hbase (auth:SIMPLE) ip=/10.177.112.145 cmd=getfileinfo src=/hbase/WALs/hghoulaslx410,16020,1421743096083/hghoulaslx410%2C16020%2C1421743096083.1422268722795 dst=null perm=null | org.apache.hadoop.hdfs.server.namenode.FSNamesystem $DefaultAuditLogger.logAuditMessage(FSNamesystem.java:7950)

12.9.32 HDFS 性能调优

12.9.32.1 提升写性能

操作场景

在HDFS中，通过调整属性的值，使得HDFS集群更适应自身的业务情况，从而提升HDFS的写性能。

说明

本章节适用于MRS 3.x及后续版本。

操作步骤

参数入口：

在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，选择“全部配置”。在搜索框中输入参数名称。

表 12-223 HDFS 写性能优化配置

参数	描述	默认值
dfs.datanode.drop.cache.behind.reads	表示是否让DataNode在将缓冲区中的数据传递给客户端后自动清除缓冲区中的所有数据。 设置为true表示丢弃缓存的数据（需要在DataNode中配置）。 当同一份数据，重复读取的次数较少时，建议设置为true，使得缓存能够被其他操作使用。重复读取的次数较多时，设置为false能够提升重复读取的速度。	false
dfs.client-write-packet-size	客户端写包的大小。当HDFS Client往DataNode写数据时，将数据生成一个包。然后将这个包在网络上传出。此参数指定传输数据包的大小，可以通过各Job来指定。单位：字节。 在万兆网部署下，可适当增大该参数值，来提升传输的吞吐量。	262144

12.9.32.2 使用客户端元数据缓存提高读取性能

操作场景

通过使用客户端缓存元数据块的位置来提高HDFS读取性能。

说明

此功能仅用于读取不经常修改的文件。因为在服务器端由某些其他客户端完成的数据修改，对于高速缓存的客户端将是不可见的，这可能导致从缓存中拿到的元数据是过期的。

本章节适用于MRS 3.x及后续版本。

操作步骤

设置参数的路径：

在FusionInsight Manager页面中，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，选择“全部配置”，并在搜索框中输入参数名称。

表 12-224 参数配置

参数	描述	默认值
dfs.client.metadata.cache.enabled	启用/禁用块位置元数据的客户端缓存。将此参数设置为“true”，搭配“dfs.client.metadata.cache.pattern”参数以启用缓存。	false

参数	描述	默认值
dfs.client.metadata.cache.pattern	需要缓存的文件路径的正则表达式模式。只有这些文件的块位置元数据被缓存，直到这些元数据过期。此配置仅在参数“dfs.client.metadata.cache.enabled”设置为“true”时有效。 示例：“/test.*”表示读取其路径是以“/test”开头的所有文件。 **说明** • 为确保一致性，配置特定模式以仅缓存其他客户端不经常修改的文件。 • 正则表达式模式将仅验证URI的path部分，而不验证在Fully Qualified路径情况下的schema和authority。	-
dfs.client.metadata.cache.expiry.sec	缓存元数据的持续时间。缓存条目在该持续时间过期后失效。即使在缓存过程中经常使用的元数据也会发生失效。 配置值可采用时间后缀s/m/h表示，分别表示秒，分钟和小时。 **说明** 若将该参数配置为“0s”，将禁用缓存功能。	60s
dfs.client.metadata.cache.max.entries	缓存一次最多可保存的非过期数据条目。	65536

说明

要在过期前完全清除客户端缓存，可调用*DFSClient#clearLocatedBlockCache()*。

用法如下所示。

```
FileSystem fs = FileSystem.get(conf);
DistributedFileSystem dfs = (DistributedFileSystem) fs;
DFSClient dfsClient = dfs.getClient();
dfsClient.clearLocatedBlockCache();
```

12.9.32.3 使用当前活动缓存提升客户端与 NameNode 的连接性能

操作场景

HDFS部署在具有多个NameNode实例的HA（High Availability）模式中，HDFS客户端需要依次连接到每个NameNode，以确定当前活动的NameNode是什么，并将其用于客户端操作。

一旦识别出来，当前活动的NameNode的详细信息就可以被缓存并共享给在客户端机器中运行的所有客户端。这样，每个新客户端可以首先尝试从缓存加载活动的NameNode的详细信息，并将RPC调用保存到备用的NameNode。在异常情况下有很多优势，例如当备用的NameNode连接长时间不响应时。

当发生故障，将另一个NameNode切换为活动状态时，缓存的详细信息将被更新为当前活动的NameNode的信息。

说明

本章节适用于MRS 3.x及后续版本。

操作步骤

设置参数的路径如下：

在FusionInsight Manager页面中，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，选择“全部配置”，并在搜索框中输入参数名称。

表 12-225 配置参数

参数	描述	默认值
dfs.client.failover.proxy.provider.[nameservice ID]	用已通过的协议创建namenode代理的Client Failover proxy provider类。配置成org.apache.hadoop.hdfs.server.namenode.ha.BlackListingFailoverProxyProvider，可在HDFS客户端使用NameNode黑名单特性。配置成org.apache.hadoop.hdfs.server.namenode.ha.ObserverReadProxyProvider，可使用从NameNode支持读的特性。	org.apache.hadoop.hdfs.server.namenode.ha.AdaptiveFailoverProxyProvider
dfs.client.failover.activeinfo.share.flag	启用缓存并将当前活动的NameNode的详细信息共享给其他客户端。若要启用缓存，需将其设置为“true”。	false
dfs.client.failover.activeinfo.share.path	指定将在机器中的所有客户端创建的共享文件的本地目录。如果要为不同用户共享缓存，该文件夹应具有必需的权限（如在给定目录中创建，读写缓存文件）。	/tmp
dfs.client.failover.activeinfo.share.io.timeout.sec	控制超时的可选配置。用于在读取或写入缓存文件时获取锁定。如果在该时间内无法获取缓存文件上的锁定，则放弃尝试读取或更新缓存。单位为秒。	5

说明

由HDFS客户端创建的缓存文件必须由其他客户端重新使用。因此，这些文件永远不会从本地系统中删除。若禁用该功能，可能需要进行手动清理。

12.9.33 HDFS 常见问题

12.9.33.1 NameNode 启动慢

问题

删除大量文件之后立刻重启NameNode（例如删除100万个文件），NameNode启动慢。

回答

由于在删除了大量文件之后，DataNode需要时间去删除对应的Block。当立刻重启NameNode时，NameNode会去检查所有DataNode上报的Block信息，发现已删除的Block时，会输出对应的INFO日志信息，如下所示：

```
2015-06-10 19:25:50,215 | INFO  | IPC Server handler 36 on 25000 | BLOCK* processReport:
blk_1075861877_2121067 on node 10.91.8.218:9866 size 10249 does not belong to any file |
org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.processReport(BlockManager.java:1854)
```

每一个被删除的Block会产生一条日志信息，一个文件可能会存在一个或多个Block。当删除的文件数过多时，NameNode会花大量的时间打印日志，然后导致NameNode启动慢。

当出现这种现象时，您可以通过如下方式提升NameNode的启动速度。

1. 删除大量文件时，不要立刻重启NameNode，待DataNode删除了对应的Block后重启NameNode，即不会存在这种情况。

   您可以通过***hdfs dfsadmin -report***命令来查看磁盘空间，检查文件是否删除完毕。

2. 如已大量出现以上日志，您可以将NameNode的日志级别修改为ERROR，NameNode不会再打印此日志信息。

   等待NameNode启动完毕后，再将此日志级别修改为INFO。修改日志级别后无需重启服务。

12.9.33.2 DataNode 状态正常，但无法正常上报数据块

问题

DataNode正常，但无法正常上报数据块，导致存在的数据块无法使用。

回答

当某个数据目录中的数据块数量超过4倍的数据块限定值(1M)时，可能会出现该错误。DataNode会产生相应的错误日志记录，如下所示：

```
2015-11-05 10:26:32,936 | ERROR | DataNode:[[[DISK]file:/srv/BigData/hadoop/data1/dn/]] heartbeating to
vm-210/10.91.8.210:8020 | Exception in BPOfferService for Block pool
BP-805114975-10.91.8.210-1446519981645
(Datanode Uuid bcada350-0231-413b-bac0-8c65e906c1bb) service to vm-210/10.91.8.210:8020 |
BPServiceActor.java:824
java.lang.IllegalStateException:com.google.protobuf.InvalidProtocolBufferException:Protocol message was
too large.May
be malicious.Use CodedInputStream.setSizeLimit() to increase the size limit. at
org.apache.hadoop.hdfs.protocol.BlockListAsLongs$BufferDecoder$1.next(BlockListAsLongs.java:369)
at org.apache.hadoop.hdfs.protocol.BlockListAsLongs$BufferDecoder$1.next(BlockListAsLongs.java:347) at
org.apache.hadoop.hdfs.
protocol.BlockListAsLongs$BufferDecoder.getBlockListAsLongs(BlockListAsLongs.java:325) at
org.apache.hadoop.hdfs.protocolPB.DatanodeProtocolClientSideTranslatorPB.
blockReport(DatanodeProtocolClientSideTranslatorPB.java:190) at
org.apache.hadoop.hdfs.server.datanode.BPServiceActor.blockReport(BPServiceActor.java:473)
at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.offerService(BPServiceActor.java:685) at
org.apache.hadoop.hdfs.server.datanode.BPServiceActor.run(BPServiceActor.java:822)
at java.lang.Thread.run(Thread.java:745) Caused
by:com.google.protobuf.InvalidProtocolBufferException:Protocol message was too large.May be
malicious.Use CodedInputStream.setSizeLimit()
to increase the size limit. at
com.google.protobuf.InvalidProtocolBufferException.sizeLimitExceeded(InvalidProtocolBufferException.java:1
10) at com.google.protobuf.CodedInputStream.refillBuffer(CodedInputStream.java:755)
at com.google.protobuf.CodedInputStream.readRawByte(CodedInputStream.java:769) at
```

```
com.google.protobuf.CodedInputStream.readRawVarint64(CodedInputStream.java:462) at
com.google.protobuf.
CodedInputStream.readSInt64(CodedInputStream.java:363) at
org.apache.hadoop.hdfs.protocol.BlockListAsLongs$BufferDecoder$1.next(BlockListAsLongs.java:363)
```

如今，数据目录中数据块的数量会显示为Metric。用户可以通过以下URL对该值进行监视http://<datanode-ip>:<http-port>/jmx，如果该值超过4倍的限定值(4*1M)，建议用户配置多个驱动器并重新启动HDFS。

恢复步骤：

1. 在DataNode上配置多个数据目录。

   示例：在原先只配置了/data1/datadir的位置

   ```
   <property> <name>dfs.datanode.data.dir</name> <value>/data1/datadir</value> </property>
   ```

   按照如下内容进行配置。

   ```
   <property> <name>dfs.datanode.data.dir</name> <value>/data1/datadir/,/data2/datadir,/data3/
   datadir</value> </property>
   ```

   说明

   建议多个数据目录应该配置到多个磁盘中，否则所有的数据都将写入同一个磁盘，对性能有很大的影响。

2. 重新启动HDFS。
3. 按照如下方法将数据移动至新的数据目录。

   mv */data1/datadir/current/finalized/subdir1 /data2/datadir/current/finalized/subdir1*

4. 重新启动HDFS。

12.9.33.3 HDFS Web UI 无法正常刷新损坏数据的信息

问题

1. 当DataNode的“dfs.datanode.data.dir”所配置的目录因权限或者磁盘损坏发生错误时，HDFS Web UI没有显示损坏数据的信息。
2. 当此错误被修复后，HDFS Web UI没有及时移除损坏数据的相关信息。

回答

1. DataNode只有在执行文件操作发生错误时，才会去检查磁盘是否正常，若发现数据损坏，则将此错误上报至NameNode，此时NameNode才会在HDFS Web UI显示数据损坏信息。
2. 当错误修复后，需要重启DataNode。当重启DataNode时，会检查所有数据状态并上传损坏数据信息至NameNode。所以当此错误被修复后，只有重启DataNode后，才会不显示损坏数据信息。

12.9.33.4 distcp 命令在安全集群上失败并抛出异常

问题

为何distcp命令在安全集群上失败并抛出异常?

客户端出现异常：

```
Invalid arguments:Unexpected end of file from server
```

服务器端出现异常：

```
javax.net.ssl.SSLException:Unrecognized SSL message, plaintext connection?
```

回答

当用户在distcp命令中使用**webhdfs://**时，会抛出上述异常，是由于集群所使用的HTTP政策为HTTPS，即配置在“core-site.xml”的“dfs.http.policy”值为“HTTPS_ONLY”。所以要避免出现此异常，应使用**swebhdfs://**替代**webhdfs://**。

例如：

./hadoop distcp swebhdfs://IP:PORT/testfile hdfs://IP:PORT/testfile1

12.9.33.5 当 dfs.datanode.data.dir 中定义的磁盘数量等于 dfs.datanode.failed.volumes.tolerated 的值时，DataNode 启动失败

问题

当“dfs.datanode.data.dir”中定义的磁盘数量等于“dfs.datanode.failed.volumes.tolerated”的值时，DataNode启动失败。

回答

默认情况下，单个磁盘的故障将会引起HDFS DataNode进程关闭，导致NameNode为每一个存在DataNode上的block调度额外的副本，在没有故障的磁盘中引起不必要的块复制。

为了防止此情况，用户可以通过配置DataNodes来承受dfs.data.dir目录的故障。在“hdfs-site.xml”中配置参数“dfs.datanode.failed.volumes.tolerated”。例如：如果该参数值为3，DataNode只有在4个或者更多个目录故障之后才会出现故障。该值会影响到DataNode的启动。

如果想要DataNode不出现故障，配置的“dfs.datanode.failed.volumes.tolerated”一定要小于所配置的卷数，也可以将“dfs.datanode.failed.volumes.tolerated”设置成-1，相当于设置该值为n-1（n为卷数），那样DataNode就不会出现启动失败。

12.9.33.6 当多个 data.dir 被配置在一个磁盘分区内，DataNode 的容量计算将会出错

问题

当多个data.dir被配置在一个磁盘分区内，DataNode的容量计算将会出错。

回答

目前容量计算是基于磁盘的，类似于Linux里面的***df***命令。理想状态下，用户不会在同一个磁盘内配置多个data.dir，否则所有的数据都将写入一个磁盘，在性能上会有很大的影响。

因此配置如下：

例如，如果机器有如下磁盘：

```
host-4:~ # df -h
Filesystem      Size  Used Avail Use% Mounted on
```

```
/dev/sda1       352G   11G     324G    4%     /
udev            190G   252K    190G    1%     /dev
tmpfs           190G   72K     190G    1%     /dev/shm
/dev/sdb1       2.7T   74G     2.5T    3%     /data1
/dev/sdc1       2.7T   75G     2.5T    3%     /data2
/dev/sdd1       2.7T   73G     2.5T    3%     /da
```

建议的配置方式：

```
<property>
<name>dfs.datanode.data.dir</name>
<value>/data1/datadir/,/data2/datadir,/data3/datadir</value>
</property>
```

不建议的配置方式：

```
<property>
<name>dfs.datanode.data.dir</name>
<value>/data1/datadir1/,/data2/datadir1,/data3/datadir1,/data1/datadir2/data1/datadir3,/data2/datadir2,/
data2/datadir3,/data3/datadir2,/data3/datadir3</value>
</property>
```

12.9.33.7 当 Standby NameNode 存储元数据（命名空间）时，出现断电的情况，Standby NameNode 启动失败

问题

当Standby NameNode存储元数据（命名空间）时，出现断电的情况，Standby NameNode启动失败并抛出如下错误信息。

```
2015-12-04 11:49:12,121 | ERROR | main | Failed to load image from FS
ImageFile(file=/srv/BigData/namenode/current/fsimage_0000000000000096
080,
cpktTxId=0000000000000096080) | FSImage.java:685
java.io.IOException: Invalid MD5 file /srv/BigData/namenode/current/f
simage_0000000000000096080.md5:
the content "锟斤拷锟斤拷锟斤拷锟斤拷锟斤拷锟[1m^A!锟 does not match the expecte
d pattern.
at org.apache.hadoop.hdfs.util.MD5FileUtils.readStoredMd5(MD5FileUtil
s.java:92)
at org.apache.hadoop.hdfs.util.MD5FileUtils.readStoredMd5ForFile(MD5F
ileUtils.java:109)
at org.apache.hadoop.hdfs.server.namenode.FSImage.loadFSImage(FSImage
.java:975)
at org.apache.hadoop.hdfs.server.namenode.FSImage.loadFSImageFile(FSI
mage.java:744)
at org.apache.hadoop.hdfs.server.namenode.FSImage.loadFSImage(FSImage
.java:682)
at org.apache.hadoop.hdfs.server.namenode.FSImage.recoverTransitionRe
ad(FSImage.java:300)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.loadFSImage(FS
Namesystem.java:968)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.loadFromDisk(F
SNamesystem.java:675)
at org.apache.hadoop.hdfs.server.namenode.NameNode.loadNamesystem(Nam
eNode.java:625)
at org.apache.hadoop.hdfs.server.namenode.NameNode.initialize(NameNod
e.java:685)
at org.apache.hadoop.hdfs.server.namenode.NameNode.<init>(NameNode.ja
va:889)
at org.apache.hadoop.hdfs.server.namenode.NameNode.<init>(NameNode.ja
va:872)
at org.apache.hadoop.hdfs.server.namenode.NameNode.createNameNode(Nam
eNode.java:1580)
at org.apache.hadoop.hdfs.server.namenode.NameNode.main(NameNode.java
:1654)
```

回答

当Standby NameNode存储元数据（命名空间）时，出现断电的情况，Standby NameNode启动失败，MD5文件会损坏。通过移除损坏的fsimage，然后启动Standby NameNode，可以修复此问题。Standby NameNode会加载先前的fsimage并重现所有的edits。

修复步骤：

1. 移除损坏的fsimage。

   rm -rf ${BIGDATA_DATA_HOME}/namenode/current/fsimage_0000000000000096

2. 启动Standby NameNode。

12.9.33.8 在存储小文件过程中，系统断电，缓存中的数据丢失

问题

在存储小文件过程中，系统断电，缓存中的数据丢失。

回答

由于断电，当写操作完成之后，缓存中的block不会立即被写入磁盘，如果要同步地将缓存的block写入磁盘，用户需要将“hdfs-site.xml”中的“dfs.datanode.synconclose”设置为“true”。

默认情况下，“dfs.datanode.synconclose”为“false”，虽然性能很高，但是断电之后，存储在缓存中的数据会丢失。将“dfs.datanode.synconclose”设置为“true”，可以解决此问题，但对性能有很大影响。请根据具体的应用场景决定是否开启该参数。

12.9.33.9 FileInputFormat split 的时候出现数组越界

问题

HDFS调用FileInputFormat的getSplit方法的时候，出现ArrayIndexOutOfBoundsException: 0，日志如下：

```
java.lang.ArrayIndexOutOfBoundsException: 0
at org.apache.hadoop.mapred.FileInputFormat.identifyHosts(FileInputFormat.java:708)
at org.apache.hadoop.mapred.FileInputFormat.getSplitHostsAndCachedHosts(FileInputFormat.java:675)
at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:359)
at org.apache.spark.rdd.HadoopRDD.getPartitions(HadoopRDD.scala:210)
at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:239)
at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:237)
at scala.Option.getOrElse(Option.scala:120)
at org.apache.spark.rdd.RDD.partitions(RDD.scala:237)
at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:35)
```

回答

每个block对应的机架信息组成为：/default/rack0/:,/default/rack0/datanodeip:port。

该问题是由于某个block块损坏或者丢失，导致该block对应的机器ip和port为空引起的，出现该问题的时候使用**hdfs fsck**检查对应文件块的健康状态，删除损坏或者恢复丢失的块，重新进行任务计算即可。

12.9.33.10 当分级存储策略为 LAZY_PERSIST 时，为什么文件的副本的存储类型都是 DISK

问题

当文件的存储策略为LAZY_PERSIST时，文件的第一副本的存储类型应为RAM_DISK，其余副本为DISK。

为什么文件的所有副本的存储类型都是DISK?

回答

当用户写入存储策略为LAZY_PERSIST的文件时，文件的三个副本会逐一写入。第一副本会优先选择客户端所在的DataNode节点，在以下情况下，当文件的存储策略为LAZY_PERSIST时，文件的所有副本的存储类型都是DISK：

- 当客户端所在的DataNode节点没有RAM_DISK时，则会写入客户端所在的DataNode节点的DISK磁盘，其余副本会写入其他节点的DISK磁盘。
- 当客户端所在的DataNode节点有RAM_DISK，但"dfs.datanode.max.locked.memory"参数值未设置或设置过小（小于“dfs.blocksize”参数值）时，则会写入客户端所在的DataNode节点的DISK磁盘，其余副本会写入其他节点的DISK磁盘。

12.9.33.11 NameNode 节点长时间满负载，HDFS 客户端无响应

问题

当NameNode节点处于满负载、NameNode所在节点的CPU 100%耗尽时，导致NameNode无法响应，对于新连接到该NameNode的HDFS客户端，能够主备切换连接到另一个NameNode，进行正常的操作，而对于已经连接到该NameNode节点的HDFS客户端可能会卡住，无法进行下一步操作。

回答

目前出现上述问题时使用的是默认配置，如**表12-226**所示，HDFS客户端到NameNode的RPC连接存在keep alive机制，保持连接不会超时，尽力等待服务器的响应，因此导致已经连接的HDFS客户端的操作会卡住。

对于已经卡住的HDFS客户端，可以进行如下操作：

- 等待NameNode响应，一旦NameNode所在节点的CPU利用率回落，NameNode可以重新获得CPU资源时，HDFS客户端即可得到响应。
- 如果无法等待更长时间，需要重启HDFS客户端所在的应用程序进程，使得HDFS客户端重新连接空闲的NameNode。

解决措施：

为了避免该问题出现，可以在客户端的配置文件“core-site.xml”中做如下配置。

表 12-226 参数说明

参数	描述	默认值
ipc.client.ping	当配置为true时，客户端会尽力等待服务端响应，定期发送ping消息，使得连接不会因为tcp timeout而断开。 当配置为false时，客户端会使用配置项“ipc.ping.interval”对应的值，作为timeout时间，在该时间内没有得到响应，即会超时。 在上述问题场景下，建议配置为false。	true
ipc.ping.interval	当“ipc.client.ping”配置为true时，表示发送ping消息的周期。 当“ipc.client.ping”设置为false时，表示连接的超时时间。 在上述问题场景下，建议配置一个较大的超时时间，避免服务繁忙时的超时，建议配置为900000，单位为ms。	60000

12.9.33.12 DataNode 禁止手动删除或修改数据存储目录

问题

- 数据块在DataNode上的存储目录由“dfs.datanode.data.dir”配置项指定，是否可以修改该配置项来修改数据存储目录?
- 是否可以手动拷贝数据存储目录下的文件?

回答

“dfs.datanode.data.dir”配置项用于指定数据块在DataNode上的存储目录，在系统安装时需要指定根目录，并且可以指定多个根目录。

- 请谨慎修改该配置项，可以添加新的数据根目录。
- 禁止删除原有存储目录，否则会造成数据块丢失，导致文件无法正常读写。
- 禁止手动删除或修改存储目录下的数据块，否则可能会造成数据块丢失。

说明

NameNode和JournalNode存在类似的配置项，也同样禁止删除原有存储目录，禁止手动删除或修改存储目录下的数据块。

- dfs.namenode.edits.dir
- dfs.namenode.name.dir
- dfs.journalnode.edits.dir

12.9.33.13 成功回滚后，为什么 NameNode UI 上显示有一些块缺失

问题

回滚成功后，为什么NameNode UI上显示有一些块缺失?

回答

原因：具有新id/genstamps的块可能存在于DataNode上。DataNode中的块文件可能具有与NameNode的回滚image中不同的生成标记和长度，所以NameNode会拒绝DataNode中的这些块，并将文件标记为已损坏。

场景如下：

1. 升级前
   客户端A - >将一些数据写入文件X（假设已写入“A”字节）
2. 升级开始了
   客户端A - >仍然将数据写入文件X（现在文件中的数据是“A + B”字节）
3. 升级完成
   客户端A - >完成写入文件。最终数据为“A + B”字节。
4. 回滚开始
   将回滚到步骤1（升级前）的状态。因此，NameNode中的文件X将具有“A”字节，但DataNode中的块文件将具有“A + B”字节。

恢复步骤：

1. 从NameNode Web UI中获取已损坏的文件列表，或者通过下面的命令获取。
   hdfs fsck <filepath> -list-corruptfileblocks
2. 对于不需要的文件，请使用以下命令删除文件。
   hdfs fsck <corrupt file path> - delete

   说明

   删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。
3. 对于所需的文件，执行fsck命令来获取块列表和块的顺序。
   - 在fsck中给出的块序列表中，使用块id搜索DataNode中的数据目录，并从DataNode下载相应的块。
   - 按照序列以追加的方式写入所有这样的块文件，并构造成原始文件。
     例如：
     File 1--> blk_1, blk_2, blk_3
     通过组合来自同一序列的所有三个块文件的内容来创建文件。
   - 从HDFS中删除旧文件并重写新构建的文件。

12.9.33.14 为什么在往 HDFS 写数据时报"java.net.SocketException: No buffer space available"异常

问题

为什么在往HDFS写数据时报"java.net.SocketException: No buffer space available"异常?

这个问题发生在往HDFS写文件时。查看客户端和DataNode的错误日志。

客户端日志如下：

图 12-23 客户端日志

DataNode日志如下：

2017-07-24 20:43:39,269 | ERROR | DataXceiver for client DFSClient_NONMAPREDUCE_996005058_86 at /192.168.164.155:40214 [Receiving block BP-1287143557-192.168.199.6-1500707719940:blk_1074269754_528941 with io weight 10] | DataNode{data=FSDataset{dirpath='[/srv/BigData/hadoop/data1/dn/current, /srv/BigData/hadoop/data2/dn/current, /srv/BigData/hadoop/data3/dn/current, /srv/BigData/hadoop/data4/dn/current, /srv/BigData/hadoop/data5/dn/current, /srv/BigData/hadoop/data6/dn/current, /srv/BigData/hadoop/data7/dn/current]'}, localName='192-168-164-155:9866', datanodeUuid='a013e29c-4e72-400c-bc7b-bbbf0799604c', xmitsInProgress=0}:Exception transfering block BP-1287143557-192.168.199.6-1500707719940:blk_1074269754_528941 to mirror 192.168.202.99:9866: java.net.SocketException: No buffer space available | DataXceiver.java:870
2017-07-24 20:43:39,269 | INFO | DataXceiver for client DFSClient_NONMAPREDUCE_996005058_86 at /192.168.164.155:40214 [Receiving block BP-1287143557-192.168.199.6-1500707719940:blk_1074269754_528941 with io weight 10] | opWriteBlock BP-1287143557-192.168.199.6-1500707719940:blk_1074269754_528941 received exception java.net.SocketException: No buffer space available | DataXceiver.java:933
2017-07-24 20:43:39,270 | ERROR | DataXceiver for client DFSClient_NONMAPREDUCE_996005058_86 at /192.168.164.155:40214 [Receiving block BP-1287143557-192.168.199.6-1500707719940:blk_1074269754_528941 with io weight 10] | 192-168-164-155:9866:DataXceiver error processing WRITE_BLOCK operation src: /192.168.164.155:40214 dst: /192.168.164.155:9866 | DataXceiver.java:304 **java.net.SocketException: No buffer space available**
at sun.nio.ch.Net.connect0(Native Method)
at sun.nio.ch.Net.connect(Net.java:454)
at sun.nio.ch.Net.connect(Net.java:446)
at sun.nio.ch.SocketChannelImpl.connect(SocketChannelImpl.java:648)
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:192)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:531)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:495)
at org.apache.hadoop.hdfs.server.datanode.DataXceiver.writeBlock(DataXceiver.java:800)
at org.apache.hadoop.hdfs.protocol.datatransfer.Receiver.opWriteBlock(Receiver.java:138)
at org.apache.hadoop.hdfs.protocol.datatransfer.Receiver.processOp(Receiver.java:74)
at org.apache.hadoop.hdfs.server.datanode.DataXceiver.run(DataXceiver.java:265)
at java.lang.Thread.run(Thread.java:748)

回答

上述问题可能是因为网络内存枯竭而导致的。

问题的解决方案是根据实际场景适当增大网络设备的阈值级别。

例如：

```
[root@xxxxx ~]# cat /proc/sys/net/ipv4/neigh/default/gc_thresh*
128
```

```
512
1024
[root@xxxxx ~]# echo 512 > /proc/sys/net/ipv4/neigh/default/gc_thresh1
[root@xxxxx ~]# echo 2048 > /proc/sys/net/ipv4/neigh/default/gc_thresh2
[root@xxxxx ~]# echo 4096 > /proc/sys/net/ipv4/neigh/default/gc_thresh3
[root@xxxxx ~]# cat /proc/sys/net/ipv4/neigh/default/gc_thresh*
512
2048
4096
```

还可以将以下参数添加到“/etc/sysctl.conf”中，即使主机重启，配置依然能生效。

```
net.ipv4.neigh.default.gc_thresh1 = 512
net.ipv4.neigh.default.gc_thresh2 = 2048
net.ipv4.neigh.default.gc_thresh3 = 4096
```

12.9.33.15 为什么主 NameNode 重启后系统出现双备现象

问题

为什么主NameNode重启后系统出现双备现象?

出现该问题时，查看Zookeeper和ZKFC的日志，发现Zookeeper服务端与客户端（ZKFC）通信时所使用的session不一致，Zookeeper服务端的sessionId为0x164cb2b3e4b36ae4，ZKFC的sessionId为0x144cb2b3e4b36ae4。这意味着Zookeeper服务端与客户端（ZKFC）之间数据交互失败。

Zookeeper日志，如下所示：

```
2015-04-15 21:24:54,257 | INFO | CommitProcessor:22 | Established session 0x164cb2b3e4b36ae4 with negotiated timeout 45000 for client /192.168.0.117:44586 | org.apache.zookeeper.server.ZooKeeperServer.finishSessionInit(ZooKeeperServer.java:623)
2015-04-15 21:24:54,261 | INFO | NIOServerCxn.Factory:192-168-0-114/192.168.0.114:2181 | Successfully authenticated client: authenticationID=hdfs/hadoop@<系统域名>; authorizationID=hdfs/hadoop@<系统域名>. | org.apache.zookeeper.server.auth.SaslServerCallbackHandler.handleAuthorizeCallback(SaslServerCallbackHandler.java:118)
2015-04-15 21:24:54,261 | INFO | NIOServerCxn.Factory:192-168-0-114/192.168.0.114:2181 | Setting authorizedID: hdfs/hadoop@<系统域名> | org.apache.zookeeper.server.auth.SaslServerCallbackHandler.handleAuthorizeCallback(SaslServerCallbackHandler.java:134)
2015-04-15 21:24:54,261 | INFO | NIOServerCxn.Factory:192-168-0-114/192.168.0.114:2181 | adding SASL authorization for authorizationID: hdfs/hadoop@<系统域名> | org.apache.zookeeper.server.ZooKeeperServer.processSasl(ZooKeeperServer.java:1009)
2015-04-15 21:24:54,262 | INFO | ProcessThread(sid:22 cport:-1): | Got user-level KeeperException when processing sessionid:0x164cb2b3e4b36ae4 type:create cxid:0x3 zxid:0x20009fafc txntype:-1 reqpath:n/a Error Path:/hadoop-ha/hacluster/ActiveStandbyElectorLock Error:KeeperErrorCode = NodeExists for /hadoop-ha/hacluster/ActiveStandbyElectorLock | org.apache.zookeeper.server.PrepRequestProcessor.pRequest(PrepRequestProcessor.java:648)
```

ZKFC日志，如下所示：

```
2015-04-15 21:24:54,237 | INFO | main-SendThread(192-168-0-114:2181) | Socket connection established to 192-168-0-114/192.168.0.114:2181, initiating session | org.apache.zookeeper.ClientCnxn$SendThread.primeConnection(ClientCnxn.java:854)
2015-04-15 21:24:54,257 | INFO | main-SendThread(192-168-0-114:2181) | Session establishment complete on server 192-168-0-114/192.168.0.114:2181, sessionid = 0x144cb2b3e4b36ae4 , negotiated timeout = 45000 | org.apache.zookeeper.ClientCnxn$SendThread.onConnected(ClientCnxn.java:1259)
2015-04-15 21:24:54,260 | INFO | main-EventThread | EventThread shut down | org.apache.zookeeper.ClientCnxn$EventThread.run(ClientCnxn.java:512)
2015-04-15 21:24:54,262 | INFO | main-EventThread | Session connected. | org.apache.hadoop.ha.ActiveStandbyElector.processWatchEvent(ActiveStandbyElector.java:547)
2015-04-15 21:24:54,264 | INFO | main-EventThread | Successfully authenticated to ZooKeeper using SASL. | org.apache.hadoop.ha.ActiveStandbyElector.processWatchEvent(ActiveStandbyElector.java:573)
```

回答

- 原因分析

  NameNode的主节点重启后，它原先在Zookeeper上建立的临时节点（/hadoop-ha/hacluster/ActiveStandbyElectorLock）就会被清理。同时，NameNode备节点发现这个信息后进行抢占希望升主，所以它重新在Zookeeper上建立了active的节点/hadoop-ha/hacluster/ActiveStandbyElectorLock。但是NameNode备节点通过客户端（ZKFC）与Zookeeper建立连接时，由于网络问题、CPU使用率高、集群压力大等原因，出现了客户端（ZKFC）的session（0x144cb2b3e4b36ae4）与Zookeeper服务端的session（0x164cb2b3e4b36ae4）不一致的问题，这就导致了NameNode备节点的watcher没有感知到自己已经成功建立临时节点，依然认为自己还是备。 而NameNode主节点启动后，发现/hadoop-ha/hacluster目录下已经有active的节点，所以也无法升主，导致两个节点都为备。

- 解决方法

  建议通过在FusionInsight Manager界面上重启HDFS的两个ZKFC加以解决。

12.9.33.16 HDFS 执行 Balance 时被异常停止，再次执行 Balance 会失败

问题

在HDFS客户端启动一个Balance进程，该进程被异常停止后，再次执行Balance操作，操作会失败。

回答

通常，HDFS执行Balance操作结束后，会自动释放“/system/balancer.id”文件，可再次正常执行Balance。

但在上述场景中，由于第一次的Balance操作是被异常停止的，所以第二次进行Balance操作时，“/system/balancer.id”文件仍然存在，则会触发***append /system/balancer.id***操作，进而导致Balance操作失败。

- 如果“/system/balancer.id”文件的释放时间超过了软租期60s，则第二次执行Balance操作的客户端的append操作会抢占租约，此时最后一个block处于under construction或者under recovery状态，会触发block的恢复操作，那么“/system/balancer.id”文件必须等待block恢复完成才能关闭，所以此次append操作失败。

  append /system/balancer.id操作失败后，会向客户端抛出RecoveryInProgressException异常：

  ```
  org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.protocol.RecoveryInProgressException): Failed to APPEND_FILE /system/balancer.id for DFSClient because lease recovery is in progress. Try again later.
  ```

- 如果该文件的释放时间没有超过默认设置60s，原有客户端会继续持有该租约，则会抛出AlreadyBeingCreatedException异常，实际上向客户端返回的是null，导致客户端出现如下异常：

  ```
  java.io.IOException: Cannot create any NameNode Connectors.. Exiting...
  ```

可通过以下方法避免上述问题：

- 方案1：等待硬租期超过1小时后，原有客户端释放租约，再执行第二次Balance操作。
- 方案2：执行第二次Balance操作之前删除“/system/balancer.id”文件。

12.9.33.17 IE 浏览器访问 HDFS 原生 UI 界面失败，显示无法显示此页

问题

通过IE 9、IE 10和IE 11浏览器访问HDFS的原生UI界面，偶尔出现访问失败情况。

现象

访问页面失败，浏览器无法显示此页，如下图所示：

原因

IE 9、IE 10、IE 11浏览器的某些版本在处理SSL握手有问题导致访问失败。

解决方法

重新刷新页面即可。

12.9.33.18 EditLog 不连续导致 NameNode 启动失败

问题

在JournalNode节点有断电，数据目录磁盘占满，网络异常时，会导致JournalNode上的EditLog不连续。此时如果重启NameNode，很可能会失败。

现象

重启NameNode会失败。在NameNode运行日志中会报如下的错误：

```
2019-11-08 16:30:28,399 | ERROR | main | Failed to start namenode. | NameNode.java:1732
java.io.IOException: There appears to be a gap in the edit log.  We expected txid 13698019, but got txid 13698088.
        at org.apache.hadoop.hdfs.server.namenode.MetaRecoveryContext.editLogLoaderPrompt(MetaRecoveryContext.java:94)
        at org.apache.hadoop.hdfs.server.namenode.FSEditLogLoader.loadEditRecords(FSEditLogLoader.java:278)
        at org.apache.hadoop.hdfs.server.namenode.FSEditLogLoader.loadFSEdits(FSEditLogLoader.java:188)
        at org.apache.hadoop.hdfs.server.namenode.FSImage.loadEdits(FSImage.java:924)
        at org.apache.hadoop.hdfs.server.namenode.FSImage.loadFSImage(FSImage.java:771)
        at org.apache.hadoop.hdfs.server.namenode.FSImage.recoverTransitionRead(FSImage.java:331)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.loadFSImage(FSNamesystem.java:1108)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.loadFromDisk(FSNamesystem.java:727)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.loadNamesystem(NameNode.java:638)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.initialize(NameNode.java:700)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.<init>(NameNode.java:943)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.<init>(NameNode.java:916)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.createNameNode(NameNode.java:1655)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.main(NameNode.java:1725)
```

解决方法

1. 找到重启前的主NameNode，进入其数据目录（查看配置项“dfs.namenode.name.dir”可获取，例如/srv/BigData/namenode/current），得到最新的FSImage文件的序号。一般如下：

```
-rw-------. 1 omm wheel     574 Oct  2 01:12 edits_0000000000013259401-00000000000
-rw-------. 1 omm wheel     575 Oct  2 01:13 edits_0000000000013259409-00000000000
-rw-------. 1 omm wheel      42 Oct  2 01:13 edits_0000000000013259417-00000000000
-rw-------. 1 omm wheel 1048576 Nov  8 16:01 edits_inprogress_0000000000013698088
-rw-------. 1 omm wheel  314803 Nov  8 15:53 fsimage_0000000000013698018
-rw-------. 1 omm wheel      62 Nov  8 15:53 fsimage_0000000000013698018.md5
-rw-------. 1 omm wheel  314803 Nov  8 15:56 fsimage_0000000000013698050
-rw-------. 1 omm wheel      62 Nov  8 15:56 fsimage_0000000000013698050.md5
-rw-------. 1 omm wheel  314803 Nov  8 15:59 fsimage_0000000000013698066
-rw-------. 1 omm wheel      62 Nov  8 15:59 fsimage_0000000000013698066.md5
-rw-------. 1 omm wheel       9 Oct  2 01:13 seen_txid
-rw-------. 1 omm wheel     187 Nov  8 15:59 VERSION
```

2. 查看各JournalNode的数据目录（查看配置项“dfs.journalnode.edits.dir”可获取，例如/srv/BigData/journalnode/hacluster/current），查看序号从第一部获取到的序号开始的edits文件，看是否有不连续的情况（即前一个edits文件的最后一个序号 和 后一个edits文件的第一个序号 不是连续的，如下图中的edits_0000000000013259231-0000000000013259237就和后一个edits_0000000000013259239-0000000000013259246就是不连续的）。

```
-rw-------. 1 omm wheel     575 Oct  2 00:43 edits_0000000000013259159-0000000000013259166
-rw-------. 1 omm wheel     576 Oct  2 00:43 edits_0000000000013259167-0000000000013259174
-rw-------. 1 omm wheel     575 Oct  2 00:45 edits_0000000000013259175-0000000000013259182
-rw-------. 1 omm wheel     575 Oct  2 00:45 edits_0000000000013259183-0000000000013259190
-rw-------. 1 omm wheel     576 Oct  2 00:47 edits_0000000000013259191-0000000000013259198
-rw-------. 1 omm wheel     575 Oct  2 00:48 edits_0000000000013259199-0000000000013259206
-rw-------. 1 omm wheel     575 Oct  2 00:49 edits_0000000000013259207-0000000000013259214
-rw-------. 1 omm wheel     575 Oct  2 00:50 edits_0000000000013259215-0000000000013259222
-rw-------. 1 omm wheel     573 Oct  2 00:51 edits_0000000000013259223-0000000000013259230
-rw-------. 1 omm wheel     571 Oct  2 00:52 edits_0000000000013259231-0000000000013259237
-rw-------. 1 omm wheel     576 Oct  2 00:53 edits_0000000000013259239-0000000000013259246
-rw-------. 1 omm wheel     575 Oct  2 00:54 edits_0000000000013259247-0000000000013259254
-rw-------. 1 omm wheel     576 Oct  2 00:55 edits_0000000000013259255-0000000000013259262
-rw-------. 1 omm wheel      42 Oct  2 00:56 edits_0000000000013259263-0000000000013259264
-rw-------. 1 omm wheel    1107 Oct  2 00:57 edits_0000000000013259265-0000000000013259278
-rw-------. 1 omm wheel      42 Oct  2 00:58 edits_0000000000013259279-0000000000013259280
-rw-------. 1 omm wheel    1109 Oct  2 00:59 edits_0000000000013259281-0000000000013259294
-rw-------. 1 omm wheel      42 Oct  2 01:00 edits_0000000000013259295-0000000000013259296
-rw-------. 1 omm wheel    1299 Oct  2 01:01 edits_0000000000013259297-0000000000013259312
-rw-------. 1 omm wheel     260 Oct  2 01:02 edits_0000000000013259313-0000000000013259316
-rw-------. 1 omm wheel     984 Oct  2 01:03 edits_0000000000013259317-0000000000013259328
-rw-------. 1 omm wheel     572 Oct  2 01:04 edits_0000000000013259329-0000000000013259336
-rw-------. 1 omm wheel     575 Oct  2 01:05 edits_0000000000013259337-0000000000013259344
```

3. 如果有这种不连续的edits文件，则需要查看其它的JournalNode的数据目录或NameNode数据目录中，有没有连续的该序号相关的连续的edits文件。如果可以找到，复制一个连续的片段到该JournalNode。
4. 如此把所有的不连续的edits文件全部都修复。
5. 重启NameNode，观察是否成功。如还是失败，请联系技术支持。

12.10 使用 Hive

12.10.1 从零开始使用 Hive

Hive是基于Hadoop的一个数据仓库工具，可将结构化的数据文件映射成一张数据库表，并提供类SQL的功能对数据进行分析处理，通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

背景信息

假定用户开发一个应用程序，用于管理企业中的使用A业务的用户信息，使用Hive客户端实现A业务操作流程如下：

普通表的操作：

- 创建用户信息表user_info。
- 在用户信息中新增用户的学历、职称信息。
- 根据用户编号查询用户姓名和地址。
- A业务结束后，删除用户信息表。

表 12-227 用户信息

编号	姓名	性别	年龄	地址
12005000201	A	男	19	A城市
12005000202	B	女	23	B城市
12005000203	C	男	26	C城市
12005000204	D	男	18	D城市
12005000205	E	女	21	E城市
12005000206	F	男	32	F城市
12005000207	G	女	29	G城市
12005000208	H	女	30	H城市
12005000209	I	男	26	I城市
12005000210	J	女	25	J城市

操作步骤

步骤1 下载客户端配置文件。

- MRS 3.x之前版本，操作如下：
  a. 登录MRS Manager页面，具体请参见**访问集群Manager**，然后选择“服务管理”。
  b. 单击“下载客户端”。
  “客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。
- MRS 3.x及后续版本，操作如下：
  a. 登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。
  b. 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 下载客户端”。

c. 下载集群客户端。

“选择客户端类型”选择“仅配置文件”，选择平台类型，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client/”。

步骤2 登录Manager的主管理节点。

- MRS 3.x之前版本，操作如下：

a. 在MRS控制台，选择“集群列表 > 现有集群”，单击集群名称，在“节点管理”页签中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。

MRS Manager的主备管理节点默认安装在集群Master节点上。在主备模式下，由于Master1和Master2之间会切换，Master1节点不一定是MRS Manager的主管理节点，需要在Master1节点中执行命令，确认MRS Manager的主管理节点。命令请参考**步骤2.d**。

b. 以**root**用户使用密码方式登录Master1节点。

c. 切换至**omm**用户。

sudo su - root

su - omm

d. 执行以下命令确认MRS Manager的主管理节点。

sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh

回显信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“mgtomsdat-sh-3-01-1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“mgtomsdat-sh-3-01-2”为备管理节点）。

```
Ha mode
double
NodeName            HostName                HAVersion          StartTime
HAActive            HAAllResOK          HARunPhase
192-168-0-30        mgtomsdat-sh-3-01-1     V100R001C01        2014-11-18 23:43:02
active              normal              Actived
192-168-0-24        mgtomsdat-sh-3-01-2     V100R001C01        2014-11-21 07:14:02
standby             normal              Deactived
```

e. 使用**root**用户登录Manager的主管理节点，例如“192-168-0-30”节点。

- MRS 3.x及后续版本，操作如下：

a. 以**root**用户登录任意部署Manager的节点。

b. 执行以下命令确认主备管理节点。

sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh

界面打印信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“node-master1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“node-master2”为备管理节点）。

```
HAMode
double
NodeName            HostName        HAVersion          StartTime           HAActive
HAAllResOK          HARunPhase
192-168-0-30        node-master1    V100R001C01        2020-05-01 23:43:02     active
normal              Actived
192-168-0-24        node-master2    V100R001C01        2020-05-01 07:14:02
standby             normal              Deactived
```

c. 以**root**用户登录主管理节点，并执行以下命令切换到**omm**用户。

sudo su - omm

步骤3 执行以下命令切换到客户端安装目录。

提前已安装集群客户端，以下客户端安装目录为举例，请根据实际情况修改。

cd /opt/client

步骤4 执行以下命令，更新主管理节点的客户端配置。

sh refreshConfig.sh /opt/client *客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/client /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

步骤5 在Master节点使用客户端。

1. 在已更新客户端的主管理节点，例如“192-168-0-30”节点，执行以下命令切换到客户端目录，客户端安装目录如：/opt/client。

   cd /opt/client

2. 执行以下命令配置环境变量。

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建Hive表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit hiveuser**。

4. 直接执行Hive组件的客户端命令。

   beeline

步骤6 运行Hive客户端命令，实现A业务。

内部表的操作：

1. 根据**表12-227**创建用户信息表user_info并添加相关数据，例如：

   create table user_info(id string,name string,gender string,age int,addr string);

   MRS 1.x和MRS 3.x及后续版本，操作如下：

   insert into table user_info(id,name,gender,age,addr) values("12005000201","A","男",19,"A城市");

   MRS 2.x版本，操作如下：

   insert into table user_info values("12005000201","A","男",19,"A城市");

2. 在用户信息表user_info中新增用户的学历、职称信息。

   以增加编号为12005000201的用户的学历、职称信息为例，其他用户类似。

   alter table user_info add columns(education string,technical string);

3. 根据用户编号查询用户姓名和地址。

   以查询编号为12005000201的用户姓名和地址为例，其他用户类似。

   select name,addr from user_info where id='12005000201';

4. 删除用户信息表。

   drop table user_info;

外部分区表的操作：

创建外部分区表并导入数据：

1. 创建外部表数据存储路径：

   hdfs dfs -mkdir /hive/

   hdfs dfs -mkdir /hive/user_info

2. 建表：

   create external table user_info(id string,name string,gender string,age int,addr string) partitioned by(year string) row format delimited fields terminated by ' ' lines terminated by '\n' stored as textfile location '/hive/user_info';

   说明

   fields terminated指明分隔的字符，如按空格分隔，' '。

   lines terminated 指明分行的字符，如按换行分隔，'\n'。

   /hive/user_info为数据文件的路径。

3. 导入数据。

   a. 使用insert语句插入数据。

      insert into user_info partition(year="2018") values ("12005000201","A","男",19,"A城市");

   b. 使用load data命令导入文件数据。

      i. 根据**表12-227**数据创建文件。如，文件名为txt.log，以空格拆分字段，以换行符作为行分隔符。

      ii. 上传文件至hdfs。

      hdfs dfs -put txt.log /tmp

      iii. 加载数据到表中。

      load data inpath '/tmp/txt.log' into table user_info partition (year='2011');

4. 查询导入数据。

   select * from user_info;

5. 删除用户信息表。

   drop table user_info;

6. 执行以下命令退出客户端。

   !q

----结束

12.10.2 配置 Hive 常用参数

参数入口

请参考**修改集群服务配置参数**进入Hive服务配置页面。

参数说明

表 12-228 Hive 参数说明

参数	参数说明	默认值
hive.auto.convert.join	Hive基于输入文件大小将普通join转为mapjoin的开关。 **说明** 在使用Hive进行联表查询，且关联的表无大小表的分别（小表数据<24M）时，建议将此参数值改为false，如果此时将此参数设置为true，执行联表查询时无法生成新的mapjoin。	取值范围： • true • false 默认值为true
hive.default.fileformat	Hive使用的默认文件格式。	MRS 3.x之前版本：TextFile MRS 3.x及后续版本：RCFile
hive.exec.reducers.max	Hive提交的MR任务中reducer的最大个数。	999
hive.server2.thrift.max.worker.threads	HiveServer内部线程池，最大能启动的线程数量。	1000
hive.server2.thrift.min.worker.threads	HiveServer内部线程池，初始化时启动的线程数量。	5
hive.hbase.delete.mode.enabled	从Hive删除HBase记录的功能开关。如果启用，用户可以使用“remove table xx where xxx”从Hive中删除HBase记录。 **说明** 本参数适用于MRS 3.x及后续版本。	true
hive.metastore.server.min.threads	MetaStore启动的用于处理连接的线程数，如果超过设置的值之后，MetaStore就会一直维护不低于设定值的线程数，即常驻MetaStore线程池的线程会维护在指定值之上。	200
hive.server2.enable.doAs	HiveServer2在与其他服务（如YARN、HDFS等）会话时是否模拟客户端用户。如果将此配置项从false改成true，会导致只有列权限的用户访问相应表权限缺失。 **说明** 本参数适用于MRS 3.x及后续版本。	true

12.10.3 Hive SQL

Hive SQL支持Hive-3.1.0版本中的所有特性，详情请参见**https://cwiki.apache.org/confluence/display/hive/languagemanual**。

系统提供的扩展Hive语句如表12-229所示。

表 12-229 扩展 Hive 语句

扩展语法	语法说明	语法示例	示例说明
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name (col_name data_type [COMMENT col_comment], ...) [ROW FORMAT row_format] [STORED AS file_format] | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (...)] ..... [TBLPROPERTIES ("groupId"=" group1 ","locatorId"="locator1")] ...;	创建一个hive表，并指定表数据文件分布的locator信息。详细说明请参见使用HDFS Colocation存储Hive表。	CREATE TABLE tab1 (id INT, name STRING) row format delimited fields terminated by '\t' stored as RCFILE TBLPROPERTIES(" groupId"=" group1 ","locatorId"="locator1");	创建表tab1，并指定tab1的表数据分布在locator1节点上。

扩展语法	语法说明	语法示例	示例说明
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name (col_name data_type [COMMENT col_comment], ...) [ROW FORMAT row_format] [STORED AS file_format] | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (...)] ... [TBLPROPERTIES (**'column.encode.columns'**='col_name1,col_name2'| **'column.encode.indices'**='col_id1,col_id2', **'column.encode.classname'**='encode_classname')]...;	创建一个hive表，并指定表的加密列和加密算法。详细说明请参见**使用Hive列加密功能**。	create table encode_test(id INT, name STRING, phone STRING, address STRING) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe' WITH SERDEPROPERTIES ('column.encode.indices'='2,3', 'column.encode.classname'='org.apache.hadoop.hive.serde2.SMS4Rewriter') STORED AS TEXTFILE;	创建表encode_test，并指定插入数据时对第2、3列加密，加密算法类为org.apache.hadoop.hive.serde2.SMS4Rewriter。
REMOVE TABLE *hbase_tablename* [WHERE where_condition];	删除hive on hbase表中符合条件的数据。详细说明请参见**删除Hive on HBase表中的单行记录**。	remove table hbase_table1 where id = 1;	删除表中符合条件“id =1”的数据。

扩展语法	语法说明	语法示例	示例说明
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name (col_name data_type [COMMENT col_comment], ...) [ROW FORMAT row_format] **STORED AS inputformat 'org.apache.hadoop.hive.contrib.fileformat.SpecifiedDelimiterInputFormat'** outputformat 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';	创建hive表，并设定表可以指定自定义行分隔符。详细说明请参见**自定义行分隔符**。	create table blu(time string, num string, msg string) row format delimited fields terminated by ',' **stored as inputformat 'org.apache.hadoop.hive.contrib.fileformat.SpecifiedDelimiterInputFormat'** outputformat 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';	创建表blu，指定inputformat为SpecifiedDelimiterInputFormat，以便查询时可以指定表的查询行分隔符。

12.10.4 权限管理

12.10.4.1 Hive 权限介绍

Hive是建立在Hadoop上的数据仓库框架，提供类似SQL的HQL操作结构化数据。

MRS提供用户、用户组和角色，集群中的各类权限需要先授予角色，然后将用户或者用户组与角色绑定。用户只有绑定角色或者加入绑定角色的用户组，才能获得权限。Hive授权相关信息请参考：**https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Authorization**。

说明

- Hive在安全模式下需要进行权限管理，在普通模式下无需进行权限管理。
- MRS 3.x及后续版本支持Ranger，如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Hive的Ranger访问权限策略**。

Hive 权限模型

使用Hive组件，必须对Hive数据库和表（含外表和视图）拥有相应的权限。在MRS中，完整的Hive权限模型由Hive元数据权限与HDFS文件权限组成。使用数据库或表时所需要的各种权限都是Hive权限模型中的一种。

- Hive元数据权限。

  与传统关系型数据库类似，MRS的Hive数据库包含“建表”和“查询”权限，Hive表和列包含“查询”、“插入”和“删除”权限。Hive中还包含拥有者权限“OWNERSHIP”和“Hive管理员权限”。

  说明

  Hive表和列的“更新”和“删除”操作只有在开启“ACID”的情况下支持，目前的版本不支持开启“ACID”。

- Hive数据文件权限，即HDFS文件权限。

  Hive的数据库、表对应的文件保存在HDFS中。默认创建的数据库或表保存在HDFS目录“/user/hive/warehouse”。系统自动以数据库名称和数据库中表的名称创建子目录。访问数据库或者表，需要在HDFS中拥有对应文件的权限，包含“读”、“写”和“执行”权限。

  说明

  MRS 3.X支持Hive多实例，多实例场景下，目录为“/user/hive*n*（*n=1～4*）/warehouse”。

用户对Hive数据库或表执行不同操作时，需要关联不同的元数据权限与HDFS文件权限。例如，对Hive数据表执行查询操作，需要关联元数据权限“查询”，以及HDFS文件权限“读”和“写”。

使用Manager界面图形化的角色管理功能来管理Hive数据库和表的权限，只需要设置元数据权限，系统会自动关联HDFS文件权限，减少界面操作，提高效率。

Hive 用户对象

MRS提供了用户和角色来使用Hive，比如创建表、在表中插入数据或者查询表。Hive中定义了“USER”类，对应用户实例；定义了“GROUP”类，对应角色实例。

使用Manager设置Hive用户对象的权限，只支持在角色中设置，用户或用户组需要绑定角色才能获得权限。支持授予管理员权限、访问数据库、表和列的权限。

Hive 使用场景及对应权限

用户使用Hive并创建数据库需要加入hive组，不需要角色授权。用户在Hive和HDFS中对自己创建的数据库或表拥有完整权限，可直接创建表、查询数据、删除数据、插入数据、更新数据以及授权他人访问表与对应HDFS目录与文件。

如果用户访问别人创建的表或数据库，需要授予权限。所以根据Hive使用场景的不同，用户需要的权限可能也不相同。

表 12-230 Hive 使用场景

主要场景	用户需要的权限
使用Hive表、列或数据库	使用其他用户创建的Hive表、列或数据库，不同的场景需要不同的Hive权限，例如： ● 创建表，需要“建表”。 ● 查询数据，需要“查询”。 ● 插入数据，需要“插入”。 ● 删除数据，需要“删除”。
关联使用其他组件	部分场景除了Hive权限，还可能需要组件的权限，例如： ● 执行部分HQL命令，例如**insert**，**count**，**distinct**，**group by**，**order by**，**sort by**或**join**等语句时，需要设置YARN权限。建议为每个Hive用户的角色添加此权限。 ● 使用Hive over HBase，例如在Hive中查询HBase表数据，需要设置HBase权限。

在一些特殊Hive使用场景下，需要单独设置其他权限。

表 12-231 Hive 授权注意事项

可能场景	用户需要的权限
创建Hive数据库、表、外表，或者为已经创建的Hive表或外表添加分区，且Hive用户指定数据文件保存在“/user/hive/warehouse”以外的HDFS目录。	需要此目录已经存在，Hive用户是目录的属主，且用户对目录拥有“读”、“写”和“执行”权限。同时用户对此目录上层的每一级目录都拥有“读”和“写”权限。然后管理员通过角色管理功能授予角色使用Hive的权限，会自动关联HDFS权限。

可能场景	用户需要的权限
Hive用户使用**load**将指定目录下所有文件或者指定文件，导入数据到Hive表。	• 数据源为Linux本地磁盘，指定目录时需要此目录已经存在，系统用户“omm”对此目录以及此目录上层的每一级目录拥有“r”和“x”的权限。指定文件时需要此文件已经存在，“omm”对此文件拥有“r”的权限，同时对此文件上层的每一级目录拥有“r”和“x”的权限。 • 数据源为HDFS，指定目录时需要此目录已经存在，Hive用户是目录属主，且用户对此目录及其子目录拥有“读”、“写”和“执行”权限，并且其上层的每一级目录拥有“读”和“写”权限。指定文件时需要此文件已经存在，Hive用户是文件属主，且用户对文件拥有“读”、“写”和“执行”权限，同时对此文件上层的每一级目录拥有“读”和“执行”权限。 **说明** 使用**load**从Linux本地磁盘导入数据时，文件需上传到执行命令的HiveServer并修改权限。建议使用客户端执行命令，可查看客户端连接的HiveServer。例如，Hive客户端显示“0: jdbc:hive2://10.172.0.43:21066/>”，表示当前连接的HiveServer节点IP地址为“10.172.0.43”。
创建函数、删除函数或者修改任意数据库。	需要授予“Hive管理员权限”。
操作Hive中所有的数据库和表。	需加入到supergroup用户组，并且授予“Hive管理员权限”。

12.10.4.2 创建 Hive 角色

操作场景

该任务指导系统管理员在Manager创建并设置Hive的角色。Hive角色可设置Hive管理员权限以及Hive数据表的数据操作权限。

用户使用Hive并创建数据库需要加入hive组，不需要角色授权。用户在Hive和HDFS中对自己创建的数据库或表拥有完整权限，可直接创建表、查询数据、删除数据、插入数据、更新数据以及授权他人访问表与对应HDFS目录与文件。默认创建的数据库或表保存在HDFS目录“/user/hive/warehouse”。

说明

- 安全模式支持创建Hive角色，普通模式不支持创建Hive角色。
- MRS 3.x及后续版本支持Ranger，如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Hive的Ranger访问权限策略**。

前提条件

- 系统管理员已明确业务需求。
- 已登录Manager。
- 已安装好Hive客户端。

操作步骤

MRS 3.x之前版本，创建Hive角色的操作如下：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 权限配置 > 角色管理”。

步骤3 单击“添加角色”，输入“角色名称”和“描述”。

步骤4 设置角色“权限”请参见**表12-232**。

- “Hive Admin Privilege”：Hive管理员权限。如需使用该权限，在执行SQL语句时需要先执行**set role admin**来设置权限。
- “Hive Read Write Privileges”：Hive数据表管理权限，可设置与管理已创建的表的数据操作权限。根据需要勾选相应database的权限，如果要精确到表，可以单击database名称，勾选相应表的权限。

说明

- Hive角色管理支持授予管理员权限、访问表和视图的权限，不支持数据库的授权。
- Hive管理员权限不支持管理HDFS的权限。
- 如果数据库中的表或者表中的文件数量比较多，在授权时可能需要等待一段时间。例如表的文件数量为1万时，可能需要等待2分钟。

表 12-232 设置角色

任务场景	角色授权操作
设置Hive管理员权限	在“权限”的表格中单击“Hive”，勾选“Hive Admin Privilege”。 **说明** 用户绑定Hive管理员角色后，在每个维护操作会话中，还需要执行以下操作： 1. 请根据客户端所在位置，登录安装客户端的节点。 2. 执行以下命令配置环境变量。 例如，Hive客户端安装目录为“/opt/hiveclient”，执行**source /opt/hiveclient/bigdata_env** 3. 执行以下命令认证用户。 **kinit** *Hive业务用户* 4. 执行以下命令登录客户端工具。 **beeline** 5. 执行以下命令更新用户的管理员权限。 **set role admin;**

任务场景	角色授权操作
设置在默认数据库中，查询其他用户表的权限	1. 在“权限”的表格中选择“Hive > Hive Read Write Privileges”。 2. 在指定表的“权限”列，勾选“Select”。
设置在默认数据库中，插入其他用户表的权限	1. 在“权限”的表格中选择“Hive > Hive Read Write Privileges”。 2. 在指定表的“权限”列，勾选“Insert”。
设置在默认数据库中，导入数据到其他用户表的权限	1. 在“权限”的表格中选择“Hive > Hive Read Write Privileges”。 2. 在指定表的“权限”列，勾选“Delete”和“Insert”。
设置提交Hql命令到Yarn执行的权限	部分业务需求使用的Hql命令将转化为MapReduce任务并提交到Yarn中执行，需要设置Yarn权限。例如运行的HQL使用了**insert**，**count**，**distinct**，**group by**，**order by**，**sort by**或**join**等语句的相关场景。 1. 在“权限”的表格中选择“Yarn > Scheduler Queue > root”。 2. 在“default”队列的“权限”列，勾选“Submit”。

步骤5 单击“确定”，返回“角色”。

步骤6 选择“系统设置 > 用户管理 > 添加用户”。

步骤7 输入用户名，在“用户类型”选择“人机”类型，设置用户密码，在用户组添加一个绑定了Hive管理员角色的用户组，并绑定新创建的Hive角色，单击“确定”完成Hive用户创建。

步骤8 待用户生成后，即可使用该用户执行相应SQL语句。

----结束

MRS 3.x及后续版本，创建Hive角色的操作如下：

步骤1 登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 选择“系统 > 权限 > 角色”。

步骤3 单击“添加角色”，输入“角色名称”和“描述”。

步骤4 设置角色“配置资源权限”请参见**表12-233**。

- 设置HDFS目录的读和执行权限。
  - 选择“*待操作集群的名称* > HDFS > 文件系统 > hdfs://hacluster/ > user”，在“hive”的“权限”列，勾选“读”和“执行”。

  - 选择“*待操作集群的名称* > HDFS > 文件系统 > hdfs://hacluster/ > user > hive”，在“warehouse”的“权限”列，勾选“读”和“执行”。
  - 选择“*待操作集群的名称* > HDFS > 文件系统 > hdfs://hacluster/ > tmp”，在“hive-scratch”的“权限”列，勾选“读”和“执行”。
- “Hive管理员权限”：Hive管理员权限。
- “Hive读写权限”：Hive数据表管理权限，可设置与管理已创建的表的数据操作权限。

说明

- MRS 3.1.0版本，Hive角色管理支持授予管理员权限、访问表和视图的权限，不支持数据库的授权。
- Hive管理员权限不支持管理HDFS的权限。
- 如果数据库中的表或者表中的文件数量比较多，在授权时可能需要等待一段时间。例如表的文件数量为1万时，可能需要等待2分钟。

表 12-233 设置角色

任务场景	角色授权操作
设置Hive管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive”，勾选“Hive管理员权限”。 **说明** 用户绑定Hive管理员角色后，在每个维护操作会话中，还需要执行以下操作： 1. 以客户端安装用户，登录安装Hive客户端的节点。 2. 执行以下命令配置环境变量。 例如，Hive客户端安装目录为“/opt/hiveclient”，执行**source /opt/hiveclient/bigdata_env** 3. 执行以下命令认证用户。 **kinit** *Hive业务用户* 4. 执行以下命令登录客户端工具。 **beeline** 5. 执行以下命令更新用户的管理员权限。 **set role admin;**
设置在默认数据库中，查询其他用户表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”。 2. 在数据库列表中单击指定的数据库名称，显示数据库中的表。 3. 在指定表的“权限”列，勾选“查询”。

任务场景	角色授权操作
设置在默认数据库中，插入其他用户表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”。 2. 在数据库列表中单击指定的数据库名称，显示数据库中的表。 3. 在指定表的“权限”列，勾选“插入”。
设置在默认数据库中，导入数据到其他用户表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”。 2. 在数据库列表中单击指定的数据库名称，显示数据库中的表。 3. 在指定表的“权限”列，勾选“删除”和“插入”。
设置提交Hql命令到Yarn执行的权限	部分业务需求使用的Hql命令将转化为MapReduce任务并提交到Yarn中执行，需要设置Yarn权限。例如运行的HQL使用了**insert**，**count**，**distinct**，**group by**，**order by**，**sort by**或**join**等语句的相关场景。 1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Yarn > 调度队列 > root”。 2. 在“default”队列的“权限”列，勾选“提交”。

步骤5 单击“确定”，返回“角色”。

----结束

12.10.4.3 配置 Hive 表、列或数据库的权限

操作场景

使用Hive表或者数据库时，如果用户访问别人创建的表或数据库，需要授予对应的权限。为了实现更严格权限控制，Hive也支持列级别的权限控制。如果要访问别人创建的表上某些列，需要授予列权限。以下介绍使用Manager角色管理功能在表授权、列授权和数据库授权三个场景下的操作。

说明

- 安全模式支持配置Hive表、列或数据库的权限，普通模式不支持配置Hive表、列或数据库的权限。
- MRS 3.x及后续版本支持Ranger，如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Hive的Ranger访问权限策略**。

前提条件

- 获取一个拥有管理员权限的用户，例如“admin”。
- 请参考**创建Hive角色**，在Manager界面创建一个角色，例如“hrole”，不需要设置Hive权限，设置提交Hql命令到Yarn执行的权限。
- 在Manager界面创建两个使用Hive的“人机”用户并加入“hive”组，例如“huser1”和“huser2”。“huser2”需绑定“hrole”。使用“huser1”创建一个数据库“hdb”，并在此数据库中创建表“htable”。

操作步骤

- 表授权

  用户在Hive和HDFS中对自己创建的表拥有完整权限，用户访问别人创建的表，需要授予权限。授予权限时只需要授予Hive元数据权限，HDFS文件权限将自动关联。以授予用户对应角色在表“htable”中查询、插入和删除数据的权限为例，操作步骤如下：

  MRS 3.x之前版本，表授权的操作如下：

  a. 在MRS Manager界面，选择“系统设置 > 权限配置 > 角色管理”。
  b. 在角色“hrole”所在行，单击“修改”。
  c. 选择“Hive > Hive Read Write Privileges”。
  d. 在数据库列表中单击指定的数据库名称“hdb”，显示数据库中的表“htable”。
  e. 在表“htable”的“权限”列，勾选“Select”、“Insert”和“Delete”。
  f. 单击“确定”完成。

  MRS 3.x及后续版本，表授权的操作如下：

  a. 在FusionInsight Manager界面，选择“系统 > 权限 > 角色”。
  b. 在角色“hrole”所在行，单击“修改”。
  c. 选择“*待操作的集群* > Hive > Hive读写权限”。
  d. 在数据库列表中单击指定的数据库名称“hdb”，显示数据库中的表“htable”。
  e. 在表“htable”的“权限”列，勾选“查询”、“插入”和“删除”。
  f. 单击“确定”完成。

  说明

  在角色管理中，授予角色在Hive外表中查询、插入和删除数据的操作与Hive表相同，授予元数据权限将自动关联HDFS文件权限。

- 列授权

  用户在Hive和HDFS中对自己创建的表拥有完整权限，用户没有权限访问别人创建的表。如果要访问别人创建的表上某些列，需要授予列权限。授予权限时只需要授予Hive元数据权限，HDFS文件权限将自动关联。以授予用户对应角色在表“htable”的列“hcol”中查询、插入数据的权限为例，操作步骤如下：

  MRS 3.x之前版本，列授权的操作如下：

  a. 在MRS Manager界面，选择“系统设置 > 权限配置 > 角色管理”。
  b. 在角色“hrole”所在行，单击“修改”。
  c. 选择“Hive > Hive Read Write Privileges”。

d. 在数据库列表中单击指定的数据库名称“hdb”，显示数据库中的表“htable”，单击表“htable”，显示表下的列“hcol”。
e. 在列“hcol”的“权限”列，勾选“Select”和“Insert”。
f. 单击“确定”完成。

MRS 3.x及后续版本，列授权的操作如下：

a. 在FusionInsight Manager界面，选择“系统 > 权限 > 角色”。
b. 在角色“hrole”所在行，单击“修改”。
c. 选择“*待操作的集群* > Hive > Hive读写权限”。
d. 在数据库列表中单击指定的数据库名称“hdb”，显示数据库中的表“htable”，单击表“htable”，显示表下的列“hcol”。
e. 在列“hcol”的“权限”列，勾选“查询”和“插入”。
f. 单击“确定”完成。

说明

在权限管理中，授予元数据权限将自动关联HDFS文件权限，所以列授权后会增加表对应所有文件的HDFS ACL权限。

- 数据库授权

  用户在Hive和HDFS中对自己创建的数据库拥有完整权限，用户访问别人创建的数据库，需要授予权限。授予权限时只需要授予Hive元数据权限，HDFS文件权限将自动关联。以授予用户对应角色在数据库“hdb”中查询和创建表的权限为例，操作步骤如下，不支持对角色授予数据库其他的操作权限：

  MRS 3.x之前版本，数据库授权的操作如下：

  a. 在MRS Manager界面，选择“系统设置 > 权限配置 > 角色管理”。
  b. 在角色“hrole”所在行，单击“修改”。
  c. 选择“Hive > Hive Read Write Privileges”。
  d. 在数据库“hdb”的“权限”列，勾选“Select”和“Create”。
  e. 单击“确定”完成。

  MRS 3.x及后续版本，数据库授权的操作如下：

  a. 在FusionInsight Manager界面，选择“系统 > 权限 > 角色”。
  b. 在角色“hrole”所在行，单击“修改”。
  c. 选择“*待操作的集群* > Hive > Hive读写权限”。
  d. 在数据库“hdb”的“权限”列，勾选“查询”和“建表”。
  e. 单击“确定”完成。

  说明

  - 在权限管理中，为了方便用户使用，授予数据库下表的任意权限将自动关联该数据库目录的HDFS权限。为了避免产生性能问题，取消表的任意权限，系统不会自动取消数据库目录的HDFS权限，但对应的用户只能登录数据库和查看表名。
  - 若为角色添加或删除数据库的查询权限，数据库中的表也将自动添加或删除查询权限。

相关概念

表 12-234 使用 Hive 表、列或数据库场景权限一览

操作场景	用户需要的权限
DESCRIBE TABLE	查询（Select）
SHOW PARTITIONS	查询（Select）
ANALYZE TABLE	查询（Select）、插入（Insert）
SHOW COLUMNS	查询（Select）
SHOW TABLE STATUS	查询（Select）
SHOW TABLE PROPERTIES	查询（Select）
SELECT	查询（Select）
EXPLAIN	查询（Select）
CREATE VIEW	查询（Select）、Select授权（Grant Of Select）、建表（Create）
SHOW CREATE TABLE	查询（Select）、Select授权（Grant Of Select）
CREATE TABLE	建表（Create）
ALTER TABLE ADD PARTITION	插入（Insert）
INSERT	插入（Insert）
INSERT OVERWRITE	插入（Insert）、删除（Delete）
LOAD	插入（Insert）、删除（Delete）
ALTER TABLE DROP PARTITION	删除（Delete）
CREATE FUNCTION	Hive管理员权限（Hive Admin Privilege）
DROP FUNCTION	Hive管理员权限（Hive Admin Privilege）
ALTER DATABASE	Hive管理员权限（Hive Admin Privilege）

12.10.4.4 配置 Hive 业务使用其他组件的权限

操作场景

Hive业务还可能需要关联使用其他组件，例如HQL语句触发MapReduce任务需要设置Yarn权限，或者Hive over HBase的场景需要HBase权限。以下介绍Hive关联Yarn和Hive over HBase两个场景下的操作。

说明

- 安全模式下Yarn和HBase的权限管理默认是开启的，因此在安全模式下默认需要配置Yarn和HBase权限。
- 在普通模式下，Yarn和HBase的权限管理默认是关闭的，即任何用户都有权限，因此普通模式下默认不需要配置Yarn和HBase权限。如果用户修改了YARN或者HBase的配置来开启权限管理，则修改后也需要配置Yarn和HBase权限。
- MRS 3.x及后续版本支持Ranger，如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Hive的Ranger访问权限策略**。

前提条件

- 完成Hive客户端的安装。例如安装目录为“/opt/client”。
- 获取一个拥有管理员权限的用户，例如“admin”。

操作步骤

MRS 3.x之前版本，Hive关联Yarn

用户如果执行**insert**，**count**，**distinct**，**group by**，**order by**，**sort by**或**join**等语句时，将触发MapReduce任务，需要设置Yarn权限。以授予角色在表“thc”执行**count**语句的权限为例，操作步骤如下：

步骤1 在MRS Manager角色界面创建一个角色。

步骤2 在“权限”的表格中选择“Yarn > Scheduler Queue > root”。

步骤3 在“default”队列的“权限”列，勾选“Submit”，单击“确定”保存。

步骤4 在“权限”的表格中选择“Hive > Hive Read Write Privileges > default”，勾选表“thc”的“Select”，单击“确定”保存。

----结束

MRS 3.x及后续版本，Hive关联Yarn

用户如果执行**insert**，**count**，**distinct**，**group by**，**order by**，**sort by**或**join**等语句时，将触发MapReduce任务，需要设置Yarn权限。以授予角色在表“thc”执行**count**语句的权限为例，操作步骤如下：

步骤1 在FusionInsight Manager角色界面创建一个角色。

步骤2 在“配置资源权限”的表格中选择“*待操作集群的名称* > Yarn > 调度队列 > root”。

步骤3 在“default”队列的“权限”列，勾选“提交”，单击“确定”保存。

步骤4 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限 > default”，勾选表“thc”的“查询”，单击“确定”保存。

----结束

MRS 3.x之前版本，Hive over HBase授权

用户如果需要使用类似SQL语句的方式来操作HBase表，授予权限后可以在Hive中使用HQL命令访问HBase表。以授予用户在Hive中查询HBase表的权限为例，操作步骤如下

步骤1 在MRS Manager角色管理界面创建一个HBase角色，例如“hive_hbase_create”，并授予创建HBase表的权限。

在“权限”的表格中选择“HBase > HBase Scope > global”，勾选命名空间“default”的“Create”，单击“确定”保存。

步骤2 在MRS Manager用户管理界面创建一个“人机”用户，例如“hbase_creates_user”，加入“hive”组，绑定角色“hive_hbase_create”，用于创建Hive表和HBase表。

步骤3 请根据客户端所在位置，登录安装客户端的节点。

步骤4 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤5 执行以下命令，认证用户。

kinit hbase_creates_user

步骤6 执行以下命令，进入Hive客户端shell环境：

beeline

步骤7 执行以下命令，同时在Hive和HBase中创建表。例如创建表“thh”。

CREATE TABLE thh(id int, name string, country string) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES("hbase.columns.mapping" = "cf1:id,cf1:name,:key") TBLPROPERTIES ("hbase.table.name" = "thh");

创建好的Hive表和HBase表分别保存在Hive的数据库“default”和HBase的命名空间“default”。

步骤8 在MRS Manager角色管理界面创建一个角色，例如“hive_hbase_select”，并授予查询Hive表“thh”和HBase表“thh”的权限。

1. 在“权限”的表格中选择“HBase > HBase Scope > global > default”，勾选表“thh”的“read”，单击“确定”保存，授予HBase角色查询表的权限。
2. 编辑角色，在“权限”的表格中选择“HBase > HBase Scope > global > hbase”，勾选表“hbase:meta”的“Execute”，单击“确定”保存。
3. 编辑角色，在“权限”的表格中选择“Hive > Hive Read Write Privileges > default”，勾选表“thh”的“Select”，单击“确定”保存。

步骤9 在MRS Manager用户管理界面创建一个“人机”用户，例如“hbase_select_user”，加入“hive”组，绑定角色“hive_hbase_select”，用于查询Hive表和HBase表。

步骤10 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤11 执行以下命令，认证用户。

kinit hbase_select_user

步骤12 执行以下命令，进入Hive客户端shell环境。

beeline

步骤13 执行以下命令，使用Hive的HQL语句查询HBase表的数据。

select * from thh;

----结束

MRS 3.x及后续版本，Hive over HBase授权

用户如果需要使用类似SQL语句的方式来操作HBase表，授予权限后可以在Hive中使用HQL命令访问HBase表。以授予用户在Hive中查询HBase表的权限为例，操作步骤如下

步骤1 在FusionInsight Manager角色管理界面创建一个HBase角色，例如“hive_hbase_create”，并授予创建HBase表的权限。

在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global”，勾选命名空间“default”的“创建”，单击“确定”保存。

步骤2 在FusionInsight Manager用户管理界面创建一个“人机”用户，例如“hbase_creates_user”，加入“hive”组，绑定角色“hive_hbase_create”，用于创建Hive表和HBase表。

步骤3 如果当前组件使用了Ranger进行权限控制，需给“hive_hbase_create”或“hbase_creates_user”配置“Create”权限，具体操作可参考**添加Hive的Ranger访问权限策略**。

步骤4 以客户端安装用户，登录安装客户端的节点。

步骤5 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤6 执行以下命令，认证用户。

kinit hbase_creates_user

步骤7 执行以下命令，进入Hive客户端shell环境：

beeline

步骤8 执行以下命令，同时在Hive和HBase中创建表。例如创建表“thh”。

CREATE TABLE thh(id int, name string, country string) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES("hbase.columns.mapping" = "cf1:id,cf1:name,:key") TBLPROPERTIES ("hbase.table.name" = "thh");

创建好的Hive表和HBase表分别保存在Hive的数据库“default”和HBase的命名空间“default”。

步骤9 在FusionInsight Manager角色管理界面创建一个角色，例如“hive_hbase_select”，并授予查询Hive表“thh”和HBase表“thh”的权限。

1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global > default”，勾选表“thh”的“读”，单击“确定”保存，授予HBase角色查询表的权限。
2. 编辑角色，在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global > hbase”，勾选表“hbase:meta”的“执行”，单击“确定”保存。
3. 编辑角色，在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限 > default”，勾选表“thh”的“查询”，单击“确定”保存。

步骤10 在FusionInsight Manager用户管理界面创建一个“人机”用户，例如“hbase_select_user”，加入“hive”组，绑定角色“hive_hbase_select”，用于查询Hive表和HBase表。

步骤11 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤12 执行以下命令，认证用户。

kinit hbase_select_user

步骤13 执行以下命令，进入Hive客户端shell环境。

beeline

步骤14 执行以下命令，使用Hive的HQL语句查询HBase表的数据。

select * from thh;

----结束

12.10.5 使用 Hive 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Hive客户端。

前提条件

- 已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。

使用 Hive 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 根据集群认证模式，完成Hive客户端登录。

- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。

  kinit *组件业务用户*

  beeline

- 普通模式，则执行以下命令，登录Hive客户端，如果不指定组件业务用户，则会以当前操作系统用户登录。

  beeline -n *组件业务用户*

  说明

  进行beeline连接后，可以编写并提交HQL语句执行相关任务。如需执行Catalog客户端命令，需要先执行**!q**命令退出beeline环境。

步骤5 使用以下命令，执行HCatalog的客户端命令。

hcat -e *"cmd"*

其中*"cmd"*必须为Hive DDL语句，如**hcat -e "show tables"**。

说明

- 若要使用HCatalog客户端，必须从“组件管理”页面单击“下载客户端”，下载全部服务的客户端。Beeline客户端不受此限制。
- 由于权限模型不兼容，使用HCatalog客户端创建的表，在HiveServer客户端中不能访问，但可以使用WebHCat客户端访问。
- 在普通模式下使用HCatalog客户端，系统将以当前登录操作系统用户来执行DDL命令。
- 退出beeline客户端时请使用**!q**命令，不要使用“Ctrl + c”。否则会导致连接生成的临时文件无法删除，长期会累积产生大量的垃圾文件。
- 在使用beeline客户端时，如果需要在一行中输入多条语句，语句之间以“;”分隔，需要将“entireLineAsCommand”的值设置为“false”。

  设置方法：如果未启动beeline，则执行**beeline --entireLineAsCommand=false**命令；如果已启动beeline，则在beeline中执行**!set entireLineAsCommand false**命令。

  设置完成后，如果语句中含有不是表示语句结束的“;”，需要进行转义，例如**select concat_ws('\;', collect_set(col1)) from tbl**。

----结束

使用 Hive 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 MRS 3.X支持Hive多实例，若安装了Hive多实例，在使用客户端连接具体Hive实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载Hive2实例变量：

source Hive2/component_env

步骤5 根据集群认证模式，完成Hive客户端登录。

- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。

  kinit *组件业务用户*

  beeline

- 普通模式，则执行以下命令，登录Hive客户端，如果不指定组件业务用户，则会以当前操作系统用户登录。

  beeline -n *组件业务用户*

步骤6 使用以下命令，执行HCatalog的客户端命令。

hcat -e *"cmd"*

其中*"cmd"*必须为Hive DDL语句，如**hcat -e "show tables"**。

说明

- 若要使用HCatalog客户端，必须从服务页面选择“更多 > 下载客户端”，下载全部服务的客户端。Beeline客户端不受此限制。
- 由于权限模型不兼容，使用HCatalog客户端创建的表，在HiveServer客户端中不能访问，但可以使用WebHCat客户端访问。
- 在普通模式下使用HCatalog客户端，系统将以当前登录操作系统用户来执行DDL命令。
- 退出beeline客户端时请使用**!q**命令，不要使用“Ctrl + C”。否则会导致连接生成的临时文件无法删除，长期会累积产生大量的垃圾文件。
- 在使用beeline客户端时，如果需要在一行中输入多条语句，语句之间以“;”分隔，需要将“entireLineAsCommand”的值设置为“false”。

  设置方法：如果未启动beeline，则执行**beeline --entireLineAsCommand=false**命令；如果已启动beeline，则在beeline中执行**!set entireLineAsCommand false**命令。

  设置完成后，如果语句中含有不是表示语句结束的“;”，需要进行转义，例如**select concat_ws('\;', collect_set(col1)) from tbl**。

----结束

Hive 客户端常用命令

常用的Hive Beeline客户端命令如下表所示。

更多命令可参考**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients#HiveServer2Clients-BeelineCommands**。

表 12-235 Hive Beeline 客户端常用命令

命令	说明
set <key>=<value>	设置特定配置变量（键）的值。 **说明** 若变量名拼错，Beeline不会显示错误。
set	打印由用户或Hive覆盖的配置变量列表。
set -v	打印Hadoop和Hive的所有配置变量。
add FILE[S] <filepath> <filepath>*add JAR[S] <filepath> <filepath>*add ARCHIVE[S] <filepath> <filepath>*	将一个或多个文件、JAR文件或ARCHIVE文件添加至分布式缓存的资源列表中。
add FILE[S] <ivyurl> <ivyurl>* add JAR[S] <ivyurl> <ivyurl>* add ARCHIVE[S] <ivyurl> <ivyurl>*	使用“ivy://goup:module:version?query_string”格式的Ivy URL，将一个或多个文件、JAR文件或ARCHIVE文件添加至分布式缓存的资源列表中。
list FILE[S]list JAR[S]list ARCHIVE[S]	列出已添加至分布式缓存中的资源。

命令	说明
list FILE[S] <filepath>*list JAR[S] <filepath>*list ARCHIVE[S] <filepath>*	检查给定的资源是否已添加至分布式缓存中。
delete FILE[S] <filepath>*delete JAR[S] <filepath>*delete ARCHIVE[S] <filepath>*	从分布式缓存中删除资源。
delete FILE[S] <ivyurl> <ivyurl>* delete JAR[S] <ivyurl> <ivyurl>* delete ARCHIVE[S] <ivyurl> <ivyurl>*	从分布式缓存中删除使用<ivyurl>添加的资源。
reload	使HiveServer2发现配置参数指定路径下JAR文件的变更“hive.reloadable.aux.jars.path”（无需重启HiveServer2）。更改操作包括添加、删除或更新JAR文件。
dfs <dfs command>	执行dfs命令。
<query string>	执行Hive查询，并将结果打印到标准输出。

12.10.6 使用 HDFS Colocation 存储 Hive 表

操作场景

HDFS Colocation（同分布）是HDFS提供的数据分布控制功能，利用HDFS Colocation接口，可以将存在关联关系或者可能进行关联操作的数据存放在相同的存储节点上。Hive支持HDFS的Colocation功能，即在创建Hive表时，设置表文件分布的locator信息，当使用insert语句向该表中插入数据时会将该表的数据文件存放在相同的存储节点上（不支持其他数据导入方式），从而使后续的多表关联的数据计算更加方便和高效。表格式只支持TextFile和RCFile。

说明

本章节适用于MRS 3.x及后续版本。

操作步骤

步骤1 使用客户端安装用户登录客户端所在节点。

步骤2 执行以下命令，切换到客户端安装目录，如：opt/client。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若集群为安全模式，执行以下命令认证用户。

kinit *MRS用户名*

步骤5 通过HDFS接口创建*<groupid>*

hdfs colocationadmin -createGroup -groupId *<groupid>* -locatorIds *<locatorid1>*,*<locatorid2>*,*<locatorid3>*

说明

其中*<groupid>*为创建的group名称，该示例语句创建的group包含三个locator，用户可以根据需要定义locator的数量。

关于hdfs创建groupid，以及HDFS Colocation的详细介绍请参考hdfs的相关说明，这里不做赘述。

步骤6 执行以下命令进入Hive客户端：

beeline

步骤7 Hive使用colocation。

假设table_name1和table_name2是相关联的两张表，创建两表的语句如下：

CREATE TABLE *<[db_name.]table_name1>[(col_name data_type , ...)]* [ROW FORMAT *<row_format>*] [STORED AS *<file_format>*] TBLPROPERTIES("groupId"=" *<group>* ","locatorId"="*<locator1>*");

CREATE TABLE *<[db_name.]table_name2> [(col_name data_type , ...)]* [ROW FORMAT *<row_format>*] [STORED AS *<file_format>*] TBLPROPERTIES("groupId"=" *<group>* ","locatorId"="*<locator1>*");

当使用insert语句分别向table_name1和table_name2插入数据后，table_name1和table_name2的数据文件就会分布在hdfs的相同存储位置上，从而方便两表进行关联操作。

----结束

12.10.7 使用 Hive 列加密功能

操作场景

Hive支持对表的某一列或者多列进行加密；在创建Hive表时，可以指定要加密的列和加密算法。当使用insert语句向表中插入数据时，即可实现将对应列加密。列加密只支持存储在HDFS上的TextFile和SequenceFile文件格式的表。Hive列加密不支持视图以及Hive over HBase场景。

Hive列加密机制目前支持的加密算法有两种，在建表时指定：

- AES(对应加密类名称为：org.apache.hadoop.hive.serde2.AESRewriter)
- SMS4(对应加密类名称为：org.apache.hadoop.hive.serde2.SMS4Rewriter)

说明

- 国密集群场景下，Hive列加密只支持创建SMS4算法的表，不支持创建AES算法类型的表。
- 将原始数据从普通Hive表导入到Hive列加密表后，在不影响其他业务情况下，建议删除普通Hive表上原始数据，因为保留一张未加密的表存在安全风险。

操作步骤

步骤1 在创建表时指定相应的加密列和加密算法：

create table*<[db_name.]table_name> (<col_name1> <data_type> ,<col_name2> <data_type>,<col_name3> <data_type>,<col_name4> <data_type>)* ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe' WITH SERDEPROPERTIES ('column.encode.columns'='*<col_name2>,<col_name3>*', 'column.encode.classname'='org.apache.hadoop.hive.serde2.AESRewriter')STORED AS TEXTFILE;

或者使用如下语句：

create table *<[db_name.]table_name> (<col_name1> <data_type> ,<col_name2> <data_type>,<col_name3> <data_type>,<col_name4> <data_type>)* ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe' WITH SERDEPROPERTIES ('column.encode.indices'='1,2', 'column.encode.classname'='org.apache.hadoop.hive.serde2.SMS4Rewriter') STORED AS TEXTFILE;

说明

- 使用序号指定加密列时，序号从0开始。0代表第1列，1代表第2列，依次类推。
- 创建列加密表时，表所在的目录必须是空目录。

步骤2 使用insert语法向设置列加密的表中导入数据。

假设test表已存在且有数据：

insert into table *<table_name>* select *<col_list>* from test;

----结束

12.10.8 自定义行分隔符

操作场景

通常情况下，Hive以文本文件存储的表会以回车作为其行分隔符，即在查询过程中，以回车符作为一行表数据的结束符。但某些数据文件并不是以回车分隔的规则文本格式，而是以某些特殊符号分割其规则文本。

MRS Hive支持指定不同的字符或字符组合作为Hive文本数据的行分隔符，即在创建表的时候，指定inputformat为SpecifiedDelimiterInputFormat，然后在每次查询前，都设置如下参数来指定分隔符，就可以以指定的分隔符查询表数据。

set hive.textinput.record.delimiter='';

说明

- 当前版本的Hue组件，不支持导入文件到Hive表时设置多个分割符。
- 本章节适用于MRS 3.x及后续版本。

操作步骤

步骤1 创建表时指定inputFormat和outputFormat：

CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] *[db_name.]table_name* **[(***col_name data_type* **[COMMENT** *col_comment***], ...)] [ROW FORMAT** *row_format***] STORED AS inputformat 'org.apache.hadoop.hive.contrib.fileformat.SpecifiedDelimiterInputFormat' outputformat 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'**

步骤2 查询之前指定配置项：

set hive.textinput.record.delimiter='!@!'

Hive会以‘!@!’为行分隔符查询数据。

----结束

12.10.9 配置跨集群互信下 Hive on HBase

两个开启Kerberos认证的互信集群中，使用Hive集群操作HBase集群，将目的端HBase集群的HBase关键配置项配置到源端Hive集群的HiveServer中。

前提条件

两个开启Kerberos认证的安全集群已完成跨集群互信配置。

跨集群配置 Hive on HBase

步骤1 下载HBase配置文件到本地，并解压。

1. 登录目的端HBase集群的FusionInsight Manager，选择“集群 > 服务 > HBase”。
2. 选择“更多 > 下载客户端”。
3. 下载HBase配置文件，客户端类型选择仅配置文件。

步骤2 登录源端Hive集群的FusionInsight Manager。

步骤3 选择“集群 > 服务 > Hive > 配置 > 全部配置”进入Hive服务配置页面，修改HiveServer角色的hive-site.xml自定义配置文件，增加HBase配置文件的如下配置项。

从已下载的HBase客户端配置文件的hbase-site.xml中，搜索并添加如下配置项及其取值到HiveServer中。

- hbase.security.authentication
- hbase.security.authorization
- hbase.zookeeper.property.clientPort
- hbase.zookeeper.quorum（域名需要转换为IP）
- hbase.regionserver.kerberos.principal
- hbase.master.kerberos.principal

步骤4 保存配置并重启Hive服务。

----结束

12.10.10 删除 Hive on HBase 表中的单行记录

操作场景

由于底层存储系统的原因，Hive并不能支持对单条表数据进行删除操作，但在Hive on HBase功能中，MRS Hive提供了对HBase表的单条数据的删除功能，通过特定的语法，Hive可以将自己的HBase表中符合条件的一条或者多条数据清除。

表 12-236 删除 Hive on HBase 表中的单行记录所需权限

集群认证模式	用户所需权限
安全模式	“SELECT”、“INSERT”和“DELETE”
普通模式	无

操作步骤

步骤1 如果要删除某张HBase表中的某些数据，可以执行HQL语句：

remove table ***<table_name>*** **where** ***<expression>*****;**

其中*<expression>*规定要删除数据的筛选条件；*<table_name>*为要删除数据的Hive on HBase表。

----结束

12.10.11 配置基于 HTTPS/HTTP 协议的 REST 接口

操作场景

WebHCat为Hive提供了对外可用的REST接口，开源社区版本默认使用HTTP协议。

MRS Hive支持使用更安全的HTTPS协议，并且可以在两种协议间自由切换。

说明

安全模式支持HTTPS和HTTP协议，普通模式只支持HTTP协议。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 修改Hive配置：

- MRS 3.x之前版本：在搜索框中输入参数名称，搜索“templeton.protocol.type”，修改参数值为HTTPS或者HTTP，修改后重启Hive服务即可使用对应的协议。
- MRS 3.x及后续版本：选择“WebHCat > 安全”，在该界面选择HTTPS或者HTTP，修改后重启Hive服务即可使用对应的协议。

----结束

12.10.12 配置是否禁用 Transform 功能

操作场景

Hive开源社区版本禁用Transform功能。

MRS Hive提供配置开关，默认为禁用Transform功能，与开源社区版本保持一致。

用户可修改配置开关，开启Transform功能，当开启Transform功能时，存在一定的安全风险。

说明

只有安全模式支持禁用Transform功能，普通模式不支持该功能。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 在搜索框中输入参数名称，搜索“hive.security.transform.disallow”，修改参数值为“true”或“false”，修改后重启所有HiveServer实例。

说明

- 选择“true”时，禁用Transform功能，与开源社区版本保持一致。
- 选择“false”时，开启Transform功能，存在一定的安全风险。

----结束

12.10.13 Hive 支持创建单表动态视图授权访问控制

操作场景

MRS中安全模式下Hive可以创建一个视图并控制用户访问权限，支持授权给不同的用户访问，又可以限定不同用户只能访问的不同数据。

在视图中，Hive可以通过获取当前客户端提交任务的用户的内置函数“current_user()”来进行过滤，这样被授权的用户，在访问视图时，即可被限定访问对应的数据。

说明

- 在普通模式下“current_user()”函数无法区别客户端提交任务的用户，因此，当前访问控制仅对安全模式下的Hive有效。
- 如果已经在实际业务逻辑中使用了“current_user()”函数，那么，在安全模式与普通模式互转时，需要充分评估可能的风险。

操作示例

- 不采用“current_user”函数，要实现不同的用户，访问不同数据，需要创建不同的视图：
  - 将视图v1授权给用户hiveuser1，hiveuser1用户可以访问表table1中“type='hiveuser1'”的数据：

    create view v1 as select * from table1 where type='hiveuser1'
  - 将视图v2授权给用户hiveuser2，hiveuser2用户可以访问表table1中“type='hiveuser2'”的数据：

    create view v2 as select * from table1 where type='hiveuser2'
- 采用“current_user”函数，则只需要创建一个视图：

  将视图v分别赋给用户hiveuser1、hiveuser2，当hiveuser1查询视图v时，“current_user()”被自动转化为hiveuser1，当hiveuser2查询视图v时，“current_user()”被自动转化为hiveuser2：

  create view v as select * from table1 where type=current_user()

12.10.14 配置创建临时函数是否需要 ADMIN 权限

操作场景

Hive开源社区版本创建临时函数需要用户具备ADMIN权限。

MRS Hive提供配置开关，默认为创建临时函数需要ADMIN权限，与开源社区版本保持一致。

用户可修改配置开关，实现创建临时函数不需要ADMIN权限。当该选项配置成false时，存在一定的安全风险。

说明

安全模式支持配置创建临时函数是否需要ADMIN权限功能，而普通模式不支持该功能。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > Hive > 配置 > 全部配置”。

步骤2 在搜索框中输入参数名称，搜索“hive.security.temporary.function.need.admin”，修改参数值为“true”或“false”，修改后重启所有HiveServer实例。

说明

- 选择“true”时，创建临时函数需要ADMIN权限，与开源社区版本保持一致。
- 选择“false”时，创建临时函数不需要ADMIN权限。

----结束

12.10.15 使用 Hive 读取关系型数据库数据

操作场景

Hive支持创建与其他关系型数据库关联的外表。该外表可以从关联到的关系型数据库中读取数据，并与Hive的其他表进行Join操作。

目前支持使用Hive读取数据的关系型数据库如下：

- DB2
- Oracle

说明

本章节适用于MRS 3.x及后续版本。

前提条件

已安装Hive客户端。

操作步骤

步骤1 以Hive客户端安装用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd *客户端安装目录*

例如安装目录为“/opt/client”，则执行以下命令：

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 集群认证模式是否为安全模式。

- 是，执行以下命令进行用户认证：
  kinit *Hive业务用户*
- 否，执行**步骤5**。

步骤5 执行以下命令，将需要关联的关系型数据库驱动Jar包上传到HDFS目录下。

hdfs dfs -put *Jar包所在目录 保存Jar包的HDFS目录*

例如将“/opt”目录下ORACLE驱动Jar包上传到HDFS的“/tmp”目录下，则执行如下命令。

hdfs dfs -put /opt/ojdbc6.jar /tmp

步骤6 按照如下示例，在Hive客户端创建关联关系型数据库的外表。

说明

如果是安全模式，建表的用户需要“ADMIN”权限，**ADD JAR**的路径请以实际路径为准。

```
-- 关联oracle linux6版本示例
-- 如果是安全模式，设置admin权限
set role admin;
-- 添加连接关系型数据库的驱动jar包,不同数据库有不同的驱动JAR
ADD JAR hdfs:///tmp/ojdbc6.jar;

CREATE EXTERNAL TABLE ora_test
-- hive表的列需比数据库返回结果多一列用于分页查询
(id STRING,rownum string)
STORED BY 'com.qubitproducts.hive.storage.jdbc.JdbcStorageHandler'
TBLPROPERTIES (
-- 关系型数据库类型
"qubit.sql.database.type" = "ORACLE",
-- 通过JDBC连接关系型数据库的url（不同数据库有不同的url格式）
"qubit.sql.jdbc.url" = "jdbc:oracle:thin:@//10.163.0.1:1521/mydb",
-- 关系型数据库驱动类名
"qubit.sql.jdbc.driver" = "oracle.jdbc.OracleDriver",
-- 在关系型数据库查询的sql语句,结果将返回hive表
"qubit.sql.query" = "select name from aaa",
-- hive表的列与关系型数据库表的列进行匹配（可忽略）
"qubit.sql.column.mapping" = "id=name",
-- 关系型数据库用户
"qubit.sql.dbcp.username" = "test",
-- 关系型数据库密码
"qubit.sql.dbcp.password" = "xxx");
```

----结束

12.10.16 Hive 支持的传统关系型数据库语法

概述

Hive支持如下传统关系型数据库语法：

- Grouping
- EXCEPT、INTERSECT

Grouping

语法简介：

- 当Group by语句带with rollup/cube选项时，Grouping才有意义。
- CUBE生成的结果集显示了所选列中值的所有组合的聚合。
- ROLLUP生成的结果集显示了所选列中值的某一层次结构的聚合。
- Grouping：当用CUBE或ROLLUP运算符添加行时，附加的列输出值为1；当所添加的行不是由CUBE或ROLLUP产生时，附加列值为0。

例如，Hive中有一张表“table_test”，表结构如下所示：

```
+----------------+-------------------+--+
| table_test.id  | table_test.value  |
```

```
+---------------+------------------+--+
| 1             | 10               |
| 1             | 15               |
| 2             | 20               |
| 2             | 5                |
| 2             | 13               |
+---------------+------------------+--+
```

执行如下语句：

select id,grouping(id),sum(value) from table_test group by id with rollup;

得到如下结果：

```
+-------+-----------------+------+--+
|  id   | groupingresult  | sum  |
+-------+-----------------+------+--+
| 1     | 0               | 25   |
| NULL  | 1               | 63   |
| 2     | 0               | 38   |
+-------+-----------------+------+--+
```

EXCEPT、INTERSECT

语法简介

- EXCEPT返回两个结果集的差（即从左查询中返回右查询没有找到的所有非重复值）。
- INTERSECT返回两个结果集的交集（即两个查询都返回的所有非重复值）。

例如，Hive中有两张表“test_table1”、“test_table2”。

“test_table1”表结构如下所示：

```
+-----------------+--+
| test_table1.id  |
+-----------------+--+
| 1               |
| 2               |
| 3               |
| 4               |
+-----------------+--+
```

“test_table2”表结构如下所示：

```
+-----------------+--+
| test_table2.id  |
+-----------------+--+
| 2               |
| 3               |
| 4               |
| 5               |
+-----------------+--+
```

- 执行如下的EXCEPT语句：

  select id from test_table1 except select id from test_table2;

  显示如下结果：

  ```
  +--------------+--+
  | _alias_0.id  |
  +--------------+--+
  | 1            |
  +--------------+--+
  ```

- 执行INTERSECT语句：

  select id from test_table1 intersect select id from test_table2;

显示如下结果：

```
+--------------+--+
| _alias_0.id  |
+--------------+--+
| 2            |
| 3            |
| 4            |
+--------------+--+
```

12.10.17 创建 Hive 用户自定义函数

当Hive的内置函数不能满足需要时，可以通过编写用户自定义函数UDF（User-Defined Functions）插入自己的处理代码并在查询中使用它们。

按实现方式，UDF分如下分类：

- 普通的UDF，用于操作单个数据行，且产生一个数据行作为输出。
- 用户定义聚集函数UDAF（User-Defined Aggregating Functions），用于接受多个输入数据行，并产生一个输出数据行。
- 用户定义表生成函数UDTF（User-Defined Table-Generating Functions），用于操作单个输入行，产生多个输出行。

按使用方法，UDF有如下分类：

- 临时函数，只能在当前会话使用，重启会话后需要重新创建。
- 永久函数，可以在多个会话中使用，不需要每次创建。

说明

用户自定义函数需要用户控制函数中变量的内存、线程等资源的占用，如果控制不当可能会导致内存溢出、CPU使用高等问题。

下面以编写一个AddDoublesUDF为例，说明UDF的编写和使用方法。

功能介绍

AddDoublesUDF主要用来对两个及多个浮点数进行相加，在该样例中可以掌握如何编写和使用UDF。

说明

- 一个普通UDF必须继承自“org.apache.hadoop.hive.ql.exec.UDF”。
- 一个普通UDF必须至少实现一个evaluate()方法，evaluate函数支持重载。
- 开发自定义函数需要在工程中添加“hive-exec-3.1.0.jar”依赖包，可从Hive服务的安装目录下获取。

样例代码

以下为UDF示例代码：

其中，*xxx*通常为程序开发的组织名称。

```
package com.xxx.bigdata.hive.example.udf;
import org.apache.hadoop.hive.ql.exec.UDF;

public class AddDoublesUDF extends UDF {
 public Double evaluate(Double... a) {
   Double total = 0.0;
```

```
    // 处理逻辑部分.
    for (int i = 0; i < a.length; i++)
      if (a[i] != null)
        total += a[i];
    return total;
  }
}
```

如何使用

步骤1 在客户端安装节点，把以上程序打包成AddDoublesUDF.jar，并上传到HDFS指定目录下（例如“/user/hive_examples_jars”）。

创建函数的用户与使用函数的用户都需要具有该文件的可读权限。

示例语句：

hdfs dfs -put ./hive_examples_jars /user/hive_examples_jars

hdfs dfs -chmod 777 /user/hive_examples_jars

步骤2 判断集群的认证模式。

- 安全模式，需要使用一个具有Hive管理权限的用户登录beeline客户端，执行如下命令：

  kinit *Hive业务用户*

  beeline

  set role admin;

- 普通模式，执行如下命令：

  beeline -n *Hive业务用户*

步骤3 在Hive Server中定义该函数，以下语句用于创建永久函数：

CREATE FUNCTION *addDoubles* **AS 'com.***xxx***.bigdata.hive.example.udf.AddDoublesUDF' using jar 'hdfs://hacluster/user/hive_examples_jars/AddDoublesUDF.jar';**

其中*addDoubles*是该函数的别名，用于SELECT查询中使用，*xxx*通常为程序开发的组织名称。

以下语句用于创建临时函数：

CREATE TEMPORARY FUNCTION addDoubles AS 'com.*xxx***.bigdata.hive.example.udf.AddDoublesUDF' using jar 'hdfs://hacluster/user/hive_examples_jars/AddDoublesUDF.jar';**

- *addDoubles*是该函数的别名，用于SELECT查询中使用。
- 关键字TEMPORARY说明该函数只在当前这个Hive Server的会话过程中定义使用。

步骤4 在Hive Server中使用该函数，执行SQL语句：

SELECT addDoubles(1,2,3);

说明

若重新连接客户端再使用函数出现[Error 10011]的错误，可执行**reload function;**命令后再使用该函数。

步骤5 在Hive Server中删除该函数，执行SQL语句：

DROP FUNCTION addDoubles;

----结束

扩展应用

无

12.10.18 beeline 可靠性增强特性介绍

操作场景

- 在批处理任务运行过程中，beeline客户端由于网络异常等问题断线时，Hive能支持beeline在断线前已经提交的任务继续运行。当再次运行该批处理任务时，已经提交过的任务不再重新执行，直接从下一个任务开始执行。
- 在批处理任务运行过程中，HiveServer服务由于某些原因导致宕机时，Hive能支持当再次运行该批处理任务时，已经成功执行完成的任务不再重新执行，直接从HiveServer2宕机时正在运行的任务开始运行。

说明

本章节适用于MRS 3.x及后续版本。

操作示例

1. beeline启动断线重连功能。
   示例：
   beeline -e "${*SQL*}" --hivevar batchid=*xxxxx*
2. beeline kill正在运行的任务。
   示例：
   beeline -e "" --hivevar batchid=*xxxxx* --hivevar kill=true
3. 登录beeline客户端，启动断线重连机制。
   登录beeline客户端后，执行“set hivevar:batchid=*xxxx*”

说明

使用说明：

- 其中“*xxxx*”表示每一次通过beeline提交任务的批次号，通过该批次号，可以识别出先提交的任务。如果提交任务时不带批次号，该特性功能不会启用。“*xxxx*”的值是执行任务时指定的，如下所示，“*xxxx*”值为“012345678901”：

  beeline -f hdfs://hacluster/user/hive/table.sql --hivevar batchid=012345678901

- 如果运行的SQL脚本依赖数据的失效性，建议不启用断点重连机制，或者每次运行时使用新的batchid。因为重复执行时，可能由于某些SQL语句已经执行过了不再重新执行，导致获取到过期的数据。
- 如果SQL脚本中使用了一些内置时间函数，建议不启用断点重连机制，或者每次运行时使用新的batchid，理由同上。
- 一个SQL脚本里面会包含一个或多个子任务。如果SQL脚本中存在先创建再删除临时表的逻辑，建议将删除临时表的逻辑放到脚本的最后。假定删除临时表子任务的后续子任务执行失败，并且删除临时表的子任务之前的子任务用到了该临时表；当下一次以相同batchid执行该SQL脚本时，因为临时表在上一次执行时已被删除，则会导致删除临时表的子任务之前用到该临时表的子任务（不包括创建该临时表的子任务，因为上一次已经执行成功，本次不会再执行，仅可编译）编译失败。这种情况下，建议使用新的batchid执行脚本。

参数说明：

- zk.cleanup.finished.job.interval：执行清理任务的间隔时间，默认隔60s执行一次。
- zk.cleanup.finished.job.outdated.threshold：节点的过期时间，每个批次的任务都会生成对应节点，从当前批次任务的结束时间开始算，如果超过60分钟，则表示已经过期了，那么就清除节点。
- batch.job.max.retry.count：单批次任务的最大重试次数，当单批次的任务失败重试次数超过这个值，就会删除该任务记录，下次运行时将从头开始运行，默认是10次。
- beeline.reconnect.zk.path：存储任务执行进度的根节点，Hive服务默认是/beeline。

12.10.19 具备表 select 权限可用 show create table 查看表结构

操作场景

此功能在MRS 3.x及后续版本适用于Hive，Spark2x。

开启此功能后，使用Hive建表时，其他用户被授予select权限后，可通过**show create table**查看表结构。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“HiveServer（角色） > 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.allow.show.create.table.in.select.nogrant”，“值”为“true”，修改后重启所有Hive实例。

步骤3 是否需要在Spark/Spark2x客户端中启用此功能?

- 是，重新下载并安装Spark/Spark2x客户端。
- 否，操作结束。

----结束

12.10.20 Hive 写目录旧数据进回收站

操作场景

此功能适用于Hive组件。

开启此功能后，执行写目录：**insert overwrite directory "***/path1***" ...**，写成功之后，会将旧数据移除到回收站，并且同时限制该目录不能为Hive元数据库中已经存在的数据库路径。

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“HiveServer（角色）> 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.overwrite.directory.move.trash”，“值”为“true”，修改后重启所有Hive实例。

----结束

12.10.21 Hive 能给一个不存在的目录插入数据

操作场景

此功能适用于Hive组件。

开启此功能后，在执行写目录：**insert overwrite directory** *“/path1/path2/path3”* ...时，其中“*/path1/path2*”目录权限为700且属主为当前用户，“*path3*”目录不存在，会自动创建“*path3*”目录，并写数据成功。

上述功能，在Hive参数“hive.server2.enable.doAs”为“true”时已经支持，本次增加当“hive.server2.enable.doAs”为“false”时的功能支持。

说明

本功能参数调整与**Hive写目录旧数据进回收站**添加的自定义参数相同。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“HiveServer（角色）> 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.overwrite.directory.move.trash”，“值”为“true”，修改后重启所有Hive实例。

----结束

12.10.22 限定仅 admin 用户能创建库和在 default 库建表

操作场景

此功能在MRS 3.x之前版本适用于Hive，Spark。在MRS 3.x及后续版本适用于Hive，Spark2x。

开启此功能后，仅有Hive管理员可以创建库和在default库中建表，其他用户需通过Hive管理员授权才可使用库。

说明

- 开启本功能之后，会限制普通用户新建库和在default库新建表。请充分考虑实际应用场景，再决定是否作出调整。
- 因为对执行用户做了限制，使用非管理员用户执行建库、表脚本迁移、重建元数据操作时需要特别注意，防止错误。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“HiveServer（角色）> 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.allow.only.admin.create”，“值”为“true”，修改后重启所有Hive实例。

步骤3 是否需要在Spark/Spark2x客户端中启用此功能?

- 是，执行**步骤4**。

- 否，操作结束。

步骤4 选择“SparkResource2x > 自定义”和“JDBCServer2x > 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.allow.only.admin.create”，“值”为“true”，修改后重启所有Spark2x实例。

步骤5 重新下载并安装Spark/Spark2x客户端。

----结束

12.10.23 限定创建 Hive 内部表不能指定 location

操作场景

此功能在MRS 3.x之前版本适用于Hive，Spark。在MRS 3.x及后续版本适用于Hive，Spark2x。

开启此功能后，在创建Hive内部表时，不能指定location。即表创建成功之后，表的location路径会被创建在当前默认warehouse目录下，不能被指定到其他目录。如果创建内部表时指定location，则创建失败。

说明

开启本功能之后，创建Hive内部表不能执行location。因为对建表语句做了限制，如果数据库中已存在建表时指向非当前默认warehouse目录的表，在执行建库、表脚本迁移、重建元数据操作时需要特别注意，防止错误。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“HiveServer（角色） > 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.internaltable.notallowlocation”，“值”为“true”，修改后重启所有Hive实例。

步骤3 是否需要在Spark/Spark2x客户端中启用此功能?

- 是，重新下载并安装Spark/Spark2x客户端。
- 否，操作结束。

----结束

12.10.24 允许在只读权限的目录建外表

操作场景

此功能在MRS 3.x之前版本适用于Hive，Spark。在MRS 3.x及后续版本适用于Hive，Spark2x。

开启此功能后，允许有目录读权限和执行权限的用户和用户组创建外部表，而不必检查用户是否为该目录的属主，并且禁止外表的location目录在当前默认warehouse目录下。同时在外表授权时，禁止更改其location目录对应的权限。

说明

开启本功能之后，外表功能变化大。请充分考虑实际应用场景，再决定是否作出调整。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“HiveServer（角色）> 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.restrict.create.grant.external.table”，“值”为“true”。

步骤3 选择“MetaStore（角色）> 自定义”，对参数文件“hivemetastore-site.xml”添加自定义参数，设置“名称”为“hive.restrict.create.grant.external.table”，“值”为“true”，修改后重启所有Hive实例。

步骤4 是否需要在Spark/Spark2x客户端中启用此功能?

- 是，重新下载并安装Spark/Spark2x客户端。
- 否，操作结束。

----结束

12.10.25 Hive 支持授权超过 32 个角色

操作场景

此功能适用于Hive。

因为操作系统用户组个数限制，导致Hive不能创建超过32个角色，开启此功能后，Hive将支持创建超过32个角色。

说明

- 开启本功能并对表库等授权后，对表库目录具有相同权限的角色将会用“|”合并。查询acl权限时，将显示合并后的结果，与开启该功能前的显示会有区别。此操作不可逆，请充分考虑实际应用场景，再决定是否作出调整。
- MRS 3.x及后续版本支持Ranger，如果当前组件使用了Ranger进行权限控制，需基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Hive的Ranger访问权限策略**。
- 开启此功能后，包括**owner**在内默认最大可支持512个角色，由MetaStore自定义参数“hive.supports.roles.max”控制，可考虑实际应用场景进行修改。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“MetaStore（角色）> 自定义”，对参数文件“hivemetastore-site.xml”添加自定义参数，设置“名称”为“hive.supports.over.32.roles”，“值”为“true”。

步骤3 选择“HiveServer（角色）> 自定义”，对参数文件“hive-site.xml”添加自定义参数，设置“名称”为“hive.supports.over.32.roles”，“值”为“true”，修改后重启所有Hive实例。

----结束

12.10.26 Hive 任务支持限定最大 map 数

操作场景

- 此功能适用于Hive。
- 此功能用于从服务端限定Hive任务的最大map数，避免HiveSever服务过载而引发的性能问题。

操作步骤

步骤1 进入Hive服务配置页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 选择“MetaStore（角色）> 自定义”，对参数文件“hivemetastore-site.xml”添加自定义参数，设置“名称”为“hive.mapreduce.per.task.max.splits”，“值”为具体设定值，一般尽量设置大，修改后重启所有Hive实例。

----结束

12.10.27 HiveServer 租约隔离使用

操作场景

- 此功能适用于Hive。
- 开启此功能可以限定指定用户访问指定节点上的HiveServer服务，实现对用户访问HiveServer服务的资源隔离。

说明

本章节适用于MRS 3.x及后续版本。

操作步骤

以对用户hiveuser设置租约隔离为例，选取Hive当前已有的或者新添加一个或者多个实例，此处选择已有的HiveServer实例：

步骤1 登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Hive > HiveServer”。

步骤3 在HiveServer列表里选择设置租约隔离的HiveServer，选择“HiveServer > 实例配置 > 全部配置”。

步骤4 在“全部配置”界面的右上角搜索“hive.server2.zookeeper.namespace”，“值”为具体设定值，比如为hiveserver2_zk。

步骤5 单击“保存”，在弹出对话框单击“确定”。

步骤6 选择“集群 > *待操作集群的名称* > 服务 > Hive”，选择“更多 > 重启服务”，输入密码开始重启服务。

步骤7 使用beeline -u 的方式登录客户端，执行以下命令：

beeline -u "jdbc:hive2:// *10.5.159.13* **:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=** *hiveserver2_zk* **;sasl.qop=auth-conf;auth=KERBEROS;principal=hive/hadoop.** ***<系统域名>*** **@** ***<系统域名>*** **"**

执行命令时将“10.5.159.13”替换为任意一个ZooKeeper实例的IP地址，查找方式为“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 实例”。

“zooKeeperNamespace=”后面的“hiveserver2_zk”为**步骤4**中参数“hive.server2.zookeeper.namespace”设置的具体设定值。

结果将只会登录到被设置租约隔离的HiveServer。

说明

- 开启本功能后，必须在登录时使用以上命令才可以访问这个被设置租约隔离的HiveServer。如果直接使用**beeline**命令登录客户端，将只会访问其他没有被设置租约隔离的HiveServer。
- 用户可登录FusionInsight Manager，选择“系统 > 权限 > 域和互信”，查看“本端域”参数，即为当前系统域名。“hive/hadoop.<系统域名>”为用户名，用户名所包含的系统域名所有字母为小写。

----结束

12.10.28 Hive 支持事务

操作场景

Hive在表以及分区级别支持事务，开启事务模式下，能够增量更新、删除、读取事务表，实现了对事务表操作的原子性、隔离性、一致性和永久性。

说明

本章节适用于MRS 3.x及后续版本。

事务特性介绍

事务（transaction）是一组单元化操作，这些操作要么都执行，要么都不执行，是一个不可分割的工作单位。事务的四个基本要素通常被称为ACID特性，分别为：

- 原子性（Atomicity）：一个事务是一个不可再分割的工作单位，事务中的所有操作要么都发生，要么都不发生。
- 一致性（Consistency）：事务开始之前和事务结束以后，数据库的完整性约束没有被破坏。
- 隔离性（Isolation）：多个事务并发访问，事务之间是隔离的，一个事务不影响其它事务运行效果。事务之间的影响有：脏读、不可重复读、幻读、丢失更新。
- 持久性（Durability）：在事务完成以后，该事务锁对数据库所做的更改将永久保存在数据库中。

事务执行特点：

- 一条语句可以写入多个分区或多个表。如果操作失败，则用户看不到部分写入或插入。即使频繁更改数据，仍然能够快速执行操作。
- Hive能够自动压缩ACID事务文件，而不会影响并发查询。当查询许多小分区文件时，自动压缩可提高查询性能和元数据占用量。
- 读取语义包括快照隔离。当读取操作开始时，Hive在逻辑上处于锁定仓库的状态。读操作不受操作期间发生的任何更改的影响。

锁机制

事务通过以下两点实现ACID特性：

- 预写日志（Write-ahead logging）保证原子性和持久性。
- 锁（locking）保证隔离性。

操作	持有锁类型
Insert overwrite	hive.txn.xlock.iow=true时持有排他锁， hive.txn.xlock.iow=false时持有半共享锁。
Insert	共享锁。执行该操作时能够对当前表或分区执行读写操作。
Update/delete	半共享锁。执行该操作时能够执行持有共享锁的操作，不能执行持有排他锁或半共享锁的操作。
Drop	排他锁。执行该操作时无法对当前表或分区执行其他任何操作。

说明

如果写操作中存在锁机制引发的冲突，优先持有锁的操作将成功，其他操作将失败。

操作步骤

开启事务

步骤1 登录FusionInsight Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > *待操作的集群* > 服务 > Hive > 配置 > 全部配置 > MetaStore（角色） > 事务”。

步骤2 将“metastore.compactor.initiator.on”设置为true。

步骤3 将“metastore.compactor.worker.threads”设置为大于0的正整数。

说明

“metastore.compactor.worker.threads”：在MetaStore上运行压缩程序工作线程个数。请根据实际业务设置合适的值，该值过小会引起事务压缩任务执行慢，过大会导致MetaStore执行性能变低。

步骤4 登录Hive客户端，执行命令开启以下参数，具体操作请参考**使用Hive客户端**。

set hive.support.concurrency=true;

set hive.exec.dynamic.partition.mode=nonstrict;

set hive.txn.manager=org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;

创建事务表

步骤5 执行以下命令创建事务表。

CREATE TABLE [IF NOT EXISTS] *[db_name.]table_name* **(***col_name data_type [COMMENT col_comment], ...***) [ROW FORMAT row_format] STORED AS orc ..... TBLPROPERTIES ('transactional'='true'[,'groupId'='group1' ...]);**

例如：

CREATE TABLE acidTbl (a int, b int) STORED AS ORC TBLPROPERTIES ('transactional'='true');

说明

- 当前事务仅支持orc格式。
- 不支持外表。
- 不支持sorted table。
- 创建事务表必须增加表属性**'transactional'='true'**。
- 只能在事务模式下读写事务表。

使用事务表

步骤6 执行命令使用事务表。以**acidTbl**表为例：

- 向已有事务表中插入数据。
  INSERT INTO acidTbl VALUES(1,1);
- 更新已有事务表
  UPDATE acidTbl SET b = 10 where a = 1;
  acidTbl内容变更为：

```
+------------+------------+
| acidtbl.a  | acidtbl.b  |
+------------+------------+
| 1          | 10         |
+------------+------------+
1 row selected (0.775 seconds)
```

- 合并新旧事务表：
  acidTbl_update表中已有数据：

```
+--------------------+--------------------+
| acidtbl_update.a   | acidtbl_update.b   |
+--------------------+--------------------+
| 1                  | 20                 |
| 2                  | 10                 |
+--------------------+--------------------+
2 rows selected (0.537 seconds)
```

  MERGE INTO acidTbl AS a
  USING acidTbl_update AS b ON a.a = b.a
  WHEN MATCHED THEN UPDATE SET b = b. b
  WHEN NOT MATCHED THEN INSERT VALUES (b.a, b.b);
  acidTbl内容变更为：

```
+------------+------------+
| acidtbl.a  | acidtbl.b  |
+------------+------------+
| 1          | 20         |
| 2          | 10         |
+------------+------------+
2 rows selected (0.666 seconds)
```

  说明

  执行merge命令时，如果出现“Error evaluating cardinality_violation”异常。请检查连接键是否有重复，或者执行**set hive.merge.cardinality.check=false;**命令用以规避。

- 删除事务表记录：
  DELETE FROM acidTbl where a = 2;

```
+------------+------------+
| acidtbl.a  | acidtbl.b  |
+------------+------------+
| 1          | 20         |
+------------+------------+
1 row selected (1.253 seconds)
```

查看事务执行状态

步骤7 执行以下命令查看事务执行状态。

- 查看锁：

  show locks;

- 查看压缩任务：

  show compactions;

- 查看事务执行状态：

  show transactions;

- 中断事务：

  abort transactions *TransactionId*;

  其中“TransactionId”即是执行**•查看事务执行状态：**命令后，结果中“Transaction ID”所在列的参数值。

----结束

配置压缩功能

HDFS不支持文件的就地更改，对于新增内容，它也不为用户提供读取的一致性。为了在HDFS上提供这些特性，遵循了在其他数据仓库工具中使用的标准方法：表或分区的数据存储在一组基本文件中，新增、更新和删除的记录存储在增量文件中。每个事务都创建一组新的增量文件以更改表或分区。在读取时，合并基础文件和增量文件并应用更新或删除的变化。

写事务表将在HDFS上产生部分小文件，Hive提供合并这些小文件的Major压缩和Minor压缩策略。

自动执行压缩操作步骤

步骤1 登录FusionInsight Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > *待操作的集群* > 服务 > Hive > 配置 > 全部配置 > MetaStore（角色） > 事务”。

步骤2 根据实际要求配置以下参数：

表 12-237 参数配置

参数	描述
hive.compactor.check.interval	压缩线程的执行间隔时间。单位：秒。默认值：300。
hive.compactor.cleaner.run.interval	清理线程的执行间隔时间。单位：毫秒。默认值：5000。
hive.compactor.delta.num.threshold	触发Minor压缩的增量文件个数阈值。默认值：10。
hive.compactor.delta.pct.threshold	触发Major压缩的增量文件（delta）大小总和占 base文件大小比例阈值，0.1表示delta文件大小之和与base文件大小之比为10%时触发Major压缩。 默认值：0.1。

参数	描述
hive.compactor.max.num.delta	压缩器将在单个作业中尝试处理的最大增量文件数。默认值：500。
metastore.compactor.initiator.on	是否在此MetaStore实例上运行启动程序线程和清理程序线程。开启事务值必须为true。默认值：false。
metastore.compactor.worker.threads	在MetaStore上运行多少个压缩程序工作线程。设置为0表示不执行压缩，使用事务必须在MetaStore服务的一个或多个实例上将此值设置为正数。单位：秒。默认值：0。

步骤3 登录Hive客户端，执行压缩，具体操作请参考**使用Hive客户端**。

```
CREATE TABLE table_name (
 id int, name string
)
CLUSTERED BY (id) INTO 2 BUCKETS STORED AS ORC
TBLPROPERTIES ("transactional"="true",
  "compactor.mapreduce.map.memory.mb"="2048",                  -- 指定紧缩map作业的属性
  "compactorthreshold.hive.compactor.delta.num.threshold"="4",  -- 如果有超过4个增量目录，则触发轻度紧缩
  "compactorthreshold.hive.compactor.delta.pct.threshold"="0.5" -- 如果增量文件的大小与基础文件的大小的比率大于50%，则触发深度紧缩
);
```

或

```
ALTER TABLE table_name COMPACT 'minor' WITH OVERWRITE TBLPROPERTIES
("compactor.mapreduce.map.memory.mb"="3072");  -- 指定紧缩map作业的属性
ALTER TABLE table_name COMPACT 'major' WITH OVERWRITE TBLPROPERTIES
("tblprops.orc.compress.size"="8192");         -- 更改任何其他Hive表属性
```

说明

执行压缩后小文件不会被立即删除，cleaner线程完成清理后文件被批量删除。

----结束

12.10.29 切换 Hive 执行引擎为 Tez

操作场景

Hive支持使用Tez引擎处理数据计算任务，用户在执行任务前可手动切换执行引擎为Tez。

前提条件

集群已安装Yarn服务的TimelineServer角色，且角色运行正常。

客户端切换执行引擎为 Tez

步骤1 安装并登录Hive客户端，具体操作请参考**使用Hive客户端**。

步骤2 执行以下命令切换引擎并开启“yarn.timeline-service.enabled”参数：

set hive.execution.engine=tez;

set yarn.timeline-service.enabled=true;

说明

- “yarn.timeline-service.enabled”参数开启后可以在Tez服务中通过TezUI查看Tez引擎执行任务的详细情况。开启后任务信息将上报TimelineServer，如果TimelineServer实例故障，会导致任务失败。
- 由于Tez使用ApplicationMaster缓冲池，“yarn.timeline-service.enabled”必须在提交Tez任务前开启，否则会导致此参数无法生效，需要重新登录客户端进行配置。
- 当执行引擎需要切换为其它引擎时，需要通过客户端执行**set yarn.timeline-service.enabled=false**命令关闭“yarn.timeline-service.enabled”参数。
- 如果需要指定Yarn运行队列，可以在客户端执行**set tez.queue.name=default**命令指定运行队列。

步骤3 提交并执行Tez任务。

步骤4 登录FusionInsight Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > *待操作的集群* > 服务 > Tez > TezUI（*主机名称*）”，在TezUI界面查看任务执行情况。

针对MRS 3.x之前版本，请登录MRS Manager界面，选择“服务管理 > Tez > Tez WebUI”，在TezUI界面查看任务执行情况。

----结束

切换 Hive 服务默认执行引擎为 Tez

步骤1 登录FusionInsight Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > *待操作的集群* > 服务 > Hive > 配置 > 全部配置 > HiveServer（角色）”，搜索“hive.execution.engine”参数。

针对MRS 3.x之前版本，请登录MRS Manager界面，选择“服务管理 > Hive > 服务配置 > 全部配置 > HiveServer”，搜索“hive.execution.engine”参数。

步骤2 将“hive.execution.engine”参数设置为“tez”。

步骤3 选择“Hive（服务） > 自定义”，搜索“yarn.site.customized.configs”。

步骤4 在“yarn.site.customized.configs”参数后添加自定义参数，名称为“yarn.timeline-service.enabled”，值为“true”。

说明

- “yarn.timeline-service.enabled”开启后可以在Tez服务中通过TezUI查看Tez引擎执行任务详细情况。开启后任务信息将上报TimelineServer，如果TimelineServer实例故障，会导致任务失败。
- 由于Tez使用ApplicationMaster缓冲池，“yarn.timeline-service.enabled”必须在提交Tez任务前开启，否则会导致此参数无法生效，需要重新登录客户端配置。
- 当执行引擎需要切换为其它引擎时，需要将自定义参数“yarn.timeline-service.enabled”的值设置为“false”。

步骤5 单击“保存”在弹出窗口单击“确定”。

针对MRS 3.x之前版本，请单击“保存配置”在弹出窗口单击“是”。

步骤6 选择“概览 > 更多 > 重启服务”，重启Hive服务，输入密码开始重启服务。

针对MRS 3.x之前版本，请在“服务状态”页签选择“更多 > 重启服务”，重启Hive服务。

步骤7 安装并登录Hive客户端，具体操作请参考**使用Hive客户端**。

步骤8 提交并执行Tez任务。

步骤9 登录FusionInsight Manager界面，选择“集群 > *待操作的集群* > 服务 > Tez > TezUI（*主机名称*）”，跳转TezUI界面查看任务执行情况。

针对MRS 3.x之前版本，请登录MRS Manager界面，选择“服务管理 > Tez > Tez WebUI”，在TezUI界面查看任务执行情况。

----结束

12.10.30 Hive 物化视图

Hive 物化视图介绍

Hive物化视图是基于Hive内部表的查询结果得到的特殊表，物化视图可以看做一张中间表，存储实际的数据，占用物理空间。物化视图赖以建立的这些表称为物化视图的基表。

物化视图主要用于预先计算并保存表连接或聚合等耗时较多的操作的结果。在执行查询时，可以将原本基于基表查询的查询语句重写成基于物化视图查询，这样就可以避免进行join、group by等耗时的操作，从而快速的得到结果。

说明

- 物化视图是特殊的表，存储实际的数据，占用物理空间。
- 删除基表之前必须先删除基于该基表所建立的物化视图。
- 物化视图创建语句是原子的，这意味着在填充所有查询结果之前，其他用户看不到物化视图。
- 不能基于物化视图的查询结果建立物化视图。
- 不能基于无表查询得到的查询结果建立物化视图。
- 不能对物化视图做增删改操作（即insert、update、delete、load、merge）。
- 能对物化视图做复杂查询操作，因其本质就是一张特殊的表。
- 当基表数据更新，需要手动对物化视图进行更新，否则物化视图将保留旧数据，即过期。
- 可通过describe语法查看基于acid表创建的物化视图是否过期。
- 基于非acid表创建的物化视图，无法通过descirbe语句查询物化视图是否过期。
- 创建物化视图只支持文件存储格式是“ORC”，并且支持事务（即“TBLPROPERTIES ('transactional'='true')”）的Hive内部表。

创建物化视图

语法

```
CREATE MATERIALIZED VIEW [IF NOT EXISTS] [db_name.]materialized_view_name
  [COMMENT materialized_view_comment]
  DISABLE REWRITE
    [ROW FORMAT row_format]
    [STORED AS file_format]
      | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (...)]
  ]
  [LOCATION hdfs_path]
  [TBLPROPERTIES (property_name=property_value, ...)]
AS
<query>;
```

说明

- 目前，物化视图文件格式支持：“PARQUET”、“TextFile”、“SequenceFile”、“RCfile”、“ORC”。如未在创建语句中使用“STORED AS”指定，则默认文件格式是ORC。
- 在同一Database下不可创建同名的物化视图，否则在新物化视图无法正常创建的同时，原物化视图的数据文件也会被新物化视图基于基表查询得到的数据文件覆盖，造成数据篡改（篡改后可通过重建物化视图进行恢复）。

案例

步骤1 登录Hive客户端，执行命令开启以下参数，具体操作请参考**使用Hive客户端**。

set hive.support.concurrency=true;

set hive.exec.dynamic.partition.mode=nonstrict;

set hive.txn.manager=org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;

步骤2 创建基表，插入数据。

```
create table tb_emp(
empno int,ename string,job string,mgr int,hiredate TIMESTAMP,sal float,comm float,deptno int
)stored as orc
tblproperties('transactional'='true');

insert into tb_emp values(7369, 'SMITH', 'CLERK',7902, '1980-12-17 08:30:09',800.00,NULL,20),
(7499, 'ALLEN', 'SALESMAN',7698, '1981-02-20 17:12:00',1600.00,300.00,30),
(7521, 'WARD', 'SALESMAN',7698, '1981-02-22 09:05:34',1250.00,500.00,30),
(7566, 'JONES', 'MANAGER', 7839, '1981-04-02 10:14:13',2975.00,NULL,20),
(7654, 'MARTIN', 'SALESMAN',7698, '1981-09-28 08:36:17',1250.00,1400.00,30),
(7698, 'BLAKE', 'MANAGER',7839, '1981-05-01 11:12:55',2850.00,NULL,30),
(7782, 'CLARK', 'MANAGER',7839, '1981-06-09 15:45:28',2450.00,NULL,10),
(7788, 'SCOTT', 'ANALYST',7566, '1987-04-19 14:05:34',3000.00,NULL,20),
(7839, 'KING', 'PRESIDENT',NULL, '1981-11-17 10:18:25',5000.00,NULL,10),
(7844, 'TURNER', 'SALESMAN',7698, '1981-09-08 09:05:34',1500.00,0.00,30),
(7876, 'ADAMS', 'CLERK',7788, '1987-05-23 15:07:44',1100.00,NULL,20),
(7900, 'JAMES', 'CLERK',7698, '1981-12-03 16:23:56',950.00,NULL,30),
(7902, 'FORD', 'ANALYST',7566, '1981-12-03 08:48:17',3000.00,NULL,20),
(7934, 'MILLER', 'CLERK',7782, '1982-01-23 11:45:29',1300.00,NULL,10);
```

步骤3 基于tb_emp的查询，创建物化视图。

```
create materialized view group_mv disable rewrite
row format serde 'org.apache.hadoop.hive.serde2.JsonSerDe'
stored as textfile
tblproperties('mv_content'='Total compensation of each department')
as select deptno,sum(sal) sum_sal from tb_emp group by deptno;
```

----结束

应用物化视图

将原本基于基表查询的查询语句重写成基于物化视图查询，从而达到提升查询效率的效果。

案例

现有查询语句如下：

select deptno,sum(sal) from tb_emp group by deptno having sum(sal)>10000;

基于所创建的物化视图，可将查询语句改写成：

select deptno, sum_sal from group_mv where sum_sal>10000;

查看物化视图

语法

SHOW MATERIALIZED VIEWS [IN database_name] ['identifier_with_wildcards’];

DESCRIBE [EXTENDED | FORMATTED] [db_name.]materialized_view_name;

案例

show materialized views;

describe formatted group_mv;

删除物化视图

语法

DROP MATERIALIZED VIEW [db_name.]materialized_view_name;

案例

drop materialized view group_mv;

重建物化视图

创建物化视图的时候，基表数据会填充到物化视图中，但是后续增删改基表的数据，这部分数据是不会自动的同步到物化视图中的。因此，在更新数据后，需要手动对视图进行重建。

语法

ALTER MATERIALIZED VIEW [db_name.]materialized_view_name REBUILD;

案例

alter materialized view group_mv rebuild;

说明

当基表数据更新，而物化视图的数据未更新，则默认物化视图的状态为过期。

基于事务表创建的物化视图，可以通过descirbe语句查看物化视图是否过期。其中“Outdated for Rewriting”值为“Yes”，表示过期，值为“No”,表示未过期。

12.10.31 Hive 日志介绍

日志描述

日志路径：Hive相关日志的默认存储路径为“/var/log/Bigdata/hive/角色名”，Hive1相关日志的默认存储路径为“/var/log/Bigdata/hive1/角色名”，以此类推。

- HiveServer：“/var/log/Bigdata/hive/hiveserver”（运行日志），“/var/log/Bigdata/audit/hive/hiveserver”（审计日志）。
- MetaStore：“/var/log/Bigdata/hive/metastore”（运行日志），“/var/log/Bigdata/audit/hive/metastore”（审计日志）。
- WebHCat：“/var/log/Bigdata/hive/webhcat”（运行日志），“/var/log/Bigdata/audit/hive/webhcat”（审计日志）。

日志归档规则：Hive的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过20MB的时候（此日志文件大小可进行配置），会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件，压缩文件保留个数和压缩文件阈值可以配置。

表 12-238 Hive 日志列表

日志类型	日志文件名	描述
运行日志	/hiveserver/hiveserver.out	HiveServer运行环境信息日志
	/hiveserver/hive.log	HiveServer进程的运行日志
	/hiveserver/hive-omm-<日期>-<PID>-gc.log.<编号>	HiveServer进程的GC日志
	/hiveserver/prestartDetail.log	HiveServer启动前的工作日志
	/hiveserver/check-serviceDetail.log	Hive服务启动是否成功的检查日志
	/hiveserver/cleanupDetail.log	HiveServer卸载的清理日志
	/hiveserver/startDetail.log	HiveServer进程启动日志
	/hiveserver/stopDetail.log	HiveServer进程停止日志
	/hiveserver/localtasklog/omm_<日期>_<任务ID>.log	Hive本地任务的运行日志
	/hiveserver/localtasklog/omm_<日期>_<任务ID>-gc.log.<编号>	Hive本地任务的GC日志
	/metastore/metastore.log	MetaStore进程的运行日志
	/metastore/hive-omm-<日期>-<PID>-gc.log.<编号>	MetaStore进程的GC日志
	/metastore/postinstallDetail.log	MetaStore安装后的工作日志
	/metastore/prestartDetail.log	MetaStore启动前的工作日志
	/metastore/cleanupDetail.log	MetaStore卸载的清理日志
	/metastore/startDetail.log	MetaStore进程启动日志
	/metastore/stopDetail.log	MetaStore进程停止日志
	/metastore/metastore.out	MetaStore运行环境信息日志

日志类型	日志文件名	描述
	/webhcat/webhcat-console.out	Webhcat进程启停正常日志
	/webhcat/webhcat-console-error.out	Webhcat进程启停异常日志
	/webhcat/prestartDetail.log	WebHCat启动前的工作日志
	/webhcat/cleanupDetail.log	Webhcat卸载时或安装前的清理日志
	/webhcat/hive-omm-<日期>-<PID>-gc.log.<编号>	WebHCat进程的GC日志
	/webhcat/webhcat.log	WebHCat进程的运行日志
审计日志	hive-audit.log hive-rangeraudit.log	HiveServer审计日志
	metastore-audit.log	MetaStore审计日志
	webhcat-audit.log	WebHCat审计日志
	jetty-<日期>.request.log	Jetty服务的请求日志

日志级别

Hive提供了如表12-239所示的日志级别。

运行日志的级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-239 日志级别

级别	描述
ERROR	ERROR表示系统运行的错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 参考修改集群服务配置参数，进入Hive服务“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别并保存。

说明

配置Hive日志级别后可立即生效，无需重启服务。

----结束

日志格式

Hive的日志格式如下所示：

表 12-240 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<LogLevel>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2014-11-05 09:45:01,242 | INFO | main | Starting hive metastore on port 21088 | org.apache.hadoop.hive.metastore.HiveMetaStore.main(HiveMetaStore.java:5198)
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<LogLevel>|<产生该日志的线程名字>|<User Name><User IP><Time><Operation><Resource><Result><Detail >|<日志事件的发生位置>	2018-12-24 12:16:25,319 | INFO | HiveServer2-Handler-Pool: Thread-185 | UserName=hive UserIP=10.153.2.204 Time=2018/12/24 12:16:25 Operation=CloseSession Result=SUCCESS Detail= | org.apache.hive.service.cli.thrift.ThriftCLIService.logAuditEvent(ThriftCLIService.java:434)

12.10.32 Hive 性能调优

12.10.32.1 建立表分区

操作场景

Hive在做Select查询时，一般会扫描整个表内容，会消耗较多时间去扫描不关注的数据。此时，可根据业务需求及其查询维度，建立合理的表分区，从而提高查询效率。

操作步骤

步骤1 MRS 3.x之前版本：

登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。选择“节点管理 > 节点名称”，进入弹性云服务器界面。单击“远程登录”按钮，完成Hive节点的登录。

MRS 3.x及后续版本：

以**root**用户登录已安装Hive客户端的节点。

步骤2 执行以下命令，进入客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤3 执行**source bigdata_env**命令，配置客户端环境变量。

步骤4 在客户端中执行如下命令，执行登录操作。

kinit *用户名*

步骤5 执行以下命令登录客户端工具。

beeline

步骤6 指定静态分区或者动态分区。

- 静态分区：

  静态分区是手动输入分区名称，在创建表时使用关键字**PARTITIONED BY**指定分区列名及数据类型。应用开发时，使用**ALTER TABLE ADD PARTITION**语句增加分区，以及使用**LOAD DATA INTO PARTITON**语句将数据加载到分区时，只能静态分区。

- 动态分区：通过查询命令，将结果插入到某个表的分区时，可以使用动态分区。

  动态分区通过在客户端工具执行如下命令来开启：

  set hive.exec.dynamic.partition=true;

  动态分区默认模式是strict，也就是必须至少指定一列为静态分区，在静态分区下建立动态子分区，可以通过如下设置来开启完全的动态分区：

  set hive.exec.dynamic.partition.mode=nonstrict;

  说明

  - 动态分区可能导致一个DML语句创建大量的分区，对应的创建大量新文件夹，对系统性能可能带来影响。
  - 在文件数量大的情况下，执行一个SQL语句启动时间较长，可以在执行SQL语句之前执行“set mapreduce.input.fileinputformat.list-status.num-threads = 100;”命令来缩短启动时间。“mapreduce.input.fileinputformat.list-status.num-threads”参数需要先添加到Hive的白名单才可设置。

----结束

12.10.32.2 Join 优化

操作场景

使用Join语句时，如果数据量大，可能造成命令执行速度和查询速度慢，此时可进行Join优化。

Join优化可分为以下方式：

- Map Join
- Sort Merge Bucket Map Join
- Join顺序优化

Map Join

Hive的Map Join适用于能够在内存中存放下的小表（指表大小小于25MB），通过“hive.mapjoin.smalltable.filesize”定义小表的大小，默认为25MB。

Map Join的方法有两种：

- 使用/*+ MAPJOIN(join_table) */。
- 执行语句前设置如下参数，当前版本中该值默认为true。

  set hive.auto.convert.join=true;

使用Map Join时没有Reduce任务，而是在Map任务前起了一个MapReduce Local Task，这个Task通过TableScan读取小表内容到本机，在本机以HashTable的形式保存并写入硬盘上传到DFS，并在distributed cache中保存，在Map Task中从本地磁盘或者distributed cache中读取小表内容直接与大表join得到结果并输出。

使用Map Join时需要注意小表不能过大，如果小表将内存基本用尽，会使整个系统性能下降甚至出现内存溢出的异常。

Sort Merge Bucket Map Join

使用Sort Merge Bucket Map Join必须满足以下2个条件：

- join的两张表都很大，内存中无法存放。
- 两张表都按照join key进行分桶（clustered by (column)）和排序（sorted by(column)），且两张表的分桶数正好是倍数关系。

通过如下设置，启用Sort Merge Bucket Map Join：

set hive.optimize.bucketmapjoin=true;

set hive.optimize.bucketmapjoin.sortedmerge=true;

这种Map Join也没有Reduce任务，是在Map任务前启动MapReduce Local Task，将小表内容按桶读取到本地，在本机保存多个桶的HashTable备份并写入HDFS，并保存在Distributed Cache中，在Map Task中从本地磁盘或者Distributed Cache中按桶一个一个读取小表内容，然后与大表做匹配直接得到结果并输出。

Join 顺序优化

当有3张及以上的表进行Join时，选择不同的Join顺序，执行时间存在较大差异。使用恰当的Join顺序可以有效缩短任务执行时间。

Join顺序原则：

- Join出来结果较小的组合，例如表数据量小或两张表Join后产生结果较少，优先执行。
- Join出来结果大的组合，例如表数据量大或两张表Join后产生结果较多，在后面执行。

例如，customer表的数据量最多，orders表和lineitem表优先Join可获得较少的中间结果。

原有的Join语句如下：

```
select
  l_orderkey,
```

```
  sum(l_extendedprice * (1 - l_discount)) as revenue,
  o_orderdate,
  o_shippriority
from
  customer,
  orders,
  lineitem
where
  c_mktsegment = 'BUILDING'
  and c_custkey = o_custkey
  and l_orderkey = o_orderkey
  and o_orderdate < '1995-03-22'
  and l_shipdate > '1995-03-22'
limit 10;
```

Join顺序优化后如下：

```
select
  l_orderkey,
  sum(l_extendedprice * (1 - l_discount)) as revenue,
  o_orderdate,
  o_shippriority
from
  orders,
  lineitem,
  customer
where
  c_mktsegment = 'BUILDING'
  and c_custkey = o_custkey
  and l_orderkey = o_orderkey
  and o_orderdate < '1995-03-22'
  and l_shipdate > '1995-03-22'
limit 10;
```

注意事项

Join数据倾斜问题

执行任务的时候，任务进度长时间维持在99%，这种现象叫数据倾斜。

数据倾斜是经常存在的，因为有少量的Reduce任务分配到的数据量和其他Reduce差异过大，导致大部分Reduce都已完成任务，但少量Reduce任务还没完成的情况。

解决数据倾斜的问题，可通过设置“set hive.optimize.skewjoin=true”并调整hive.skewjoin.key的大小。hive.skewjoin.key是指Reduce端接收到多少个key即认为数据是倾斜的，并自动分发到多个Reduce。

12.10.32.3 Group By 优化

操作场景

优化Group by语句，可提升命令执行速度和查询速度。

Group by的时候， Map端会先进行分组， 分组完后分发到Reduce端， Reduce端再进行分组。可采用Map端聚合的方式来进行Group by优化，开启Map端初步聚合，减少Map的输出数据量。

操作步骤

在Hive客户端进行如下设置：

set hive.map.aggr=true;

注意事项

Group By数据倾斜

Group By也同样存在数据倾斜的问题，设置hive.groupby.skewindata为true，生成的查询计划会有两个MapReduce Job，第一个Job的Map输出结果会随机的分布到Reduce中，每个Reduce做聚合操作，并输出结果，这样的处理会使相同的Group By Key可能被分发到不同的Reduce中，从而达到负载均衡，第二个Job再根据预处理的结果按照Group By Key分发到Reduce中完成最终的聚合操作。

Count Distinct聚合问题

当使用聚合函数count distinct完成去重计数时，处理值为空的情况会使Reduce产生很严重的数据倾斜，可以将空值单独处理，如果是计算count distinct，可以通过where字句将该值排除掉，并在最后的count distinct结果中加1。如果还有其他计算，可以先将值为空的记录单独处理，再和其他计算结果合并。

12.10.32.4 数据存储优化

操作场景

“ORC”是一种高效的列存储格式，在压缩比和读取效率上优于其他文件格式。

建议使用“ORC”作为Hive表默认的存储格式。

前提条件

已登录Hive客户端，具体操作请参见**使用Hive客户端**。

操作步骤

- 推荐：使用“SNAPPY”压缩，适用于压缩比和读取效率要求均衡场景。

  Create table *xx (col_name data_type)* stored as orc tblproperties ("orc.compress"="SNAPPY");
- 可用：使用“ZLIB”压缩，适用于压缩比要求较高场景。

  Create table *xx (col_name data_type)* stored as orc tblproperties ("orc.compress"="ZLIB");

说明

xx为具体使用的Hive表名。

12.10.32.5 SQL 优化

操作场景

在Hive上执行SQL语句查询时，如果语句中存在“(a&b) or (a&c)”逻辑时，建议将逻辑改为“a & (b or c)”。

样例

假设条件a为“p_partkey = l_partkey”，优化前样例如下所示：

```
select
    sum(l_extendedprice* (1 - l_discount)) as revenue
```

```
from
	lineitem,
	part
where
	(
		p_partkey = l_partkey
		and p_brand = 'Brand#32'
		and p_container in ('SM CASE', 'SM BOX', 'SM PACK', 'SM PKG')
		and l_quantity >= 7 and l_quantity <= 7 + 10
		and p_size between 1 and 5
		and l_shipmode in ('AIR', 'AIR REG')
		and l_shipinstruct = 'DELIVER IN PERSON'
	)
	or
	(	p_partkey = l_partkey
		and p_brand = 'Brand#35'
		and p_container in ('MED BAG', 'MED BOX', 'MED PKG', 'MED PACK')
		and l_quantity >= 15 and l_quantity <= 15 + 10
		and p_size between 1 and 10
		and l_shipmode in ('AIR', 'AIR REG')
		and l_shipinstruct = 'DELIVER IN PERSON'
	)
	or
	(	p_partkey = l_partkey
		and p_brand = 'Brand#24'
		and p_container in ('LG CASE', 'LG BOX', 'LG PACK', 'LG PKG')
		and l_quantity >= 26 and l_quantity <= 26 + 10
		and p_size between 1 and 15
		and l_shipmode in ('AIR', 'AIR REG')
		and l_shipinstruct = 'DELIVER IN PERSON'
	)
```

优化后样例如下所示：

```
select
	sum(l_extendedprice* (1 - l_discount)) as revenue
from
	lineitem,
	part
where	p_partkey = l_partkey and
	((
		p_brand = 'Brand#32'
		and p_container in ('SM CASE', 'SM BOX', 'SM PACK', 'SM PKG')
		and l_quantity >= 7 and l_quantity <= 7 + 10
		and p_size between 1 and 5
		and l_shipmode in ('AIR', 'AIR REG')
		and l_shipinstruct = 'DELIVER IN PERSON'
	)
	or
	(
		p_brand = 'Brand#35'
		and p_container in ('MED BAG', 'MED BOX', 'MED PKG', 'MED PACK')
		and l_quantity >= 15 and l_quantity <= 15 + 10
		and p_size between 1 and 10
		and l_shipmode in ('AIR', 'AIR REG')
		and l_shipinstruct = 'DELIVER IN PERSON'
	)
	or
	(
		p_brand = 'Brand#24'
		and p_container in ('LG CASE', 'LG BOX', 'LG PACK', 'LG PKG')
		and l_quantity >= 26 and l_quantity <= 26 + 10
		and p_size between 1 and 15
		and l_shipmode in ('AIR', 'AIR REG')
		and l_shipinstruct = 'DELIVER IN PERSON'
	))
```

12.10.32.6 使用 Hive CBO 优化查询

操作场景

在Hive中执行多表Join时，Hive支持开启CBO（Cost Based Optimization），系统会自动根据表的统计信息，例如数据量、文件数等，选出合适计划提高多表Join的效率。Hive需要先收集表的统计信息后才能使CBO正确的优化。

说明

- CBO优化器会基于统计信息和查询条件，尽可能地使join顺序达到更优。但是也可能存在特殊情况导致join顺序调整不准确。例如数据存在倾斜，以及查询条件值在表中不存在等场景，可能调整出非优化的join顺序。
- 开启列统计信息自动收集时，需要在reduce侧做聚合统计。对于没有reduce阶段的insert任务，将会多出reduce阶段，用于收集统计信息。
- 本章节适用于MRS 3.x及后续版本。

前提条件

已登录Hive客户端，具体操作请参见**使用Hive客户端**。

操作步骤

步骤1 在Manager界面Hive组件的配置中搜索“hive.cbo.enable”参数，选中“true”永久开启功能。

步骤2 手动收集Hive表已有数据的统计信息。

执行以下命令，可以手动收集统计信息。仅支持统计一张表，如果需要统计不同的表需重复执行。

ANALYZE TABLE [db_name.]tablename [PARTITION(partcol1[=val1], partcol2[=val2], ...)]

COMPUTE STATISTICS

[FOR COLUMNS]

[NOSCAN];

说明

- 指定FOR COLUMNS时，收集列级别的统计信息。
- 指定NOSCAN时，将只统计文件大小和个数，不扫描具体文件。

例如：

analyze table table_name compute statistics;

analyze table table_name compute statistics for columns;

步骤3 配置Hive自动收集统计信息。开启配置后，执行**insert overwrite/into**命令插入数据时才自动统计新数据的信息。

- 在Hive客户端执行以下命令临时开启收集：

  set hive.stats.autogather = true;开启表/分区级别的统计信息自动收集。

  set hive.stats.column.autogather = true; 开启列级别的统计信息自动收集。

说明

- 列级别统计信息的收集不支持复杂的数据类型，例如Map，Struct等。
- 表级别统计信息的自动收集不支持Hive on HBase表。

- 在Manager界面Hive的服务配置中，搜索参数“hive.stats.autogather”和“hive.stats.column.autogather”，选中“true”永久开启收集功能。

步骤4 执行以下命令可以查看统计信息。

DESCRIBE FORMATTED table_name[.column_name] PARTITION partition_spec;

例如：

desc formatted table_name;

desc formatted table_name id;

desc formatted table_name partition(time='2016-05-27');

说明

分区表仅支持分区级别的统计信息收集，因此分区表需要指定分区来查询统计信息。

----结束

12.10.33 Hive 常见问题

12.10.33.1 如何在多个 HiveServer 之间同步删除 UDF

问题

如果需要删除永久函数（Permanent UDF），如何在多个HiveServer之间同步删除?

回答

因为多个HiveServer之间共用一个MetaStore存储数据库，所以MetaStore存储数据库和HiveServer的内存之间数据同步有延迟。如果在单个HiveServer上删除永久函数，操作结果将无法同步到其他HiveServer上。

遇到如上情况，需要登录Hive客户端，连接到每个HiveServer，并分别删除永久函数。具体操作如下：

步骤1 以Hive客户端安装用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd *客户端安装目录*

例如安装目录为“/opt/client”，则执行以下命令：

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 执行以下命令进行用户认证。

kinit *Hive业务用户*

说明

登录的用户需具备Hive admin权限。

步骤5 执行如下命令，连接指定的HiveServer。

beeline -u "jdbc:hive2:// *10.39.151.74*:*21066*/**default;sasl.qop=auth-conf;auth=KERBEROS;principal=***hive*/*hadoop.<系统域名>*@*<系统域名>*"

说明

- *10.39.151.74*为HiveServer所在节点的IP地址。
- *21066*为HiveServer端口。HiveServer端口默认范围为21066～21070，用户需根据实际配置端口进行修改。
- *hive*为用户名。例如，使用Hive1实例时，则使用hive1。
- 用户可登录FusionInsight Manager，选择“系统 > 权限 > 域和互信”，查看“本端域”参数，即为当前系统域名。
- “hive/hadoop.*<系统域名>*”为用户名，用户的用户名所包含的系统域名所有字母为小写。

步骤6 执行如下命令，启用Hive admin权限。

set role admin;

步骤7 执行如下命令，删除永久函数。

drop function *function_name*;

说明

- *function_name*为永久函数的函数名。
- 如果永久函数是在Spark中创建的，在Spark中删除该函数后需要在HiveServer中删除，即执行上述删除命令。

步骤8 确定是否已连接所有HiveServer并删除永久函数。

- 是，操作完毕。
- 否，执行**步骤5**。

----结束

12.10.33.2 已备份的 Hive 表无法执行 drop 操作

问题

为什么已备份的Hive表执行drop操作会失败?

回答

由于已备份Hive表对应的HDFS目录创建了快照，导致HDFS目录无法删除，造成Hive表删除失败。

Hive表在执行备份操作时，会创建表对应的HDFS数据目录快照。而HDFS的快照机制有一个约束：如果一个HDFS目录已创建快照，则在快照完全删除之前，该目录无法删

除或修改名称。Hive表（除EXTERNAL表外）执行drop操作时，会尝试删除该表对应的HDFS数据目录，如果目录删除失败，系统会提示表删除失败。

如果确实需要删除该表，可手动删除涉及到该表的所有备份任务。

12.10.33.3 如何在 Hive 自定义函数中操作本地文件

问题

在Hive自定义函数中需要操作本地文件，例如读取文件的内容，需要如何操作?

回答

默认情况下，可以在UDF中用文件的相对路径来操作文件，如下示例代码：

```
public String evaluate(String text) {
  // some logic
  File file = new File("foo.txt");
  // some logic
  // do return here
}
```

在Hive中使用时，将UDF中用到的文件“foo.txt”上传到HDFS上，如上传到“hdfs://hacluster/tmp/foo.txt”，使用以下语句创建UDF，在UDF中就可以直接操作“foo.txt”文件了：

create function testFunc as 'some.class' using jar 'hdfs://hacluster/somejar.jar', file 'hdfs://hacluster/tmp/foo.txt';

例外情况下，如果“hive.fetch.task.conversion”参数的值为“more”，在UDF中不能再使用相对路径来操作文件，而要使用绝对路径，并且保证所有的HiveServer节点和NodeManager节点上该文件是存在的且**omm**用户对该文件有相应的权限，才能正常在UDF中操作本地文件。

12.10.33.4 如何强制停止 Hive 执行的 MapReduce 任务

问题

在Hive执行MapReduce任务长时间卡住的情况下想手动停止任务，需要如何操作?

回答

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作的集群名称* > 服务 > Yarn”。

步骤3 单击左侧页面的“ResourceManager(*主机名称*，主)”按钮，登录Yarn界面。

步骤4 单击对应任务ID的按钮进入任务页面，单击界面左上角的“Kill Application”按钮，在弹框中单击“确认”停止任务。

----结束

12.10.33.5 如何对 Hive 表大小数据进行监控

问题

如何对Hive中的表大小数据进行监控?

回答

当用户要对Hive表大小数据进行监控时，可以通过HDFS的精细化监控对指定表目录进行监控，从而到达监控指定表大小数据的目的。

前提条件

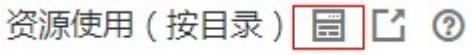
- Hive、HDFS组件功能正常
- HDFS精细化监控功能正常

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 通过“集群 > *待操作集群的名称* > 服务 > HDFS > 资源 ”，进入HDFS精细化页面。

步骤3 找到“资源使用（按目录）”监控项，单击该监控项左上角第一个图标。



步骤4 进入配置空间监控子页面，单击“添加”。

步骤5 在名称空格中填写监控的表名称（或其他用户自定义的别名），在路径中填写需要监控表的路径。单击“确定”。该监控的横坐标为时间，纵坐标为监控目录的大小。

----结束

12.10.33.6 如何对重点目录进行保护，防止“insert overwrite”语句误操作导致数据丢失

问题

如何对重点目录进行保护，防止“insert overwrite”语句误操作导致数据丢失?

回答

当用户要对Hive重点数据库、表或目录进行监控，防止“insert overwrite”语句误操作导致数据丢失时，可以利用Hive配置中的“hive.local.dir.confblacklist”进行目录保护。

该配置项已对“/opt/”，“/user/hive/warehouse”等目录进行了默认配置。

前提条件

Hive、HDFS组件功能正常。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置 > 全部配置”，搜索“hive.local.dir.confblacklist”配置项。

步骤3 在该配置项中添加用户要重点保护的数据库、表或目录路径。

步骤4 输入完成后，单击“保存”，保存配置项。

----结束

12.10.33.7 未安装 HBase 时 Hive on Spark 任务卡顿处理

操作场景

此功能适用于Hive组件。

按如下操作步骤设置参数后，在未安装HBase的环境执行Hive on Spark任务时，可避免任务卡顿。

说明

Hive on Spark任务的Spark内核版本已经升级到Spark2x，可以支持在不安装Spark2x的情况下，执行Hive on Spark任务。如果没有安装HBase，默认在执行Spark任务时，会尝试去连接Zookeeper访问HBase，直到超时，这样会造成任务卡顿。

在未安装HBase的环境，要执行Hive on Spark任务，可以按如下操作处理。如果是从已有HBase低版本环境升级上来的，升级完成之后可不进行设置。

操作步骤

步骤1 登录FusionInsight Manager 。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置 > 全部配置”。

步骤3 选择“HiveServer（角色） > 自定义”，对参数文件“spark-defaults.conf”添加自定义参数，设置“名称”为“spark.security.credentials.hbase.enabled”，“值”为“false” 。

步骤4 单击“保存”，在弹出对话框单击“确定”。

步骤5 选择“集群 > *待操作集群的名称* > 服务 > Hive > 实例”，勾选所有Hive实例，选择“更多 > 重启实例”，输入密码，单击“确定”。

----结束

12.10.33.8 FusionInsight Hive 使用 WHERE 条件查询超过 3.2 万分区的表报错

问题

Hive创建超过3.2万分区的表，执行带有WHERE分区的条件查询时出现异常，且“metastore.log”中打印的异常信息包含以下信息：

```
Caused by: java.io.IOException: Tried to send an out-of-range integer as a 2-byte value: 32970
        at org.postgresql.core.PGStream.SendInteger2(PGStream.java:199)
        at org.postgresql.core.v3.QueryExecutorImpl.sendParse(QueryExecutorImpl.java:1330)
        at org.postgresql.core.v3.QueryExecutorImpl.sendOneQuery(QueryExecutorImpl.java:1601)
```

```
at org.postgresql.core.v3.QueryExecutorImpl.sendParse(QueryExecutorImpl.java:1191)
at org.postgresql.core.v3.QueryExecutorImpl.execute(QueryExecutorImpl.java:346)
```

回答

带有分区条件的查询，Hiveserver会对分区进行优化，避免全表扫描，需要查询元数据符合条件的所有分区，而gaussDB中提供的接口sendOneQuery，调用的sendParse方法中对参数的限制为32767，如果分区条件数超过32767就异常。

12.10.33.9 使用 IBM 的 jdk 访问 Beeline 客户端出现连接 hiveserver 失败

操作场景

查看客户端使用的jdk版本，如果是IBM JDK，则需要对Beeline客户端进行改造，否则会造成连接hiveserver失败。

操作步骤

步骤1 登录FusionInsight Manager 页面，选择“系统 > 权限 > 用户”，在待操作用户的“操作”栏下选择“更多 > 下载认证凭据”，选择集群信息后单击“确定”，下载keytab文件。

步骤2 解压keytab文件，使用WinSCP工具将解压得到的“user.keytab”文件上传到待操作节点的Hive客户端安装目录下，例如：“/opt/client”。

步骤3 使用以下命令打开hive客户端目录下面的配置文件Hive/component_env：

vi *Hive客户端安装目录*/**Hive/component_env**

在变量“export CLIENT_HIVE_URI”所在行后面添加如下内容：

```
\;user.principal=用户名@HADOOP.COM\;user.keytab=user.keytab文件所在路径/user.keytab
```

----结束

12.10.33.10 关于 Hive 表的 location 支持跨 OBS 和 HDFS 路径的说明

问题

Hive表的location支持跨OBS和HDFS路径吗?

回答

1. Hive存储在OBS上的普通表，支持表location配置为hdfs路径。
2. 同一个Hive服务中可以分别创建存储在OBS上的表和存储在HDFS上的表。
3. Hive存储在OBS上的分区表，不支持将分区location配置为hdfs路径（存储在HDFS上的分区表也不支持修改分区location为OBS）。

12.10.33.11 通过 Tez 引擎执行 union 相关语句写入的数据，切换 MR 引擎后查询不出来。

问题

Hive通过Tez引擎执行union相关语句写入的数据，切换到Mapreduce引擎后进行查询，发现数据没有查询出来。

回答

由于Hive使用Tez引擎在执行union语句时，生成的输出文件会存在HIVE_UNION_SUBDIR目录，切回Mapreduce引擎后默认不读取目录下的文件，所以没有读取到HIVE_UNION_SUBDIR目录下的数据。

此时可以设置参数**set mapreduce.input.fileinputformat.input.dir.recursive=true**，开启union优化，决定是否读取目录下的数据。

12.10.33.12 Hive 不支持对同一张表或分区进行并发写数据

问题

为什么通过接口并发对Hive表进行写数据会导致数据不一致?

回答

Hive不支持对同一张表或同一个分区进行并发数据插入，这样会导致多个任务操作同一个数据临时目录，一个任务将另一个任务的数据移走，导致任务数据异常。解决方法是修改业务逻辑，单线程插入数据到同一张表或同一个分区。

12.10.33.13 Hive 不支持向量化查询

问题

当设置向量化参数hive.vectorized.execution.enabled=true时，为什么执行hive on Tez/Mapreduce/Spark时会偶现一些空指针或类型转化异常?

回答

当前Hive不支持向量化执行，向量化执行有很多社区问题引入目前没有稳定修复，默认hive.vectorized.execution.enabled=false，不建议将次参数打开。

12.10.33.14 Hive 表 HDFS 数据目录被误删，但是元数据仍然存在，导致执行任务报错处理

问题

Hive表HDFS数据目录被误删，但是元数据仍然存在，导致执行任务报错。

回答

这是一种误操作的异常情况，需要手动删除对应表的元数据后重试。

例如：

执行以下命令进入控制台：

source ${BIGDATA_HOME}/FusionInsight_BASE_8.1.0.1/install/FusionInsight-dbservice-2.7.0/.dbservice_profile

gsql -p 20051 -U hive -d hivemeta -W HiveUser@

执行 **delete from tbls where tbl_id='xxx';**

12.10.33.15 如何关闭 Hive 客户端日志

问题

如何关闭Hive客户端的运行日志?

回答

步骤1 使用**root**用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 根据集群认证模式，完成Hive客户端登录。

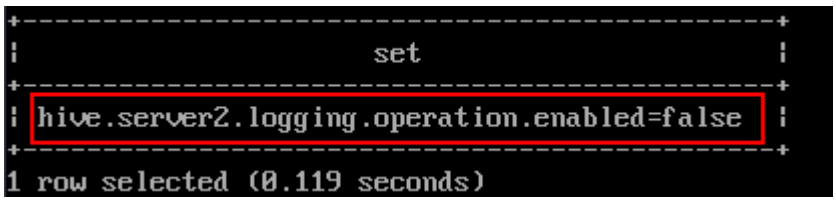
- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。
  kinit *组件业务用户*
  beeline
- 普通模式，则执行以下命令，登录Hive客户端。
  - 使用指定组件业务用户登录Hive客户端。
    beeline -n *组件业务用户*
  - 不指定组件业务用户登录Hive客户端，则会以当前操作系统用户登录。
    beeline

步骤5 执行以下命令关闭客户端日志：

set hive.server2.logging.operation.enabled=false;

步骤6 执行以下命令查看客户端日志是否已关闭，如下图所示即为关闭成功。

set hive.server2.logging.operation.enabled;



----结束

12.10.33.16 Hive 快删目录配置类问题

问题

在配置MRS多用户访问OBS细粒度权限的场景中，在Hive自定义配置中添加OBS快删目录的配置后，删除Hive表，执行结果为成功，但是OBS目录没有删掉。

回答

由于没有给用户配置快删目录的权限，导致数据不能被删除。需要修改用户对应的委托的IAM自定义策略，在策略内容上，配置Hive快删目录的权限。

12.10.33.17 Hive 配置类问题

- Hive SQL执行报错：java.lang.OutOfMemoryError: Java heap space.

  解决方案：
  - 对于MapReduce任务，增大下列参数：

    set mapreduce.map.memory.mb=8192;

    set mapreduce.map.java.opts=-Xmx6554M;

    set mapreduce.reduce.memory.mb=8192;

    set mapreduce.reduce.java.opts=-Xmx6554M;
  - 对于Tez任务，增大下列参数：

    set hive.tez.container.size=8192;
- Hive SQL对列名as为新列名后，使用原列名编译报错：Invalid table alias or column reference 'xxx'.

  解决方案：**set hive.cbo.enable=true;**
- Hive SQL子查询编译报错：Unsupported SubQuery Expression 'xxx': Only SubQuery expressions that are top level conjuncts are allowed.

  解决方案：**set hive.cbo.enable=true;**
- Hive SQL子查询编译报错：CalciteSubquerySemanticException [Error 10249]: Unsupported SubQuery Expression Currently SubQuery expressions are only allowed as Where and Having Clause predicates.

  解决方案：**set hive.cbo.enable=true;**
- Hive SQL编译报错：Error running query: java.lang.AssertionError: Cannot add expression of different type to set.

  解决方案：**set hive.cbo.enable=false;**
- Hive SQL执行报错：java.lang.NullPointerException at org.apache.hadoop.hive.ql.udf.generic.GenericUDAFComputeStats $GenericUDAFNumericStatsEvaluator.init.

  解决方案：**set hive.map.aggr=false;**
- Hive SQL设置hive.auto.convert.join = true（默认开启）和 hive.optimize.skewjoin=true执行报错：ClassCastException org.apache.hadoop.hive.ql.plan.ConditionalWork cannot be cast to org.apache.hadoop.hive.ql.plan.MapredWork.

  解决方案：**set hive.optimize.skewjoin=false;**

- Hive SQL设置hive.auto.convert.join=true（默认开启）、hive.optimize.skewjoin=true和hive.exec.parallel=true执行报错：java.io.FileNotFoundException: File does not exist:xxx/reduce.xml.

  解决方案：
  - 方法一：切换执行引擎为Tez，详情请参考**切换Hive执行引擎为Tez**。
  - 方法二：**set hive.exec.parallel=false;**
  - 方法三：**set hive.auto.convert.join=false;**
- Hive on Tez执行Bucket表Join报错：NullPointerException at org.apache.hadoop.hive.ql.exec.CommonMergeJoinOperator.mergeJoinComputeKeys

  解决方案：**set tez.am.container.reuse.enabled=false;**

12.11 使用 Hue（MRS 3.x 之前版本）

12.11.1 从零开始使用 Hue

Hue提供了文件浏览器功能，使用户可以通过界面图形化的方式查看Hive上文件及目录功能。

前提条件

已安装Hive以及Hue组件，且状态为运行中的Kerberos认证的集群。

操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 打开Hue WebUI，然后选择“Query Editors > Hive”。

步骤3 在“Databases”选择一个Hive中的数据库，默认数据库为“default”。

系统将自动显示数据库中的所有表。可以输入表名关键字，系统会自动搜索包含此关键字的全部表。

步骤4 单击指定的表名，可以显示表中所有的列。

步骤5 在HiveQL语句编辑区输入HiveQL语句。

create table hue_table(id int,name string,company string) row format delimited fields terminated by ',' stored as textfile;

单击 并选择“Explain”，编辑器将分析输入的HiveQL语句是否有语法错误以及执行计划，如果存在语法错误则显示“Error while compiling statement”。

步骤6 单击 ，选择HiveQL语句执行的引擎。

步骤7 单击 开始执行HiveQL语句。

步骤8 在命令输入框内输入**show tables;**，单击 按钮，查看结果中有**步骤5**创建的表hue_table。

----结束

12.11.2 访问 Hue 的 WebUI

操作场景

MRS集群安装Hue组件后，用户可以通过Hue的WebUI，在图形化界面使用Hadoop与Hive。

该任务指导用户在MRS集群中打开Hue的WebUI。

说明

Internet Explorer浏览器可能存在兼容性问题，建议更换兼容的浏览器访问Hue WebUI，例如Google Chrome浏览器50版本。

对系统的影响

第一次访问Manager和Hue WebUI，需要在浏览器中添加站点信任以继续打开Hue WebUI。

前提条件

启用Kerberos认证时，MRS集群管理员已分配用户使用Hive的权限。例如创建一个“人机”用户“hueuser”，并加入“hive”、“hadoop”、“supergroup”组和“System_administrator”角色，主组为“hive”。

该用户用于登录Hue WebUI。

操作步骤

步骤1 登录服务页面：单击集群名称，登录集群详情页面，选择“组件管理”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 选择“Hue”，在“Hue WebUI”右侧，单击链接，打开Hue的WebUI，以创建的“hueuser”用户登录Hue WebUI。

Hue的WebUI支持以下功能：

- 使用“Query Editors”执行Hive的查询语句。需要MRS集群已安装Hive。
- 使用“Data Browsers”管理Hive中的表。需要MRS集群已安装Hive。
- 使用  查看HDFS中的目录和文件。需要MRS集群已安装HDFS。
- 使用  查看MRS集群中所有作业。需要MRS集群已安装YARN。

说明

- 使用创建的用户第一次登录Hue WebUI，需修改密码。
- 用户获取Hue WebUI的访问地址后，可以给其他无法访问Manager的用户用于访问Hue WebUI。
- 在Hue的WebUI操作但不操作Manager页面，重新访问Manager时需要输入已登录的帐号密码。

----结束

12.11.3 Hue 常用参数

参数入口

参数入口，请参考**修改集群服务配置参数**。

参数说明

表 12-241 Hue 常用参数

配置参数	说明	缺省值	范围
HANDLER_ACCESSLOG_LEVEL	表示Hue的访问日志级别。	DEBUG	• ERROR • WARN • INFO • DEBUG
HANDLER_AUDITSLOG_LEVEL	表示Hue的审计日志级别。	DEBUG	• ERROR • WARN • INFO • DEBUG
HANDLER_ERRORLOG_LEVEL	表示Hue的错误日志级别。	ERROR	• ERROR • WARN • INFO • DEBUG
HANDLER_LOGFILE_LEVEL	表示Hue的运行日志级别。	INFO	• ERROR • WARN • INFO • DEBUG
HANDLER_LOGFILE_MAXBACKUPINDEX	表示Hue日志文件最大个数。	20	1～999
HANDLER_LOGFILE_SIZE	表示Hue日志文件最大大小。	5MB	-

12.11.4 在 Hue WebUI 使用 HiveQL 编辑器

操作场景

用户需要使用图形化界面在集群中执行HiveQL语句时，可以通过Hue完成任务。

访问“Query Editors”

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 选择“Query Editors > Hive”，进入“Hive”。

“Hive”支持以下功能：

- 执行和管理HiveQL语句。
- 在“Saved Queries”中查看当前访问用户已保存的HiveQL语句。
- 在“Query History”中查看当前访问用户执行过的HiveQL语句。
- 单击，在“Databases”下可以显示Hive中所有的数据库。

----结束

执行 HiveQL 语句

步骤1 选择“Query Editors > Hive”，进入“Hive”。

步骤2 单击，在“Databases”下选择一个数据库，默认数据库为“default”。

系统将自动显示数据库中的所有表。可以输入表名关键字，系统会自动搜索包含此关键字的全部表。

步骤3 单击指定的表名，可以显示表中所有的列。

光标移动到表所在的行，单击可以查看列的详细信息。

步骤4 在HiveQL语句编辑区输入查询语句。

单击并选择“Explain”，编辑器将分析输入的查询语句是否有语法错误以及执行计划，如果存在语法错误则显示“Error while compiling statement”。

步骤5 单击，选择HiveQL语句执行的引擎。

- “mr”表示语句使用MapReduce计算框架执行语句。
- “spark”表示语句使用Spark计算框架执行语句。
- “tez”表示语句使用Tez计算框架执行语句。

说明

tez适用于MRS 1.9.x及以后版本。

步骤6 单击开始执行HiveQL语句。

说明

- 如果希望下次继续使用已输入的HiveQL语句，请单击 保存。
- 格式化HiveQL语句，请单击 选择“Format”。
- 删除已输入的HiveQL语句，请单击 选择“Clear”。
- 清空已输入的语句并执行一个新的语句，请单击 选择“New query”。
- 查看历史：
  单击“Query History”，可查看HiveQL运行情况，支持显示所有语句或只显示保存的语句的运行情况。历史记录存在多个结果时，可以在输入框使用关键字进行搜索。
- 高级查询配置：
  单击右上角的，对文件、函数、设置等信息进行配置。
- 查看快捷键：
  单击右上角的，可查看所有快捷键信息。

----结束

查看执行结果

步骤1 在“Hive”的执行区，默认显示“Query History”。

步骤2 单击“Results”查看已执行语句的执行结果。

----结束

管理查询语句

步骤1 选择“Query Editors > Hive”，进入“Hive”。

步骤2 单击“Saved Queries”。

单击一条已保存的语句，系统会自动将其填充至编辑区中。

----结束

修改在 Hue 使用“Query Editors”的会话配置

步骤1 在“Hive”页签，单击 。

步骤2 在“Files”的右侧单击 ，然后单击 指定该文件的存储目录。

可以单击 新增加一个文件资源。

步骤3 在“Functions”的右侧单击 ，输入用户自定义的名称和函数的类名称。

可以单击 新增加一个自定义函数。

步骤4 在“Settings”的右侧单击 ，在“Key”输入Hive的参数名，在“Value”输入对应的参数值，则当前Hive会话会以用户定义的配置连接Hive。

可以单击 新增加一个参数。

----结束

12.11.5 在 Hue WebUI 使用元数据浏览器

操作场景

用户需要使用图形化界面在集群中管理Hive的元数据，可以通过Hue完成任务。

Metastore 管理器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

选择“Data Browsers > Metastore Tables”，进入“Metastore Manager”。

- 查看Hive表的元数据

  在左侧导航栏中，将鼠标放在某一表上，单击显示在其右侧的图标 ，界面将显示Hive表的元数据信息。
- 管理Hive表的元数据

  在Hive表的元数据信息界面，单击右上角的 可导入数据，单击 可浏览数据，单击 可查看表文件的位置信息。

  注意

  Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。
- 管理Hive元数据表

  选择右上角的 可在数据库中根据上传的文件创建一个新表，选择右上角的 可手动创建一个新表。

访问“Metastore Manager”

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 选择“Data Browsers > Metastore Tables”，进入“Metastore Manager”。

“Metastore Manager”支持以下功能：

- 使用文件创建一个Hive表
- 手动创建一个Hive表
- 查看Hive表元数据

----结束

使用文件创建一个 Hive 表

步骤1 访问“Metastore Manager”，在“Databases”选择一个数据库。

默认数据库为“default”。

步骤2 单击  ，进入“Create a new table from a file”页面。

步骤3 选择文件。

1. 在“Table Name”填写Hive表的名称。

   支持字母、数字、下划线，首位必须为字母或数字，且长度不能超过128位。
2. 根据需要，在“Description”填写Hive表的描述信息。
3. 在“Input File or Location”单击，在HDFS中选择一个用于创建Hive表文件。此文件将存储Hive表的新数据。

   如果文件未在HDFS中保存，可以单击“Upload a file”从本地选择文件并上传。支持同时上传多个文件，文件不可为空。
4. 如果需要将文件中的数据导入Hive表，选择“Import data”作为“Load method”。默认选择“Import data”。

   选择“Create External Table”时，创建的是Hive外部表。

   说明

   当选择“Create External Table”时，参数“Input File or Location”需要选择为路径。

   选择“Leave Empty”则创建空的Hive表。
5. 单击“Next”。

步骤4 设置分隔符。

1. 在“Delimiter”选择一个分隔符。

   如果分隔符不在列表中，选择“Other..”，然后输入新定义的分隔符。
2. 单击“Preview”查看数据处理预览。
3. 单击“Next”。

步骤5 定义字段列。

1. 单击“Use first row as column names”右侧的  ，则使用文件中第一行数据作为列名称。取消则不使用数据作为列名称。
2. 在“Column name”编辑每个列的名称。

   支持字母、数字、下划线，首位必须为字母或数字，且长度不能超过128位。

   说明

   单击“Bulk edit column names”右侧的  ，可批量对列重新命名。输入所有列的名称并使用逗号分隔。
3. 在“Column Type”选择每个列的类型。

步骤6 单击“Create Table”创建表，等待Hue显示Hive表的信息。

----结束

手工创建一个 Hive 表

步骤1 访问“Metastore Manager”，在“Databases”选择一个数据库。

默认数据库为“default”。

步骤2 单击，进入“Create a new table manually”页面。

步骤3 设置表名称。

1. 在“Table Name”填写Hive表的名称。
   支持字母、数字、下划线，首位必须为字母或数字，且长度不能超过128位。
2. 根据需要，在“Description”填写Hive表的描述信息。
3. 单击“Next”。

步骤4 选择一个存储数据的格式。

- 需要使用分隔符分隔数据时，选择“Delimited”，然后执行**步骤5**。
- 需要使用序列化格式保存数据时，选择“SerDe”，执行**步骤6**。

步骤5 配置分隔符。

1. 在“Field terminator”设置一个列分隔符。
   如果分隔符不在列表中，选择“Other..”，然后输入新定义的分隔符。
2. 在“Collection terminator”设置一个分隔符，用于分隔Hive中类型为“array”的列的数据集合。例如一个列为array类型，其中一个值需要保存“employee”和“manager”，用户指定分隔符为“:”，则最终的值为“employee:manager”。
3. 在“Map key terminator”设置一个分隔符，用于分隔Hive中类型为“map”的列的数据。例如某个列为map类型，其中一个值需要保存描述为“aaa”的“home”，和描述为“bbb”的“company”，用户指定分隔符为“|”，则最终的值为“home|aaa:company|bbb”。
4. 单击“Next”，执行**步骤7**。

步骤6 设置序列化属性。

1. 在“SerDe Name”输入序列化格式的类名称“org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe”。
   用户可扩展Hive支持更多自定义的序列化类。
2. 在“Serde properties”输入序列化的样式的值：“"field.delim"="," "colelction.delim"=":" "mapkey.delim"="|"”。
3. 单击“Next”，执行执行**步骤7**。

步骤7 选择一个数据表的格式，并单击“Next”。

- “TextFile”表示使用文本类型文件存储数据。
- “SequenceFile”表示使用二进制类型文件存储数据。
- “InputFormat”表示使用自定义的输入输出格式来使用文件中的数据。
  用户可扩展Hive支持更多的自定义格式化类。
  a. 在“InputFormat Class”填写输入数据使用的类“org.apache.hadoop.hive.ql.io.RCFileInputFormat”。
  b. 在“OutputFormat Class”填写输出数据使用的类“org.apache.hadoop.hive.ql.io.RCFileOutputFormat”。

步骤8 选择一个文件保存位置，并单击“Next”。

默认勾选“Use default location”。如果需要自定义存储位置，请取消选中状态并在“External location”单击 指定一个文件存储位置。

步骤9 设置Hive表的字段。

1. 在“Column name”设置列的名称。
   支持字母、数字、下划线，首位必须为字母或数字，且长度不能超过128位。
2. 在“Column type”选择一个数据类型。
   单击“Add a column”可增加新的列。
3. 单击“Add a partition”为Hive表增加分区，可提高查询效率。

步骤10 单击“Create Table”创建表，等待Hue显示Hive表的信息。

----结束

管理 Hive 表

步骤1 访问“Metastore Manager”，在“Databases”选择一个数据库，页面显示数据库中所有的表。

默认数据库为“default”。

步骤2 单击数据库中的表名称，打开表的详细信息。

支持导入数据、浏览数据或查看文件存储位置。查看数据库所有的表时，可以直接勾选表然后执行查看、浏览数据操作。

注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

----结束

12.11.6 在 Hue WebUI 使用文件浏览器

操作场景

用户需要使用图形化界面管理HDFS中文件时，可以通过Hue完成任务。

注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

访问文件浏览器（File Browser）

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击  ，进入“File Browser”。

默认进入当前登录用户的主目录。

文件浏览器将显示目录中的子目录或文件以下信息：

表 12-242 HDFS 文件属性介绍

属性名	描述
“Name”	表示目录或文件的名称。
“Size”	表示文件的大小。
“User”	表示目录或文件的属主。
“Group”	表示目录或文件的属组。
“Permissions”	表示目录或文件的权限设置。
“Date”	表示目录或文件创建时间。

步骤3 在搜索框输入关键字，系统会在当前目录自动搜索目录或文件。

步骤4 清空搜索框的内容，系统会重新显示所有目录和文件。

----结束

执行动作

步骤1 单击  ，选择一个或多个目录或文件。

步骤2 单击“Actions”，在弹出菜单选择一个操作。

- “Rename”：表示重新命名一个目录或文件。
- “Move”：表示移动文件，在“移至”选择新的目录并单击“移动”完成移动。
- “Copy”：表示复制选中的文件或目录。
- “Change permissions”：表示修改选中目录或文件的访问权限。
  - 可以为属主、属组和其他用户设置“Read”、“Write”和“Excute”权限。
  - “Sticky”表示禁止HDFS的管理员、目录属主或文件属主以外的用户在目录中移动文件。
  - “Recursive”表示递归设置权限到子目录。
- “Storage policies”：表示设置目录或文件在HDFS中的存储策略。
- “Summary”：表示查看选中文件或目录的HDFS存储信息。

----结束

访问其他目录

步骤1 单击目录名并输入需要访问的目录完整路径，例如“/mr-history/tmp”并按回车键进入目录。

需要当前登录Hue WebUI的用户拥有其他目录的访问权限。

步骤2 单击“Home”可进入用户的主目录。

步骤3 单击“History”可以显示最近访问目录的历史记录，并重新访问。

步骤4 单击“Trash”可以访问当前目录的回收站空间。

单击“Empty Trash”可清空回收站。

----结束

上传用户文件

步骤1 单击，单击Upload。

步骤2 选择一个操作。

- “Files”：表示上传用户文件到当前用户。
- “Zip/Tgz/Bz2 file”：表示上传了一个压缩文件，在弹出框单击“Select ZIP, TGZ or BZ2 files”选择需要上传的压缩文件。系统会自动在HDFS中对文件解压。支持“ZIP”、“TGZ”和“BZ2”格式的压缩文件。

----结束

创建新文件或者目录

步骤1 单击，单击“New”。

步骤2 选择一个操作。

- “File”：表示创建一个文件，输入文件名后单击“Create”完成。
- “Directory”：表示创建一个目录，输入目录名后单击“Create”完成

----结束

存储策略定义使用介绍

说明

若Hue的服务配置参数“fs_defaultFS”配置为“viewfs://ClusterX”时，不能启用存储策略定义功能。

步骤1 登录MRS Manager。

步骤2 在MRS Manager界面，选择“系统设置 > 权限配置 > 角色管理 > 添加角色”：

1. 设置“角色名称”。
2. 选择“权限 > Hue”，勾选"Storage Policy Admin"，单击“确定”，为该角色赋予存储策略管理员的权限。

步骤3 选择“系统设置 > 权限配置 > 用户组管理 > 添加用户组”，设置“组名”，单击“角色”后的“选择添加角色”，在弹出的界面选择刚创建的角色，单击“确定”将该角色添加到组中。

步骤4 选择“系统设置 > 权限配置 > 用户管理 > 添加用户”：

1. 设置可以登录Hue的WebUI界面且有存储策略管理员权限的用户的“用户名”。
2. “用户类型”选择“人机”。
3. 设置登录Hue的WebUI界面的“密码”、“确认密码”。
4. 单击“用户组”后的“选择添加的用户组”，在弹出的界面选择创建的用户组、supergroup、hadoop和hive用户组，单击“确定”。
5. “主组”选择“hive”。
6. 单击“分配角色权限”右侧的“选择并绑定角色”，在弹出的界面选择刚刚创建的角色和System_administrator角色，单击“确定”。
7. 再单击“确定”成功添加该用户。

步骤5 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤6 单击右上角的  。

步骤7 勾选目录的复选框，单击页面上方的“Action”，选择“Storage policies”。

步骤8 在弹出的对话框中设置新的存储策略，单击“OK”。

----结束

12.11.7 在 Hue WebUI 使用作业浏览器

操作场景

用户需要使用图形化界面查看集群中所有作业时，可以通过Hue完成任务。

访问“Job Browser”

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击“Job Browser”。

默认显示当前集群的所有作业。

说明

“Job Browser”显示的数字表示集群中所有作业的总数。

“Job Browser”将显示作业以下信息：

表 12-243 MRS 作业属性介绍

属性名	描述
“Logs”	表示作业的日志信息。如果作业有输出日志，则显示  。
“ID”	表示作业的编号，由系统自动生成。
“Name”	表示作业的名称。
“Application Type”	表示作业的类型。

属性名	描述
“Status”	表示作业的状态，包含“RUNNING”、“SUCCEEDED”、“FAILED”和“KILLED”。
“User”	表示启动该作业的用户。
“Maps”	表示作业执行Map过程的进度。
“Reduces”	表示作业执行Reduce过程的进度。
“Queue”	表示作业运行时使用的YARN队列。
“Priority”	表示作业运行时的优先级。
“Duration”	表示作业运行使用的时间。
“Submitted”	表示作业提交到MRS集群的时间。

说明

如果MRS集群安装了Spark组件，则默认会启动一个作业“Spark-JDBCServer”，用于执行任务。

----结束

搜索作业

步骤1 在“Job Browser”的“Username”或“Text”，输入指定的字符，系统会自动搜索包含此关键字的全部作业。

步骤2 清空搜索框的内容，系统会重新显示所有作业。

----结束

查看作业详细信息

步骤1 在“Job Browser”的作业列表，单击作业所在的行，可以打开作业详情。

步骤2 在“Metadata”页签，可查看作业的元数据。

说明

单击可打开作业运行时的日志。

----结束

12.12 使用 Hue（MRS 3.x 及之后版本）

12.12.1 从零开始使用 Hue

Hue汇聚了与大多数Apache Hadoop组件交互的接口，致力让用户通过界面图形化的方式轻松使用Hadoop组件。目前Hue支持HDFS、Hive、HBase、Yarn、MapReduce、Oozie和SparkSQL等组件的可视化操作。

前提条件

已安装Hue组件，且状态为运行中的Kerberos认证的集群。

操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在左侧导航栏单击编辑器  ，然后选择“Hive”。

步骤3 在“Database”右侧下拉列表选择一个Hive中的数据库，默认数据库为“default”。

系统将自动显示数据库中的所有表。可以输入表名关键字，系统会自动搜索包含此关键字的全部表。

步骤4 单击指定的表名，可以显示表中所有的列。

步骤5 在HiveQL语句编辑区输入HiveQL语句。

create table hue_table(id int,name string,company string) row format delimited fields terminated by ',' stored as textfile;

步骤6 单击 ▶ 开始执行HiveQL语句。

步骤7 在命令输入框内输入**show tables;**，单击 ▶ 按钮，查看“结果”中有**步骤5**创建的表hue_table。

----结束

12.12.2 访问 Hue 的 WebUI

操作场景

MRS集群安装Hue组件后，用户可以通过Hue的WebUI，在图形化界面使用Hadoop生态相关组件。

该任务指导用户在MRS集群中打开Hue的WebUI。

说明

Internet Explorer浏览器可能存在兼容性问题，建议更换兼容的浏览器访问Hue WebUI，例如Google Chrome浏览器50版本。

对系统的影响

第一次访问Manager和Hue WebUI，需要在浏览器中添加站点信任以继续打开Hue WebUI。

前提条件

启用Kerberos认证时，MRS集群管理员已分配用户使用Hive的权限。具体操作请参见“用户指南 > MRS操作指导 > 权限管理 > 创建用户”章节。例如创建一个“人机”用户“hueuser”，并加入“hive”、“hadoop”、“supergroup”组和“System_administrator”角色，主组为“hive”。

该用户用于登录Manager。

操作步骤

步骤1 登录服务页面：

MRS 3.x之前版本，在MRS控制台单击集群名称，选择“组件管理 > Hue”。

MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Hue”。

步骤2 在“Hue WebUI”右侧，单击链接，打开Hue的WebUI。

Hue的WebUI支持以下功能：

- 使用编辑器  执行Hive、SparkSql的查询语句以及Notebook代码段。需要MRS集群已安装Hive、Spark2x。
- 使用计划程序  提交Workflow任务、计划任务、Bundle任务。
- 使用文档  查看、导入、导出在Hue页面上操作的任务，例如保存的Workflow任务、定时任务、Bundle任务等。
- 使用表  管理Hive、SparkSql中的元数据。需要MRS集群已安装Hive、Spark2x。
- 使用文件  查看HDFS中的目录和文件。需要MRS集群已安装HDFS。
- 使用作业  查看MRS集群中所有作业。需要MRS集群已安装Yarn。
- 使用HBase  创建/查询HBase表。需要MRS集群已安装HBase组件并添加Thrift1Server实例。
- 使用导入器  通过“.csv”，“.txt”等格式的文件导入数据。

说明

- 使用创建的用户第一次登录Hue WebUI，需修改密码。
- 用户获取Hue WebUI的访问地址后，可以给其他无法访问Manager的用户用于访问Hue WebUI。
- 在Hue的WebUI操作但不操作Manager页面，重新访问Manager时需要输入已登录的帐号密码。

----结束

12.12.3 Hue 常用参数

参数入口

参数入口，请参考**修改集群服务配置参数**进入Hue服务“全部配置”页面。

参数说明

Hue常用参数请参见**表12-244**。

表 12-244 Hue 常用参数

配置参数	说明	缺省值	范围
HANDLER_ACCESSLOG_LEVEL	Hue的访问日志级别。	DEBUG	• ERROR • WARN • INFO • DEBUG
HANDLER_AUDITSLOG_LEVEL	Hue的审计日志级别。	DEBUG	• ERROR • WARN • INFO • DEBUG
HANDLER_ERRORLOG_LEVEL	Hue的错误日志级别。	ERROR	• ERROR • WARN • INFO • DEBUG
HANDLER_LOGFILE_LEVEL	Hue的运行日志级别。	INFO	• ERROR • WARN • INFO • DEBUG
HANDLER_LOGFILE_MAXBACKUPINDEX	Hue日志文件最大个数。	20	1～999
HANDLER_LOGFILE_SIZE	Hue日志文件最大大小。	5MB	-

12.12.4 在 Hue WebUI 使用 HiveQL 编辑器

操作场景

用户需要使用图形化界面在集群中执行HiveQL语句时，可以通过Hue完成任务。

访问编辑器

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在左侧导航栏单击  ，然后选择“Hive”，进入“Hive”。

“Hive”支持以下功能：

- 执行和管理HiveQL语句。
- 在“保存的查询”中查看当前访问用户已保存的HiveQL语句。

- 在“查询历史记录”中查看当前访问用户执行过的HiveQL语句。

----结束

执行 HiveQL 语句

步骤1 在“Database”右侧下拉列表选择一个Hive中的数据库，默认数据库为“default”。

系统将自动显示数据库中的所有表。可以输入表名关键字，系统会自动搜索包含此关键字的全部表。

步骤2 单击指定的表名，可以显示表中所有的列。

光标移动到表或列所在的行，单击 可以查看详细信息。

步骤3 在HiveQL语句编辑区输入查询语句。

步骤4 单击 开始执行HiveQL语句。

说明

- 如果希望下次继续使用已输入的HiveQL语句，请单击保存。
- 高级查询配置：

  单击右上角的，对文件、功能、设置等信息进行配置。
- 查看快捷键：

  单击右上角的，可查看语法和键盘快捷方式信息。
- 删除已输入的HiveQL语句，请单击 后的三角选择“清除”。
- 查看历史：

  单击“查询历史记录”，可查看HiveQL运行情况，支持显示所有语句或只显示保存的语句的运行情况。历史记录存在多个结果时，可以在输入框使用关键字进行搜索。

----结束

查看执行结果

步骤1 在“Hive”的执行区，默认显示“查询历史记录”。

步骤2 单击结果查看已执行语句的执行结果。

----结束

管理查询语句

步骤1 单击“保存的查询”。

步骤2 单击一条已保存的语句，系统会自动将其填充至编辑区中。

----结束

修改在 Hue 使用编辑器的会话配置

步骤1 在编辑器页面，单击 。

步骤2 在“文件”的右侧单击 ，然后单击 选择文件。

可以单击“文件”后的 新增加一个文件资源。

步骤3 在“功能” ，输入用户自定义的名称和函数的类名称。

可以单击“功能”后的 新增加一个自定义函数。

步骤4 在“设置” ，在“设置”的“键”输入Hive的参数名，在“值”输入对应的参数值，则当前Hive会话会以用户定义的配置连接Hive。

可以单击 新增加一个参数。

----结束

12.12.5 在 Hue WebUI 使用 SparkSql 编辑器

操作场景

用户需要使用图形化界面在集群中执行SparkSql语句时，可以通过Hue完成任务。

配置 Spark2x

使用SparkSql编辑器之前需要先修改Spark2x配置。

步骤1 进入Spark2x的全部配置页面，具体操作请参考**修改集群服务配置参数**。

步骤2 设置Spark2x多实例模式，搜索并修改Spark2x服务的以下参数：

参数名称	值
spark.thriftserver.proxy.enabled	false
spark.scheduler.allocation.file	#{conf_dir}/fairscheduler.xml

步骤3 进入JDBCServer2x自定义界面，在spark.core-site.customized.configs参数内，添加两个自定义项：

名称为：hadoop.proxyuser.hue.groups，值为：*

名称为：hadoop.proxyuser.hue.hosts，值为：*

步骤4 保存配置，重启Spark2x服务。

----结束

访问编辑器

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在左侧导航栏单击 ，然后选择“SparkSql”，进入“SparkSql”。

“SparkSql”支持以下功能：

- 执行和管理SparkSql语句。
- 在“保存的查询”中查看当前访问用户已保存的SparkSql语句。
- 在“查询历史记录”中查看当前访问用户执行过的SparkSql语句。

----结束

执行 SparkSql 语句

步骤1 在“Database”右侧下拉列表选择一个SparkSql中的数据库，默认数据库为“default”。

系统将自动显示数据库中的所有表。可以输入表名关键字，系统会自动搜索包含此关键字的全部表。

步骤2 单击指定的表名，可以显示表中所有的列。

光标移动到表所在的行，单击 可以查看列的详细信息。

步骤3 在SparkSql语句编辑区输入查询语句。

单击 后的三角并选择“解释”，编辑器将分析输入的查询语句是否有语法错误以及执行计划，如果存在语法错误则显示“Error while compiling statement”。

步骤4 单击 开始执行SparkSql语句。

说明

- 如果希望下次继续使用已输入的SparkSql语句，请单击 保存。
- 高级查询配置：

  单击右上角的 ，对文件、功能、设置等信息进行配置。
- 查看快捷键：

  单击右上角的 ，可查看语法和键盘快捷方式信息。
- 格式化SparkSql语句，请单击 后的三角选择“格式”
- 删除已输入的SparkSql语句，请单击 后的三角选择“清除”
- 查看历史：

  单击“查询历史记录”，可查看SparkSql运行情况，支持显示所有语句或只显示保存的语句的运行情况。历史记录存在多个结果时，可以在输入框使用关键字进行搜索。

----结束

查看执行结果

步骤1 在“SparkSql”的执行区，默认显示“查询历史记录”。

步骤2 单击结果查看已执行语句的执行结果。

----结束

管理查询语句

步骤1 单击“保存的查询”。

步骤2 单击一条已保存的语句，系统会自动将其填充至编辑区中。

----结束

12.12.6 在 Hue WebUI 使用元数据浏览器

操作场景

用户需要使用图形化界面在集群中管理Hive的元数据，可以通过Hue完成任务。

元数据管理器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

- 查看Hive表的元数据

  在左侧导航栏单击表  ，单击某一表名称，界面将显示Hive表的元数据信息。
- 管理Hive表的元数据

  在Hive表的元数据信息界面：
  - 单击右上角的“导入”可导入数据。
  - 单击“概述”，在“属性”域可查看表文件的位置信息。

    可查看Hive表各列字段的信息，并手动添加描述信息，注意此处添加的描述信息并不是Hive表中的字段注释信息（comment）。
  - 单击“样本”可浏览数据。
- 管理Hive元数据表

  单击左侧列表中的 ➕ 可在数据库中根据上传的文件创建一个新表，也可手动创建一个新表。

⚠ 注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

12.12.7 在 Hue WebUI 使用文件浏览器

操作场景

用户需要使用图形化界面管理HDFS中文件时，可以通过Hue完成任务。

⚠ 注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

访问文件浏览器

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在左侧导航栏单击文件  。进入“文件浏览器”页面。

“文件浏览器”的“主页”默认进入当前登录用户的主目录。界面将显示目录中的子目录或文件的以下信息：

表 12-245 HDFS 文件属性介绍

属性名	描述
名称	表示目录或文件的名称。
大小	表示文件的大小。
用户	表示目录或文件的属主。
组	表示目录或文件的属组。
权限	表示目录或文件的权限设置。
日期	表示目录或文件创建时间。

步骤3 在搜索框输入关键字，系统会在当前目录自动搜索目录或文件。

步骤4 清空搜索框的内容，系统会重新显示所有目录和文件。

----结束

执行动作

步骤1 在“文件浏览器”界面，勾选一个或多个目录或文件。

步骤2 单击“操作”，在弹出菜单选择一个操作。

- 重命名：表示重新命名一个目录或文件。
- 移动：表示移动文件，在“移至”页面选择新的目录并单击“移动”完成移动。
- 复制：表示复制选中的文件或目录。
- 更改权限：表示修改选中目录或文件的访问权限。
  - 可以为属主、属组和其他用户设置“读取”、“写”和“执行”权限。
  - “易贴”表示禁止HDFS的管理员、目录属主或文件属主以外的用户在目录中移动文件。

  - “递归”表示递归设置权限到子目录。
- 存储策略：表示设置目录或文件在HDFS中的存储策略。
- 摘要：表示查看选中文件或目录的HDFS存储信息。

----结束

上传用户文件

步骤1 在“文件浏览器”界面，单击“上传”。

步骤2 在弹出的上传文件窗口中单击“选择文件”或将文件拖至窗口中，完成文件上传。

----结束

创建新文件或者目录

步骤1 在“文件浏览器”界面，单击“新建”。

步骤2 选择一个操作。

- 文件：表示创建一个文件，输入文件名后单击“创建”完成。
- 目录：表示创建一个目录，输入目录名后单击“创建”完成。

----结束

存储策略定义使用介绍

说明

若Hue的服务配置参数“fs_defaultFS”配置为“viewfs://ClusterX”时，不能启用存储策略定义功能。

步骤1 登录FusionInsight Manager。

步骤2 在FusionInsight Manager界面，选择“系统 > 权限 > 角色 > 添加角色”：

1. 设置“角色名称”。
2. 在“配置资源权限”下选择“*待操作集群名称*>Hue”，勾选“存储策略管理员”，单击“确定”，为该角色赋予存储策略管理员的权限。

步骤3 选择“系统 > 权限 > 用户组 > 添加用户组”，设置“组名”，单击“角色”后的“添加”，在弹出的界面选择**步骤2**创建的角色，单击“确定”将该角色添加到组中，单击“确定”完成用户组的创建。

步骤4 选择“系统 > 权限 > 用户 > 添加用户”：

1. “用户名”填写待添加的用户名。
2. “用户类型”设置为“人机”。
3. 设置登录Hue的WebUI界面的“密码”、“确认密码”。
4. 单击“用户组”后的“添加”，在弹出的界面选择**步骤3**创建的用户组、supergroup、hadoop和hive用户组，单击“确定”。
5. “主组”选择“hive”。
6. 单击“角色”后的“添加”，在弹出的界面选择**步骤2**创建的角色和System_administrator角色，单击“确定”。

7. 再单击“确定”，成功添加该用户。

步骤5 使用创建的用户访问Hue WebUI，具体操作请参考**访问Hue的WebUI**。

步骤6 左侧导航栏单击文件  。进入“文件浏览器”页面。

步骤7 勾选目录的复选框，单击页面上方的“操作”，单击“存储策略”。

步骤8 在弹出的对话框中设置新的存储策略，单击“保存”。

----结束

12.12.8 在 Hue WebUI 使用作业浏览器

操作场景

用户需要使用图形化界面查看集群中所有作业时，可以通过Hue完成任务。

访问作业浏览器

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击作业  。

默认显示当前集群的所有作业。

说明

作业浏览器显示的数字表示集群中所有作业的总数。

“作业浏览器”将显示作业以下信息：

表 12-246 MRS 作业属性介绍

属性名	描述
名称	表示作业的名称。
用户	表示启动该作业的用户。
类型	表示作业的类型。
状态	表示作业的状态，包含“成功”、“正在运行”、“失败”。
进度	表示作业运行进度。
组	表示作业所属组。
开始	表示作业开始时间。
持续时间	表示作业运行使用的时间。
Id	表示作业的编号，由系统自动生成。

说明

如果MRS集群安装了Spark组件，则默认会启动一个作业“Spark-JDBCServer”，用于执行任务。

----结束

搜索作业

步骤1 在“作业浏览器”的搜索栏，输入指定的字符，系统会按照ID、名称、用户自动搜索包含此关键字的全部作业。

步骤2 清空搜索框的内容，系统会重新显示所有作业。

----结束

查看作业详细信息

步骤1 在“作业浏览器”的作业列表，单击作业所在的行，可以打开作业详情。

步骤2 在“元数据”页签，可查看作业的元数据。

说明

单击“日志”可打开作业运行时的日志。

----结束

12.12.9 在 Hue WebUI 使用 HBase

操作场景

用户需要使用图形化界面在集群中创建或查询HBase表时，可以通过Hue完成任务。

需要MRS集群已安装HBase组件并添加Thrift1Server实例。

访问作业浏览器

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击HBase  ，进入“HBase Browser”页面。

----结束

新建 HBase 表

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击HBase  ，进入“HBase Browser”页面。

步骤3 单击右侧“新建表”按钮，输入表名和列族参数，单击“提交”，完成HBase表创建。

----结束

查询 HBase 表数据

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击HBase  ，进入“HBase Browser”页面。

步骤3 单击需要查询的HBase表。可在上方的搜索栏后单击健值，对HBase表进行查询。

----结束

12.12.10 典型场景

12.12.10.1 HDFS on Hue

Hue提供了文件浏览器功能，使用户可以通过界面图形化的方式使用HDFS。

⚠ 注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

文件浏览器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

然后单击  ，进入“文件浏览器”页面。您可以进行以下操作。

- 查看文件和目录

  默认显示登录用户的目录及目录中的文件，可查看目录或文件的“名称”、“大小”、“用户”、“组”、“权限”和“日期”信息。

  单击文件名，可查看文本文件的文本信息或二进制数据。支持编辑文件内容。

  如果文件和目录数量比较多，可以在搜索框输入关键字，搜索特定的文件或目录。

- 创建文件或目录

  单击右上角的“新建”，选择“文件”创建文件，选择“目录”创建目录。

- 管理文件或目录

  勾选文件或目录的复选框，单击“操作”，选择“重命名”、“移动”、“复制”和“更改权限”等，实现文件或目录的重命名、移动、复制、更改权限等功能。

- 上传文件

  单击右上角的“上传”，单击“选择文件”或将文件拖至窗口中可进行文件上传。

存储策略定义使用介绍

说明

若Hue的服务配置参数“fs_defaultFS”配置为“viewfs://ClusterX”时，不能启用存储策略定义功能。

存储策略定义在Hue的WebUI界面上分为两大类：

- 静态存储策略

  当前存储策略

  根据HDFS的文档访问频率、重要性，为HDFS目录指定存储策略，例如ONE_SSD、ALL_SSD等，此目录下的文件可被迁移到相应存储介质上保存。

- 动态存储策略

  为HDFS目录设置规则，系统可以根据文件的最近访问时间、最近修改时间自动修改存储策略、修改文件副本数、移动文件目录。

  在Hue的WebUI界面设置动态存储策略之前，需先在Manager界面设置冷热数据迁移的CRON表达式，并启动自动冷热数据迁移特性。

  操作方法为：

  修改HDFS服务的NameNode的“dfs.auto.data.mover.cron.expression”的参数值。参数修改方法请参考**修改集群服务配置参数**。

  说明

  - “dfs.auto.data.mover.cron.expression”表示触发检测HDFS数据是否满足动态存储策略规则的CRON表达式，用于控制数据迁移操作的开始时间。其默认值是“0 * * * *”，表示在整点检测。当满足动态存储策略规则时，在该整点执行冷热数据迁移任务。
  - “dfs.auto.data.mover.enable”的默认值是“false”。仅当“dfs.auto.data.mover.enable”设置为“true”时该值才有效。

  修改此参数时，表达式介绍如**表12-247**所示。支持“*”表示连续的时间段。

  表 12-247 执行表达式参数解释

列	说明
第1列	分钟，参数值为0～59。
第2列	小时，参数值为0～23。
第3列	日期，参数值为1～31。
第4列	月份，参数值为1～12。
第5列	星期，参数值为0～6，0表示星期日。

存储策略定义在WebUI界面上的操作如下：

步骤1 登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 在FusionInsight Manager界面，选择“系统 > 权限 > 角色 > 添加角色”：

1. 设置“角色名称”。

2. 在“配置资源权限”下选择“*待操作集群名称*>Hue”，勾选“存储策略管理员”，单击“确定”，为该角色赋予存储策略管理员的权限。

步骤3 选择“系统 > 权限 > 用户组 > 添加用户组”，设置“组名”，单击“角色”后的“添加”，在弹出的界面选择**步骤2**创建的角色，单击“确定”将该角色添加到组中，单击“确定”完成用户组的创建。

步骤4 选择“系统 > 权限 > 用户 > 添加用户”：

1. “用户名”填写待添加的用户名。
2. “用户类型”设置为“人机”。
3. 设置登录Hue的WebUI界面的“密码”、“确认密码”。
4. 单击“用户组”后的“添加”，在弹出的界面选择**步骤3**创建的用户组、supergroup、hadoop和hive用户组，单击“确定”。
5. “主组”选择“hive”。
6. 单击“角色”后的“添加”，在弹出的界面选择**步骤2**创建的角色和System_administrator角色，单击“确定”。
7. 再单击“确定”，成功添加该用户。

步骤5 使用创建的用户访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤6 左侧导航栏单击文件  。进入“文件浏览器”页面。

步骤7 勾选目录的复选框，单击页面上方的“操作”，单击“存储策略”。

步骤8 在弹出的对话框中设置新的存储策略，单击“确定”。

- 在“静态存储策略”页签设置静态存储策略，单击“保存”。
- 在“动态存储策略”页签可创建、删除、修改动态存储策略，详细的参数介绍如**表12-248**所示。

表 12-248 动态存储策略参数介绍

分类	参数	说明
规则	文件最近访问时间	按照该文件最近一次访问时间。
	文件最近修改时间	按照该文件最近一次修改时间。
操作	修改副本数	设置文件副本数。
	修改存储策略	修改存储策略，包括HOT、WARM、COLD、ONE_SSD、ALL_SSD。
	移动到目录	移动该文件到其他目录。

说明

- 设置规则需要用户充分考虑合理性，例如多条规则之间是否有冲突，是否会对系统造成破坏等。
- 一个目录设置多个规则和动作时，规则被先触发的放在规则/动作列表的下面，规则被后触发的放在规则/动作列表的上面，避免动作反复执行。
- 系统每个小时整点扫描动态存储策略指定的目录下的文件是否符合规则，如果满足，则触发执行动作。执行日志记录在主NameNode的“/var/log/Bigdata/hdfs/nn/hadoop.log”目录下。

----结束

典型场景

通过Hue界面对HDFS以文本或二进制查看和编辑文件的操作如下：

查看文件

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 左侧导航栏单击文件  。进入“文件浏览器”页面。

步骤3 单击需要查看的文件名。

步骤4 单击“以二进制格式查看”，可以切换视图从文本到二进制；单击“以文本格式查看”，可以切换视图从二进制到文本。

编辑文件

步骤5 单击“编辑文件”，显示文件内容可编辑。

步骤6 单击“保存”或“另存为”保存文件。

----结束

12.12.10.2 配置 HDFS 冷热数据迁移

配置场景

冷热数据迁移工具根据配置的策略移动HDFS文件。配置策略是条件或非条件规则的集合。如果规则匹配文件集，则该工具将对该文件执行一组行为操作。

冷热数据迁移工具支持以下规则和行为。

- 迁移规则：
  - 根据文件的最后访问时间迁移数据
  - 根据年龄时间迁移数据（修改时间）
  - 无条件迁移数据

表 12-249 规则条件标签

条件标签	描述
<age operator="lt">	定义年龄/修改时间的条件。

条件标签	描述
<atime operator="gt">	定义访问时间的条件。

说明

对于手动迁移规则，不需要条件。

- 行为列表：
  - 将存储策略设置为给定的数据层名称
  - 迁移到其他文件夹
  - 为文件设置新的副本数
  - 删除文件
  - 设置节点标签（NodeLabel）

表 12-250 行为类型

行为类型	描述	所需参数
MARK	为确定数据的冷热度并设置相应的数据存储策略。	<param> <name>targettier</name> <value>STORAGE_POLICY</value> <param>
MOVE	为设置数据存储策略或NodeLabel并调用HDFS Mover工具。	<param> <name>targettier</name> <value>STORAGE_POLICY</value> <param> <param> <name>targetnodelabels</name> <value>SOME_EXPRESSION</value> <param> **说明** 用户可以配置其中任一参数或两者都配置。
SET_REPL	为文件设置新的副本数。	<param> <name>replcount</name> <value>INTEGER</value> <param>

行为类型	描述	所需参数
MOVE_TO_FOLDER	将文件移动到目标文件夹。如果“overwrite”参数为“true”，则目标路径将被覆盖。	<param> <name>target</name> <value>PATH</value> <param> <param> <name>overwrite</name> <value>true/false</value> <param> **说明** “overwrite”是可选参数，如果未配置，则默认值为“false”。
DELETE	删除文件。	NA

配置描述

必须定期调用迁移工具，并需要在客户端的“hdfs-site.xml”文件中进行以下配置。

表 12-251 参数描述

参数	描述	默认值
dfs.auto-data-movement.policy.class	用于指定默认的数据迁移策略。 **说明** 当前只支持DefaultDataMovementPolicy。	com.xxx.hadoop.hdfs.datamovement.policy.DefaultDataMovementPolicy
dfs.auto.data.mover.id	冷热数据迁移输出（行为状态）文件的名称。	当前系统时间（毫秒）
dfs.auto.data.mover.output.dir	冷热数据迁移输出在HDFS中的目录名称。迁移工具将在此处写入行为状态文件。	/system/datamovement

DefaultDataMovementPolicy拥有配置文件“default-datamovement-policy.xml”。用户需要定义所有基于age/accessTime的规则和在此文件中采取的行为操作，此文件必须存储在客户端的classpath中。

如下为“default-datamovement-policy.xml”配置文件的示例：

```
<policies>
  <policy>
    <fileset>
      <file>
        <name>/opt/data/1.txt</name>
      </file>
      <file>
        <name>/opt/data/*/subpath/</name>
        <excludes>
          <name>/opt/data/some/subpath/sub1</name>
```

```
    </excludes>
   </file>
  </fileset>
  <rules>
   <rule>
    <age>2w</age>
    <action>
     <type>MOVE</type>
     <params>
      <param>
       <name>targettier</name>
       <value>HOT</value>
      </param>
     </params>
    </action>
   </rule>
  </rules>
 </policy>
</policies>
```

说明

在策略，规则和行为操作中使用的标签中，可以添加其他属性，例如“name”可用于管理用户界面（例如：Hue UI）和工具输入xml之间的映射。

示例： <policy name="Manage_File1">

标签（Tag）说明如下：

表 12-252 配置标签（Tag）描述

标签（Tag）名称	描述	是否可重复使用
<policy>	定义单一策略。 ● idempotent属性：指定当策略中有多个规则时，如果满足当前规则，是否检查下一个规则。 示例：<policy name ="policy2" idempotent ="true">。 其默认值为“true”，表示其中的规则和行为操作是幂等的，可以继续检查下一个规则。如果值为“false”，则将在当前规则处停止评估。 ● hours_allowed属性：配置是否根据系统时间执行策略评估。 hours_allowed的值是以逗号分隔的数字，范围从0到23，表示系统时间。 示例：<policy name ="policy1" hours_allowed ="2-6,13-14"> 如果当前系统时间在配置的范围内，则继续评估。否则，将跳过评估。 **说明** 在输入XML中，每个文件仅支持一个策略。因此，文件中的所有规则必须由一个策略标签覆盖。	Yes
<fileset>	为每个策略定义一组文件/文件夹。	No（在policy标签内）

标签（Tag）名称	描述	是否可重复使用
<file>	定义文件和/或文件夹在<file>标签内被配置一个或者多个<name>标签。文件/文件夹名支持POSIX globs配置。	Yes（在fileset标签内）
<excludes>	在<file>标签内定义该标签，该标签下可以包含多个<name>标签，在<file>标签中配置的文件或文件夹范围下，<name>标签所包含的文件或文件夹将会被排除。文件或文件夹名支持POSIX globs配置。	No（在fileset标签内）
<rules>	针对策略定义多个规则。	No（在policy标签内）
<rule>	定义单一规则。	Yes（在rules标签内）
<age>or<atime>	定义在<fileset>中定义的文件的age/accesstime。策略将匹配该age。age可以用[num]y[num]m[num]w[num]d[num]h的格式表示。其中num表示数字。 其中字母的意思如下： * y--年（一年是365天）。 * m--月（一个月是30天）。 * w--周（一周是7天）。 * d--天。 * h--小时。 可以单独使用年，月，周，天或小时，也可以将他们组合。比如，1y2d表示1年零2天或者367天。 如果没有单位（即数字后面没有任何上述字母），默认单位为天。 **说明** 用户可以在<age>和<atime>标签中配置“gt”（greater）和“lt”（less），默认运算符为“gt”。 示例：<age operator="lt">	No（在rule标签内）
<action>	如果规则匹配，这个标签定义了要执行的action。	No (在rule标签内)
<type>	定义了action类型。当前支持的action类型是MOVE和MARK。	No (在action标签内)
<**params**>	定义与每个action相关的参数。	No (在action标签内)

标签（Tag）名称	描述	是否可重复使用
<param>	定义单个使用<name>和<value>标签的name-value格式参数。 对于MARK和MOVE，只支持参数名“targettier”。该参数表示如果满足age规则，则指定数据存储策略。 如果多个param中具有相同name的参数，则采用第一个参数值。 对于MARK，支持的“targettier”参数值为“ALL_SSD”，“ONE_SSD”，“HOT”，“WARM”，“COLD”。 对于MOVE，支持的“targettier”参数值为“ALL_SSD”，“ONE_SSD”，“HOT”，“WARM”和“COLD”。	Yes (在params标签内).

对于在<file>标签下的文件/文件夹使用FileSystem#globStatus API，对于其他使用GlobPattern类（被GlobFilter使用）。参照支持的API的细节。例如，对于globStatus，“/opt/hadoop/*”将匹配“/opt/hadoop”文件夹下的一切。“/opt/*/hadoop”将匹配/“opt”目录的子目录下的所有hadoop文件夹。

对于globStatus，分别匹配每个路径组件的glob模式，而对于其他的，直接匹配glob模式。

https://hadoop.apache.org/docs/r3.1.1/api/org/apache/hadoop/fs/FileSystem.html#globStatus(org.apache.hadoop.fs.Path)

Glob	Name	Matches
*	*asterisk*	Matches zero or more characters
?	*question mark*	Matches a single character
[ab]	*character class*	Matches a single character in the set {a, b}
[^ab]	*negated character class*	Matches a single character that is not in the set {a, b}
[a-b]	*character range*	Matches a single character in the (closed) range [a, b], where a is lexicographically less than or equal to b
[^a-b]	*negated character range*	Matches a single character that is not in the (closed) range [a, b], where a is lexicographically less than or equal to b
{a,b}	*alternation*	Matches either expression a or b
\c	*escaped character*	Matches character c when it is a metacharacter

行为操作示例

- MARK

```
<action>
  <type>MARK</type>
  <params>
    <param>
      <name>targettier</name>
      <value>HOT</value>
```

```
    </param>
  </params>
</action>
```

- MOVE

```
<action>
  <type>MOVE</type>
  <params>
    <param>
      <name>targettier</name>
      <value>HOT</value>
    </param>
    <param>
      <name>targetnodelabels</name>
      <value>SOME_EXPRESSION</value>
    </param>
  </params>
</action>
```

- SET_REPL

```
<action>
  <type>SET_REPL</type>
  <params>
    <param>
      <name>replcount</name>
      <value>5</value>
    </param>
  </params>
</action>
```

- MOVE_TO_FOLDER

```
<action>
  <type>MOVE_TO_FOLDER</type>
  <params>
    <param>
      <name>target</name>
      <value>path</value>
    </param>
    <param>
      <name>overwrite</name>
      <value>true</value>
    </param>

  </params>
</action>
```

说明

MOVE_TO_FOLDER操作只是将文件路径更改为目标文件夹，不会更改块位置。如果想要移动块，则需要配置一个独立的move策略。

- DELETE

```
<action>
  <type>DELETE</type>
</action>
```

说明

- 在编写xml文件时，用户应该注意行为操作的配置和顺序。冷热数据迁移工具按照输入xml中给定的顺序执行规则。
- 如果只希望运行基于atime/age的一个规则，则按照时间逆序排列，且将idempotent属性设置为false。
- 如果为文件集配置删除操作，则在删除操作后不能再配置其他规则。
- 支持使用"-fs"选项，用于指定客户端默认的文件系统地址。

审计日志

冷热数据迁移工具支持以下操作的审计日志。

- 工具启动状态
- 行为类型及参数详细信息和状态
- 工具完成状态

对于启用审计日志工具，在“<HADOOP_CONF_DIR>/log4j.property”文件中添加以下属性。

```
autodatatool.logger=INFO, ADMTRFA
autodatatool.log.file=HDFSAutoDataMovementTool.audit
log4j.logger.com.xxx.hadoop.hdfs.datamovement.HDFSAutoDataMovementTool.audit=${autodatatool.logger}
log4j.additivity.com.xxx.hadoop.hdfs.datamovement.HDFSAutoDataMovementTool-audit=false
log4j.appender.ADMTRFA=org.apache.log4j.RollingFileAppender
log4j.appender.ADMTRFA.File=${hadoop.log.dir}/${autodatatool.log.file}
log4j.appender.ADMTRFA.layout=org.apache.log4j.PatternLayout
log4j.appender.ADMTRFA.layout.ConversionPattern=%d{ISO8601} %p %c: %m%n
log4j.appender.ADMTRFA.MaxBackupIndex=10
log4j.appender.ADMTRFA.MaxFileSize=64MB
```

说明

具体请参考“<HADOOP_CONF_DIR>/log4j_autodata_movment_template.properties”文件。

12.12.10.3 Hive on Hue

Hue提供了Hive图形化管理功能，使用户可以通过界面的方式查询Hive的不同数据。

查询编辑器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

在左侧导航栏单击编辑器，然后选择“Hive”，进入“Hive”。

- 执行Hive HQL语句

  在左侧选中目标数据库，也可通过单击右上角的 default ，输入目标数据库的名称以搜索目标数据库。

  在文本编辑框输入Hive HQL语句，单击 或者按“Ctrl+Enter”，运行HQL语句，执行结果将在“结果”页签显示。

- 分析HQL语句

  在左侧选中目标数据库，在文本编辑框输入Hive HQL语句，单击 编译HQL语句并显示语句是否正确，执行结果将在文本编辑框下方显示。

- 保存HQL语句

  在文本编辑框输入Hive HQL语句，单击右上角的 ，并输入名称和描述。已保存的语句可以在“保存的查询”页签查看。

- 查看历史

  单击“查询历史记录”，可查看HQL运行情况，支持显示所有语句或只显示保存的语句的运行情况。历史记录存在多个结果时，可以在输入框使用关键字进行搜索。

- 高级查询配置

单击右上角的，对文件、函数、设置等信息进行配置。

- 查看快捷键

  单击右上角的，可查看所有快捷键信息。

元数据浏览器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

- 查看Hive表的元数据

  在左侧导航栏单击表，单击某一表名称，界面将显示Hive表的元数据信息。

- 管理Hive表的元数据

  在Hive表的元数据信息界面：

  - 单击右上角的“导入”可导入数据。
  - 单击“概述”，在“属性”域可查看表文件的位置信息。

    可查看Hive表各列字段的信息，并手动添加描述信息，注意此处添加的描述信息并不是Hive表中的字段注释信息（comment）。
  - 单击“样本”可浏览数据。

- 管理Hive元数据表

  单击左侧列表中的可在数据库中根据上传的文件创建一个新表，也可手动创建一个新表。

注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

典型场景

通过Hue界面对Hive进行创建表的操作如下：

步骤1 单击Hue的WebUI界面左上角的，选择要操作的Hive实例，进入Hive命令的执行页面。

步骤2 在命令输入框内输入一条HQL语句，例如：

create table hue_table(id int,name string,company string) row format delimited fields terminated by ',' stored as textfile;

单击 执行HQL。

步骤3 在命令输入框内输入：

show tables;

单击 ，查看“结果”中有创建的表hue_table。

----结束

12.12.10.4 Oozie on Hue

Hue提供了Oozie作业管理器功能，使用户可以通过界面图形化的方式使用Oozie。

注意

Hue界面主要用于文件、表等数据的查看与分析，禁止通过Hue界面对操作对象进行删除等高危管理操作。如需操作，建议在确认对业务没有影响后通过各组件的相应操作方法进行处理，例如使用HDFS客户端对HDFS文件进行操作，使用Hive客户端对Hive表进行操作。

Oozie 作业设计器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

在左侧导航栏单击  ，选择“Workflow”。

在作业设计器，支持用户创建MapReduce、Java、Streaming、Fs、Ssh、Shell和DistCp作业。

仪表板使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

选择右上角“作业”，进入“作业浏览器”。

支持查看Workflow、Coordinator和Bundles作业的运行情况。

编辑器使用介绍

访问Hue WebUI，请参考**访问Hue的WebUI**。

在左侧导航栏单击  ，然后选择“Workflow”。

支持创建Workflow、计划和Bundles的操作。支持提交运行、共享、复制和导出已创建的应用。

- 每个Workflow可以包含一个或多个作业，形成完整的工作流，用于实现指定的业务。

  创建Workflow时，可直接在Hue的编辑器设计作业，并添加到Workflow中。

- 每个计划可定义一个时间触发器，用于定时触发执行一个指定的Workflow。不支持多个Workflow。
- 每个Bundles可定义一个集合，用于触发执行多个计划，使不同Workflow批量执行。

12.12.11 Hue 日志介绍

日志描述

日志路径：Hue相关日志的默认存储路径为“/var/log/Bigdata/hue”（运行日志），“/var/log/Bigdata/audit/hue”（审计日志）。

日志归档规则：Hue的日志启动了自动压缩归档功能，默认情况下，当“access.log”、“error.log”、“runcpserver.log”和“hue-audits.log”大小超过5MB的时候，会自动压缩。最多保留最近的20个压缩文件，压缩文件保留个数和压缩文件阈值可以配置。

表 12-253 Hue 日志列表

日志类型	日志文件名	描述
运行日志	access.log	访问日志。
	error.log	错误日志。
	gsdb_check.log	gaussDB检查日志。
	kt_renewer.log	Kerberos认证日志。
	kt_renewer.out.log	Kerberos认证日志的异常输出日志。
	runcpserver.log	操作记录日志。
	runcpserver.out.log	进程运行异常日志。
	supervisor.log	进程启动日志。
	supervisor.out.log	进程启动异常日志。
	dbDetail.log	数据库初始化日志
	initSecurityDetail.log	keytab文件下载初始化日志。
	postinstallDetail.log	Hue服务安装后工作日志。
	prestartDetail.log	Prestart日志。
	statusDetail.log	Hue服务健康状态日志。
	startDetail.log	启动日志。
	get-hue-ha.log	Hue HA状态日志。
	hue-ha-status.log	Hue HA状态监控日志。
	get-hue-health.log	Hue健康状态日志。

日志类型	日志文件名	描述
	hue-health-check.log	Hue健康检查日志。
	hue-refresh-config.log	Hue配置刷新日志。
	hue-script-log.log	Manager界面的Hue操作日志。
	hue-service-check.log	Hue服务状态监控日志。
	db_pwd.log	Hue连接DBService数据库密码修改日志
	modifyDBPwd_*日期*.log	-
	watch_config_update.log	参数更新日志。
审计日志	hue-audits.log	审计日志。

日志级别

Hue提供了如**表12-254**所示的日志级别。

日志的级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-254 日志级别

级别	描述
ERROR	ERROR表示系统运行的错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 参考**修改集群服务配置参数**进入Hue服务“全部配置”页面。

步骤2 在左侧导航栏选择需修改的角色所对应的“日志”菜单。

步骤3 在右侧选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

步骤5 重新启动配置过期的服务或实例以使配置生效。

----结束

日志格式

Hue的日志格式如下所示：

表 12-255 日志格式

日志类型	格式	示例
运行日志	<dd-MM-yy HH:mm:ss,SSS><日志事件的发生位置><log level><log中的message>	[03/Nov/2014 11:57:19] middleware | INFO | Unloading MimeTypeJSFileFixStreamingMiddleware.
	<Log Level><时间格式><yyyy-MM-dd HH:mm:ss,SSS><日志事件的发生位置><log中的message>	INFO : CST 2014-11-06 11:22:52 hue-ha-status.sh : update 4 <= 15:myHostName=10.0.0.250 ACTIVE=10.0.0.250
审计日志	<UserName><yyyy-MM-dd HH:mm:ss,SSS><审计操作描述><资源参数><url><是否允许><审计操作><ip地址>	{"username": "admin", "eventTime": "2014-11-06 10:28:34", "operationText": "Successful login for user: admin", "service": "accounts", "url": "/accounts/login/", "allowed": true, "operation": "USER_LOGIN", "ipAddress": "10.0.0.250"}

12.12.12 Hue 常见问题

12.12.12.1 如何解决使用 IE 浏览器在 Hue 中执行 HQL 失败的问题

问题

遇到使用IE浏览器在Hue中访问Hive Editor并执行所有HQL失败，界面提示“There was an error with your query.”，如何解决并正常执行HQL?

回答

IE浏览器存在功能问题，不支持在307重定向中处理含有form data的AJAX POST请求，建议更换兼容的浏览器，例如Google Chrome浏览器。

12.12.12.2 在使用 Hive 时，输入 use database 语句失效了

问题

使用Hive的时候，在输入框中输入了**use database**的语句切换数据库，重新在输入框内输入其他语句，为什么数据库没有切换过去?

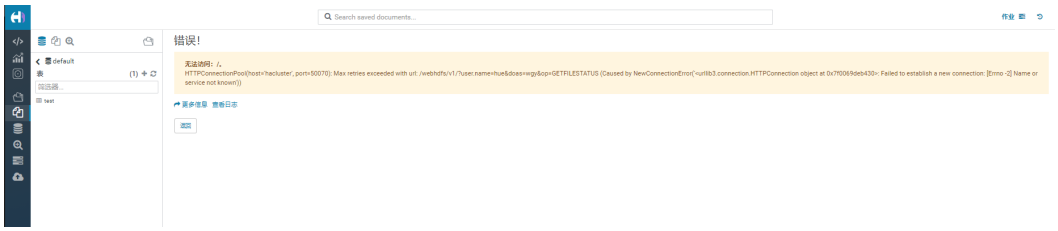
回答

在Hue上使用Hive有区别于用Hive客户端使用Hive，Hue界面上有选择数据库的按钮，当前SQL执行的数据库以界面上显示的数据库为准。与此相关的还有设置参数等session级别的一次性操作，都应该使用界面功能进行设置，不建议使用输入语句进行操作。若是必须使用输入语句进行操作，需保证所有语句在同一个输入框内。

12.12.12.3 如何处理使用 Hue WebUI 访问 HDFS 文件失败的问题

问题

在使用Hue WebUI访问HDFS文件时，报如下图所示无法访问的错误提示，该如何处理?



回答

1. 查看登录Hue WebUI的用户是否具有“hadoop”用户组权限。
2. 查看HDFS服务是否安装了HttpFS实例且运行正常。如果未安装HttpFS实例，需手动安装并重启Hue服务。

12.12.12.4 Hue 页面上传大文件失败如何处理

问题

通过Hue页面上传大文件时，上传失败。

回答

1. 不建议使用Hue文件浏览器上传大文件，大文件建议使用客户端通过命令上传。
2. 如果必须使用Hue上传，参考以下步骤修改Httpd的参数：
   a. 以**omm**用户登录主管理节点。
   b. 执行以下命令编辑“httpd.conf”配置文件。
      vi $BIGDATA_HOME/om-server/Apache-httpd-*/conf/httpd.conf
   c. 搜索21201，在</VirtualHost>配置中加上“RequestReadTimeout handshake=0 header=0 body=0”，如下所示。

```
...
<VirtualHost *:21201>
    ServerName https://10.112.16.93:21201
    AllowEncodedSlashes On
    SSLProxyEngine On
    ProxyRequests Off
    TraceEnable off
    ProxyTimeout  1200
    RewriteEngine on
    RewriteMap proxylist dbm:${BIGDATA_ROOT_HOME}/om-server_*/Apache-httpd-*/conf/
proxylist.dbm
```

```
    RewriteRule ^(\/.*)$  ${proxylist:/Hue/Hue/21201}$1 [E=TARGET_PATH:$1,L,P]

    Header edit Location ^(?!https://10.112.16.93:20009|https://
10.112.16.93:21201)http[s]?://[^/]*(.*)$  https://10.112.16.93:21201$1

    ProxyPassReverseCookiePath / / interpolate

    SSLEngine On
    SSLProxyProtocol  All +TLSv1.2 -SSLv2 -SSLv3 -TLSv1 -TLSv1.1
    SSLProtocol ALL +TLSv1.2 -SSLv2 -SSLv3 -TLSv1 -TLSv1.1
    SSLCipherSuite ECDHE-RSA-AES256-GCM-SHA384:ECDHE-ECDSA-AES256-GCM-
SHA384:ECDHE-RSA-AES128-GCM-SHA256:ECDHE-ECDSA-AES128-GCM-SHA256:DHE-DSS-
AES256-GCM-SHA384:DHE-RSA-AES256-GCM-SHA384:DHE-DSS-AES128-GCM-SHA256:DHE-
RSA-AES128-GCM-SHA256
    SSLProxyCheckPeerName off
    SSLProxyCheckPeerCN off
    SSLCertificateFile "${BIGDATA_ROOT_HOME}/om-server_*/Apache-httpd-*/conf/security/
proxy_ssl.cert"
    SSLCertificateKeyFile "${BIGDATA_ROOT_HOME}/om-server_*/Apache-httpd-*/conf/security/
server.key"
    SSLProxyCACertificateFile ${BIGDATA_ROOT_HOME}/om-server_*/apache-tomcat-*/conf/
security/tomcat.crt
    SSLCertificateChainFile "${BIGDATA_ROOT_HOME}/om-server_*/Apache-httpd-2.4.39/conf/
security/proxy_chain.cert"
    RequestReadTimeout handshake=0 header=0 body=0
</VirtualHost>
...
```

d. 执行**ps -ef|grep httpd|grep -v grep|xargs kill -9**命令重启httpd。

12.12.12.5 集群未安装 Hive 服务时 Hue 原生页面无法正常显示

问题

集群没有安装Hive服务时，Hue服务原生页面显示空白。

回答

MRS 3.x版本存在Hue依赖Hive组件，如果出现此情况，首先需要检查当前集群是否安装了Hive组件，如果没有，需要安装Hive。

12.13 使用 Impala

12.13.1 从零开始使用 Impala

Impala是用于处理存储在Hadoop集群中的大量数据的MPP（大规模并行处理）SQL查询引擎。 它是一个用C++和Java编写的开源软件。 与其他Hadoop的SQL引擎相比，它拥有高性能和低延迟的特点。

背景信息

假定用户开发一个应用程序，用于管理企业中的使用A业务的用户信息，使用Impala客户端实现A业务操作流程如下：

普通表的操作：

- 创建用户信息表user_info。

- 在用户信息中新增用户的学历、职称信息。
- 根据用户编号查询用户姓名和地址。
- A业务结束后，删除用户信息表。

表 12-256 用户信息

编号	姓名	性别	年龄	地址
12005000201	A	男	19	A城市
12005000202	B	女	23	B城市
12005000203	C	男	26	C城市
12005000204	D	男	18	D城市
12005000205	E	女	21	E城市
12005000206	F	男	32	F城市
12005000207	G	女	29	G城市
12005000208	H	女	30	H城市
12005000209	I	男	26	I城市
12005000210	J	女	25	J城市

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 运行Impala客户端命令，实现A业务。

直接执行Impala组件的客户端命令：

impala-shell

说明

默认情况下，**impala-shell**尝试连接到localhost的21000端口上的Impala守护程序。如需连接到其他主机，请使用**-i <host:port>**选项，例如：impala-shell -i xxx.xxx.xxx.xxx:21000。要自动连接到特定的Impala数据库，请使用**-d <database>**选项。例如，如果您的所有Kudu表都位于数据库“impala_kudu”中，则**-d impala_kudu**可以使用此数据库。要退出Impala Shell，请使用**quit**命令。

内部表的操作：

1. 根据表12-256创建用户信息表user_info并添加相关数据。
   ```
   create table user_info(id string,name string,gender string,age int,addr string);
   insert into table user_info(id,name,gender,age,addr) values("12005000201","A","男",19,"A城市");
   ```
   .....（其他语句相同）
2. 在用户信息表user_info中新增用户的学历、职称信息。
   以增加编号为12005000201的用户的学历、职称信息为例，其他用户类似。
   ```
   alter table user_info add columns(education string,technical string);
   ```
3. 根据用户编号查询用户姓名和地址。
   以查询编号为12005000201的用户姓名和地址为例，其他用户类似。
   ```
   select name,addr from user_info where id='12005000201';
   ```
4. 删除用户信息表。
   ```
   drop table user_info;
   ```

外部分区表的操作：

创建外部分区表并导入数据

1. 创建外部表数据存储路径。
   - 安全模式（集群开启了Kerberos认证）：
     cd /opt/hadoopclient
     source bigdata_env
     kinit *hive*

     说明

     用户hive需要具有Hive管理员权限。

     impala-shell
     hdfs dfs -mkdir /hive
     hdfs dfs -mkdir /hive/user_info
   - 普通模式（集群关闭了Kerberos认证）：
     su - omm
     cd /opt/hadoopclient
     source bigdata_env
     impala-shell
     hdfs dfs -mkdir /hive
     hdfs dfs -mkdir /hive/user_info
2. 建表。
   ```
   create external table user_info(id string,name string,gender string,age int,addr string) partitioned by(year string) row format delimited fields terminated by ' ' lines terminated by '\n' stored as textfile location '/hive/user_info';
   ```

说明

fields terminated指明分隔的字符,如按空格分隔，' '。

lines terminated 指明分行的字符，如按换行分隔，'\n'。

/hive/user_info为数据文件的路径。

3. 导入数据。
   a. 使用insert语句插入数据。
   ```
   insert into user_info partition(year="2018") values ("12005000201","A","男",19,"A城市");
   ```
   b. 使用load data命令导入文件数据。
      i. 根据表12-256数据创建文件。如，文件名为txt.log，以空格拆分字段，以换行符作为行分隔符。
      ii. 上传文件至hdfs。
      hdfs dfs -put txt.log /tmp
      iii. 加载数据到表中。
      load data inpath '/tmp/txt.log' into table user_info partition (year='2018');
4. 查询导入数据。
   ```
   select * from user_info;
   ```
5. 删除用户信息表。
   ```
   drop table user_info;
   ```

----结束

12.13.2 访问 Impala 的 WebUI

用户可以通过Impala的WebUI，在图形化界面查看Impala作业的相关信息。Impala的WebUI根据实例不同分为如下三种：

- StateStore WebUI：用于管理节点。
- Catalog WebUI：用于查看元数据。
- Impalad WebUI：用于查看每个SQL执行的详细信息。

前提条件

已安装Impala服务的集群。

访问 StateStore WebUI

步骤1 登录Manager页面，请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 选择“服务管理 > Impala”。

步骤3 在“Impala 概述”的“StateStore WebUI”中单击“StateStore(Statestore)”，打开StateStore的WebUI页面。

----结束

访问 Catalog WebUI

步骤1 登录Manager页面，请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 选择“服务管理 > Impala”。

步骤3 在“Impala 概述”的“Catalog WebUI”中单击“Catalog(Catalog)”，打开Catalog的WebUI页面。

----结束

访问 Impalad WebUI

步骤1 登录Manager页面，请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 选择“服务管理 > Impala > 实例”。

步骤3 移动鼠标至“角色”列的Impalad实例上，在页面左下角显示如下链接，获取null后的数值，例如本例中的82。

https://*EIP*:9022/mrsmanager/index.jsp?locale=zh-cn#/app/services/Impala/Impalad/null/82/*EIP*/STARTED/status/detail

其中82为样例值，实际值请以实际环境为准。

步骤4 参考**访问StateStore WebUI**。

步骤5 修改StateStore WebUI的URL地址中的“StateStore/xx”为“Impalad/**xx**”并访问修改后的URL，其中**xx**为**步骤3**中获取的数值。

----结束

12.13.3 使用 Impala 操作 Kudu

您可以使用Impala的SQL语法插入、查询、更新和删除Kudu中的数据，作为使用Kudu API构建自定义Kudu应用程序的替代方案。

前提条件

已安装集群完整客户端。例如安装目录为“/opt/Bigdata/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

Impala on Kudu

步骤1 登录安装客户端的节点。

步骤2 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤3 若集群开启Kerberos认证，请执行如下步骤认证用户。若集群未开启Kerberos认证请跳过该步骤。

kinit *业务用户*

步骤4 执行如下命令登录impala客户端。

impala-shell

说明

默认情况下，**impala-shell**尝试连接到localhost的21000端口上的Impala守护程序。如需连接到其他主机，请使用**-i <host:port>**选项。要自动连接到特定的Impala数据库，请使用**-d <database>**选项。例如，如果您的所有Kudu表都位于数据库“impala_kudu”中，则**-d impala_kudu**可以使用此数据库。要退出Impala Shell，请使用以下命令**quit**。

步骤5 执行如下命令创建Impala表并导入已准备好的数据，例如/tmp/data10。

create table dataorigin (name string,age string,pt string, date_p date) row format delimited fields terminated by ',' stored as textfile;

load data inpath '/tmp/data10' overwrite into table dataorigin;

步骤6 执行如下命令创建Kudu表，其中kudu.master_addresses地址为KuduMaster实例的IP，请根据实际集群地址填写。

create table dataorigin2 (name string,age string,pt string, date_p date, primary key(name)) stored as kudu TBLPROPERTIES('kudu.master_addresses'='192.168.190.164:7051,192.168.204.178:7051,192.168.244.63:7051');

步骤7 执行如下命令操作Kudu表。

1. 插入数据

   insert into dataorigin2 select * from dataorigin;

2. 更新数据

   UPDATE dataorigin2 SET date_p="2021-03-31" where age="73";

3. 更新或插入行

   UPSERT INTO dataorigin2 VALUES ("spjted","75","28","2021-03-32");

   UPSERT INTO dataorigin2 VALUES ("kwhakb","92","29","2021-03-33");

   UPSERT INTO dataorigin2 VALUES ("oftrkf","13","30","2021-03-34");

   UPSERT INTO dataorigin2 VALUES ("kiewti","36","31","2021-03-35");

   UPSERT INTO dataorigin2 VALUES ("rknmql","98","32","2021-03-36");

   UPSERT INTO dataorigin2 VALUES ("fwcoij","52","33","2021-03-37");

   UPSERT INTO dataorigin2 VALUES ("pgvpdo","37","34","2021-03-35");

4. 删除行

   DELETE FROM dataorigin2 WHERE date_p="2021-03-31";

----结束

12.13.4 Impala 对接外部 LDAP

本操作适用于MRS 3.1.0及之后版本。

步骤1 登录Manager。

步骤2 在Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Impala > 配置 > 全部配置 > Impalad（角色） > LDAP”。

步骤3 配置如下参数的值。

表 12-257 参数配置

参数名称	参数描述	备注
--enable_ldap_auth	是否开启LDAP认证	【 取值范围 】 true或false
--ldap_bind_pattern	LDAP userDNPattern	例如：cn=#UID,ou=People,dc=xxx,dc=com或cn=%s,ou=People,dc=xxx,dc=com
--ldap_passwords_in_clear_ok	LDPA 密码是否以明文发送	如果设置为true，将允许LDAP密码在网络上明文发送 【 取值范围 】 true或false **说明** 当“--enable_ldap_auth”设置为“true”时，认证时默认没有开启Ldap TLS协议，所以需要将“--ldap_passwords_in_clear_ok”参数设置为“true”，否则会导致Impalad角色启动失败。 如需开启Ldap TLS协议则需要在Impalad角色的自定义配置中添加配置项“--ldap_tls”为“true”，配置之后密码将支持用密文传输。
--ldap_uri-ip	LDAP IP	-
--ldap_uri-port	LDAP 端口	【 默认值 】389

步骤4 修改完成后，单击左上方“保存”，在弹出的对话框中单击“确定”保存配置。

步骤5 选择“集群 > *待操作集群的名称* > 服务 > Impala > 实例”，勾选配置状态为“配置过期”的实例，选择“更多 > 重启实例”重启受影响的Impala实例。

----结束

12.14 使用 Kafka

12.14.1 从零开始使用 Kafka

操作场景

用户可以在集群客户端完成Topic的创建、查询、删除等基本操作。

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

使用 Kafka 客户端（MRS 3.x 之前版本）

步骤1 进入ZooKeeper实例页面：

单击集群名称，登录集群详情页面，选择“组件管理 > ZooKeeper > 实例”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient/Kafka/kafka/bin”。

cd /opt/hadoopclient/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/hadoopclient/bigdata_env

步骤6 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Kafka用户*

步骤7 创建一个Topic：

sh kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份个数* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤8 执行以下命令，查询集群中的Topic信息：

sh kafka-topics.sh --list --zookeeper *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤9 删除**步骤7**中创建的Topic：

sh kafka-topics.sh --delete --topic *主题名称* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

输入 "y"，回车。

----结束

使用 Kafka 客户端（MRS 3.x 及之后版本）

步骤1 进入ZooKeeper实例页面：

登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 实例”。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient/Kafka/kafka/bin”。

cd /opt/hadoopclient/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/hadoopclient/bigdata_env

步骤6 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Kafka用户*

步骤7 登录FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 配置 > 全部配置”，搜索参数“clientPort”,记录“clientPort”的参数值。

步骤8 创建一个Topic：

sh kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份个数* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤9 执行以下命令，查询集群中的Topic信息：

sh kafka-topics.sh --list --zookeeper *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

步骤10 删除步骤8中创建的Topic：

sh kafka-topics.sh --delete --topic *主题名称* **--zookeeper** *ZooKeeper角色实例所在节点IP地址.clientPort*/**kafka**

----结束

12.14.2 管理 Kafka 主题

操作场景

用户可以根据业务需要，使用集群客户端管理Kafka的主题。启用Kerberos认证的集群，需要拥有管理Kafka主题的权限。

前提条件

已安装客户端。

操作步骤

步骤1 进入ZooKeeper实例页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > ZooKeeper > 实例”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > ZooKeeper > 实例”。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/client/Kafka/kafka/bin”。

cd /opt/client/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤6 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤7 MRS 3.x之前版本：分别执行以下命令，管理Kafka主题。

- 创建主题

  sh kafka-topics.sh --create --topic ***主题名称*** **--partitions** ***主题占用的分区数*** **--replication-factor** ***主题的备份个数*** **--zookeeper** ***ZooKeeper角色实例所在节点IP地址:clientPort*****/kafka**

- 删除主题

  sh kafka-topics.sh --delete --topic ***主题名称*** **--zookeeper** ***ZooKeeper角色实例所在节点IP地址:clientPort*****/kafka**

说明

- 主题分区数和主题备份个数不能大于Kafka角色实例数量。
- 默认情况下，ZooKeeper的“clientPort”为“2181”。
- ZooKeeper角色实例所在节点IP地址，填写三个角色实例其中任意一个的IP地址即可。
- 使用Kafka主题管理消息，请参见**管理Kafka主题中的消息**。

步骤8 MRS 3.x及后续版本：使用**kafka-topics.sh**管理Kafka主题。

- 创建主题：

  Topic的Partition自动划分时，默认根据节点及磁盘上已有的Partition数进行均衡划分，如果期望根据磁盘容量进行Partition划分，那么需要修改Kafka服务配置“log.partition.strategy”为“capacity”。

  Kafka创建Topic时，支持基于“机架感知” 和 “跨AZ特性” 两种选项组合生成分区及副本的分配方案且支持 “--zookeeper” 和 “--bootstrap-server”两种方式

  – 禁用机架策略 & 禁用跨AZ特性 (默认策略)。

    基于此策略新建的Topic的副本会**完**全随机分配到集群中任意节点上。

./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka**

./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties**

其中，使用“--bootstrap-server”方式创建Topic时，需配置“rack.aware.enable=false”和“az.aware.enable=false”。

- 启用机架策略 & 禁用跨AZ特性。

  基于此策略新建的Topic的各个Partition的Leader会在集群节点上随机分配，但会确保同一Partition的不同Replica会分配在不同的机架上，所以当使用此策略时，需保证各个机架内的节点个数一致，否则会导致节点少的机架上的机器负载远高于集群平均水平。

  ./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka --enable-rack-aware**

  ./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties**

  其中，使用 “--bootstrap-server”方式创建Topic时，需配置“rack.aware.enable=true”和“az.aware.enable=false”。

- 禁用机架策略 & 启用跨AZ特性。

  基于此策略新建的Topic的各个Partition的Leader会在集群节点上随机分配，但会确保同一Partition的不同Replica会分配在不同的AZ上，所以当使用此策略时，需保证各个AZ内的节点个数一致，否则会导致节点少的AZ上的机器负载远高于集群平均水平。

  ./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka --enable-az-aware**

  ./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties**

  其中，使用 “--bootstrap-server”方式创建Topic时，需配置“rack.aware.enable=false”和“az.aware.enable=true”。

- 启用机架策略 & 启用跨AZ特性。

  基于此策略新建的Topic的各个Partition的Leader会在集群节点上随机分配，但会确保同一Partition的不同Replica会分配到不同AZ内的不同RACK上，使用此策略需保证每个AZ内的每个RACK上的节点个数一致，否则会导致集群内负载不均衡。

  ./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka --enable-rack-aware --enable-az-aware**

  ./kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份数* **--bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties**

  使用 “--bootstrap-server”方式创建Topic时，需配置“rack.aware.enable=true”和“az.aware.enable=true”。

说明

- Kafka创建Topic支持 “--zookeeper” 和 “--bootstrap-server”两种方式，区别如下：
  - “--zookeeper”方式由客户端生成副本分配方案，社区从一开始就支持这种方式，为了降低对Zookeeper组件的依赖，社区将在后续版本中删除对这种方式的支持。基于这种方式创建Topic时，可以通过 "--enable-rack-aware" 和 “--enable-az-aware”这两个选项自由组合来选用副本分配策略。注意：使用“--enable-az-aware”选项的前提是服务端开启了跨AZ特性，即服务端启动参数“az.aware.enable”为“true”,否则会执行失败。
  - “--bootstrap-server”方式由服务端生成副本分配方案，后续版本，社区将只支持这种方式来进行Topic管理。基于这种方式创建Topic时，不支持 "--enable-rack-aware" 和 "--enable-az-aware"选项来控制副本分配策略，支持“rack.aware.enable”和“az.aware.enable”这两个服务启动参数组合来控制副本分配策略，需注意的是“az.aware.enable”参数不可修改，在创建集群时，如果开启跨AZ特性，会自动配置为“true”；“rack.aware.enable”参数支持用户自定义修改。

- 罗列主题：
  - **./kafka-topics.sh --list --zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka**
  - **./kafka-topics.sh --list --bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties**
- 查看主题：
  - **./kafka-topics.sh --describe --zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka** --**topic** *主题名称*
  - **./kafka-topics.sh --describe --bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties --topic** *主题名称*
- 修改主题：
  - **./kafka-topics.sh --alter --topic** *主题名称* **--config** *配置项=配置值* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka**
- 扩展分区：
  - **./kafka-topics.sh --alter --topic** *主题名称* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka --command-config Kafka/kafka/config/client.properties --partitions** *扩展后分区个数*
  - **./kafka-topics.sh --alter --topic** *主题名称* **--bootstrap-server** *Kafka集群IP:21007* **--command-config Kafka/kafka/config/client.properties --partitions** *扩展后分区个数*
- 删除主题：
  - **./kafka-topics.sh --delete --topic** *主题名称* **--zookeeper** *ZooKeeper的任意一个节点的业务IP:clientPort*/**kafka**
  - **./kafka-topics.sh --delete --topic** *主题名称* **--bootstrap-server** *Kafka集群IP:21007* **--command-config ../config/client.properties**

----结束

12.14.3 查看 Kafka 主题

操作场景

用户可以在MRS上查看Kafka已创建的主题信息。

操作步骤

步骤1 进入Kafka服务页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Kafka”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Kafka”。

步骤2 单击“KafkaTopic监控”。

主题列表默认显示所有主题。可以查看主题的分区数和备份数。

步骤3 在主题列表单击指定主题的名称，可查看详细信息。

----结束

12.14.4 管理 Kafka 用户权限

操作场景

在启用Kerberos认证的集群中，用户使用Kafka前需要拥有对应的权限。MRS集群支持将Kafka的使用权限，授予不同用户。

Kafka默认用户组如**表12-258**所示。

说明

在MRS 3.x及之后版本中，Kafka支持两种鉴权插件：“Kafka开源自带鉴权插件”和“Ranger鉴权插件”。

本章节描述的是基于“Kafka开源自带鉴权插件”的用户权限管理。若想使用 “Ranger鉴权插件”，请参考**添加Kafka的Ranger访问权限策略**。

表 12-258 Kafka 默认用户组

用户组名称	描述
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有主题的创建，删除，授权及读写权限。
kafkasuperuser	Kafka高级用户组。添加入本组的用户，拥有所有主题的读写权限。

用户组名称	描述
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定主题的读写权限，才能访问对应主题。

前提条件

- 已安装客户端。
- 用户已明确业务需求，并准备一个属于kafkaadmin组的用户，作为Kafka管理员用户。例如“admin”。

操作步骤

步骤1 进入ZooKeeper实例页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > ZooKeeper > 实例”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > ZooKeeper > 实例”。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/client/Kafka/kafka/bin”。

cd /opt/client/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤6 执行以下命令，进行用户认证。

kinit *组件业务用户*

步骤7 MRS 3.x之前版本：选择业务需要对应的场景，管理Kafka用户权限。

- 查看某个主题的权限控制列表

  sh kafka-acls.sh --authorizer-properties zookeeper.connect=*ZooKeeper角色实例所在节点IP地址*:2181/kafka --list --topic *主题名称*

- 为某个用户添加生产者的权限

  sh kafka-acls.sh --authorizer-properties zookeeper.connect=*ZooKeeper角色实例所在节点IP地址*:2181/kafka --add --allow-principal User:*用户名* --producer --topic *主题名称*

- 删除某个用户的生产者权限

  sh kafka-acls.sh --authorizer-properties zookeeper.connect=*ZooKeeper角色实例所在节点IP地址*:2181/kafka --remove --allow-principal User:*用户名* --producer --topic *主题名称*

- 为某个用户添加消费者的权限

  sh kafka-acls.sh --authorizer-properties zookeeper.connect=*ZooKeeper角色实例所在节点IP地址*:2181/kafka --add --allow-principal User:*用户名* --consumer --topic *主题名称* --group *消费者组名称*

- 删除某个用户的消费者权限

  sh kafka-acls.sh --authorizer-properties zookeeper.connect=*ZooKeeper角色实例所在节点IP地址*:2181/kafka --remove --allow-principal User:*用户名* --consumer --topic *主题名称* --group *消费者组名称*

  说明

  删除权限时需要输入两次“y”确认删除权限。

步骤8 MRS 3.x及后续版本：使用“kafka-acl.sh”进行用户授权常用命令如下。

- 查看某Topic权限控制列表：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--list --topic *<Topic名称>***

  ./kafka-acls.sh --bootstrap-server <*Kafka集群IP:21007*> **--command-config ../config/client.properties --list --topic** <*Topic名称*>

- 添加给某用户Producer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--add --allow-principal User:***<用户名>* **--producer --topic** *<Topic名称>*

  ./kafka-acls.sh --bootstrap-server <*Kafka集群IP:21007*> **--command-config ../config/client.properties --add --allow-principal User:***<用户名>* **--producer --topic** <*Topic名称*>

- 给某用户批量添加Producer权限

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--add --allow-principal User:***<用户名>* **--producer --topic** *<Topic名称>* **--resource-pattern-type prefixed**

  ./kafka-acls.sh --bootstrap-server <*Kafka集群IP:21007*> **--command-config ../config/client.properties --add --allow-principal User:***<用户名>* **--producer --topic** *<Topic名称>***--resource-pattern-type prefixed**

- 删除某用户Producer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--remove --allow-principal User:***<用户名>* **--producer --topic** *<Topic名称>*

  ./kafka-acls.sh --bootstrap-server <*Kafka集群IP:21007*> **--command-config ../config/client.properties --remove --allow-principal User:***<用户名>* **--producer --topic** <*Topic名称*>

- 批量删除某用户Producer权限:

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--remove --allow-principal User:***<用户名>* **--producer --topic** *<Topic名称>* **--resource-pattern-type prefixed**

./kafka-acls.sh --bootstrap-server *<Kafka集群IP:21007>* **--command-config ../config/client.properties --remove --allow-principal User:***<用户名>* **--producer --topic** *<Topic名称>***--resource-pattern-type prefixed**

- 添加给某用户Consumer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--add --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>*

  ./kafka-acls.sh --bootstrap-server *<Kafka集群IP:21007>* **--command-config ../config/client.properties --add --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>*

- 给某用户批量添加Consumer权限

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--add --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>* **--resource-pattern-type prefixed**

  ./kafka-acls.sh --bootstrap-server *<Kafka集群IP:21007>* **--command-config ../config/client.properties --add --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>* **--resource-pattern-type prefixed**

- 删除某用户Consumer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--remove --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>*

  ./kafka-acls.sh --bootstrap-server *<Kafka集群IP:21007>* **--command-config ../config/client.properties --remove --allow-principal User:***<用户名>* ***--consumer --topic*** *<Topic名称>* **--group** *<消费者组名称>*

- 批量删除某用户Consumer权限:

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=*<ZooKeeper的任意一个节点的业务IP:21812181/kafka >* **--remove --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>* **--resource-pattern-type prefixed**

  ./kafka-acls.sh --bootstrap-server *<Kafka集群IP:21007>* **--command-config ../config/client.properties --remove --allow-principal User:***<用户名>* **--consumer --topic** *<Topic名称>* **--group** *<消费者组名称>* **--resource-pattern-type prefixed**

----结束

12.14.5 管理 Kafka 主题中的消息

操作场景

用户可以根据业务需要，使用MRS集群客户端，在Kafka主题中产生消息，或消费消息。启用Kerberos认证的集群，需要用户拥有在Kafka主题中执行相应操作的权限。

前提条件

已安装客户端。

操作步骤

步骤1 进入Kafka服务页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理 > Kafka”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名称* > 服务 > Kafka”。

步骤2 单击“实例”，查看Kafka角色实例的IP地址。

记录Kafka角色实例其中任意一个的IP地址即可。

步骤3 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/client/Kafka/kafka/bin”。

cd /opt/client/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

步骤6 启用Kerberos认证的集群，执行以下命令认证用户身份。未启用Kerberos认证的集群无需执行。

kinit *Kafka用户*

例如：

kinit admin

步骤7 根据业务需要，管理Kafka主题中的消息。

- 在主题中产生消息

  sh kafka-console-producer.sh --broker-list *Kafka角色实例所在节点的IP地址*:9092 --topic *主题名称* --producer.config /opt/client/Kafka/kafka/config/producer.properties

  用户可以输入指定的内容作为生产者产生的消息，输入完成后按回车发送消息。如果需要结束产生消息，使用“Ctrl + C”退出任务。

- 消费主题中的消息

  sh kafka-console-consumer.sh --topic *主题名称* --bootstrap-server *Kafka角色实例所在节点的IP地址*:9092 --consumer.config /opt/client/Kafka/kafka/config/consumer.properties

  配置文件中“group.id”指定的消费者组默认为“example-group1”。用户可根据业务需要，自定义其他消费者组。每次消费时生效。

  执行命令时默认会读取当前消费者组中未被处理的消息。如果在配置文件指定了新的消费者组且命令中增加参数“--from-beginning”，则会读取所有Kafka中未被自动删除的消息。

说明

----结束

12.14.6 基于 binlog 的 MySQL 数据同步到 MRS 集群中

本章节为您介绍使用Maxwell同步工具将线下基于binlog的数据迁移到MRS Kafka集群中的指导。

Maxwell是一个开源程序（**https://maxwells-daemon.io**），通过读取MySQL的binlog日志，将增删改等操作转为JSON格式发送到输出端(如控制台/文件/Kafka等)。Maxwell可部署在MySQL机器上，也可独立部署在其他与MySQL网络可通的机器上。

Maxwell运行在Linux服务器上，常见的有EulerOS、Ubuntu、Debian、CentOS、OpenSUSE等，且需要Java 1.8+支持。

同步数据具体内容如下。

1. **配置MySQL**
2. **安装Maxwell**
3. **配置Maxwell**
4. **启动Maxwell**
5. **验证Maxwell**
6. **停止Maxwell**
7. **Maxwell生成的数据格式及常见字段含义**

配置 MySQL

步骤1 开启binlog，在MySQL中打开my.cnf文件，在[mysqld] 区块检查是否配置server_id，log-bin与binlog_format，若没有配置请执行如下命令添加配置项并重启MySQL，若已经配置则忽略此步骤。

```
$ vi my.cnf

[mysqld]
server_id=1
log-bin=master
binlog_format=row
```

步骤2 Maxwell需要连接MySQL，并创建一个名称为maxwell的数据库存储元数据，且需要能访问需要同步的数据库，所以建议新创建一个MySQL用户专门用来给Maxwell使用。使用root登录MySQL之后，执行如下命令创建maxwell用户（其中XXXXXX是密码，请修改为实际值）。

- 若Maxwell程序部署在非MySQL机器上，则创建的maxwell用户需要有远程登录数据库的权限，此时创建命令为

  mysql> GRANT ALL on maxwell.* to 'maxwell'@'%' identified by 'XXXXXX';

  mysql> GRANT SELECT, REPLICATION CLIENT, REPLICATION SLAVE on *.* to 'maxwell'@'%';

- 若Maxwell部署在MySQL机器上，则创建的maxwell用户可以设置为只能在本机登录数据库，此时创建命令为

  mysql> GRANT SELECT, REPLICATION CLIENT, REPLICATION SLAVE on *.* to 'maxwell'@'localhost' identified by 'XXXXXX';

mysql> GRANT ALL on maxwell.* to 'maxwell'@'localhost';

----结束

安装 Maxwell

步骤1 下载安装包，下载路径为**https://github.com/zendesk/maxwell/releases**，选择名为maxwell-XXX.tar.gz的二进制文件下载，其中XXX为版本号。

步骤2 将tar.gz包上传到任意目录下（本示例路径为Master节点的/opt）。

步骤3 登录部署Maxwell的服务器，并执行如下命令进入tar.gz包所在目录。

cd /opt

步骤4 执行如下命令解压“maxwell-XXX.tar.gz”压缩包，并进入“maxwell-XXX”文件夹。

tar -zxvf maxwell-XXX.tar.gz

cd maxwell-XXX

----结束

配置 Maxwell

在maxwell-XXX文件夹下若有conf目录则配置config.properties文件，配置项说明请参见**表12-259**。若没有conf目录，则是在maxwell-XXX文件夹下将config.properties.example修改成config.properties。

表 12-259 Maxwell 配置项说明

配置项	是否必填	说明	默认值
user	是	连接MySQL的用户名，即**步骤2**中新创建的用户	-
password	是	连接MySQL的密码	-
host	否	MySQL地址	localhost
port	否	MySQL端口	3306
log_level	否	日志打印级别，可选值为 • debug • info • warn • error	info
output_ddl	否	是否发送DDL(数据库与数据表的定义修改)事件 • true：发送DDL事件 • false：不发送DDL事件	false

配置项	是否必填	说明	默认值
producer	是	生产者类型，配置为kafka ● stdout：将生成的事件打印在日志中 ● kafka：将生成的事件发送到kafka	stdout
producer_partition_by	否	分区策略，用来确保相同一类的数据写入到kafka同一分区 ● database：使用数据库名称做分区，保证同一个数据库的事件写入到kafka同一个分区中 ● table：使用表名称做分区，保证同一个表的事件写入到kafka同一个分区中	databa
ignore_producer_error	否	是否忽略生产者发送数据失败的错误 ● true：在日志中打印错误信息并跳过错误的数据，程序继续运行 ● false：在日志中打印错误信息并终止程序	true
metrics_slf4j_interval	否	在日志中输出上传kafka成功与失败数据的数量统计的时间间隔，单位为秒	60
kafka.bootstrap.servers	是	kafka代理节点地址，配置形式为HOST:PORT[,HOST:PORT]	-
kafka_topic	否	写入kafka的topic名称	maxwell
dead_letter_topic	否	当发送某条记录出错时，记录该条出错记录主键的kafka topic	-
kafka_version	否	Maxwell使用的kafka producer版本号，不能在config.properties中配置，需要在启动命令时用-- kafka_version xxx参数传入	-
kafka_partition_hash	否	划分kafka topic partition的算法，支持default或murmur3	default
kafka_key_format	否	Kafka record的key生成方式，支持array或Hash	Hash
ddl_kafka_topic	否	当output_ddl配置为true时，DDL操作写入的topic	{kafka_topic}

配置项	是否必填	说明	默认值
filter	否	过滤数据库或表。 • 若只想采集mydatabase的库，可以配置为exclude: *.*,include: mydatabase.* • 若只想采集mydatabase.mytable的表，可以配置为exclude: *.*,include: mydatabase.mytable • 若只想采集mydatabase库下的mytable，mydate_123, mydate_456表，可以配置为exclude: *.*,include: mydatabase.mytable, include: mydatabase./mydate_\\d*/	-

启动 Maxwell

步骤1 登录Maxwell所在的服务器。

步骤2 执行如下命令进入Maxwell安装目录。

cd /opt/maxwell-1.21.0/

说明

如果是初次使用Maxwell，建议将conf/config.properties中的log_level改为debug(调试级别)，以便观察启动之后是否能正常从MySQL获取数据并发送到kafka，当整个流程调试通过之后，再把log_level修改为info，然后先停止再启动Maxwell生效。

```
# log level [debug | info | warn | error]
log_level=debug
```

步骤3 执行如下命令启动Maxwell。

source /opt/client/bigdata_env

bin/Maxwell

**bin/maxwell --user='maxwell' --password='XXXXXX' --host='127.0.0.1' **

--producer=kafka --kafka.bootstrap.servers=kafkahost:9092 --kafka_topic=Maxwell

其中，user，password和host分别表示MySQL的用户名，密码和IP地址，这三个参数可以通过修改配置项配置也可以通过上述命令配置，kafkahost为流式集群的Core节点的IP地址。

显示类似如下信息，表示Maxwell启动成功。

```
Success to start Maxwell [78092].
```

----结束

验证 Maxwell

步骤1 登录Maxwell所在的服务器。

步骤2 查看日志。如果日志里面没有ERROR日志，且有打印如下日志，表示与MySQL连接正常。

```
BinlogConnectorLifecycleListener - Binlog connected.
```

步骤3 登录MySQL数据库，对测试数据进行更新/创建/删除等操作。操作语句可以参考如下示例。

```
-- 创建库
create database test;
-- 创建表
create table test.e (
  id int(10) not null primary key auto_increment,
  m double,
  c timestamp(6),
  comment varchar(255) charset 'latin1'
);
-- 增加记录
insert into test.e set m = 4.2341, c = now(3), comment = 'I am a creature of light.';
-- 更新记录
update test.e set m = 5.444, c = now(3) where id = 1;
-- 删除记录
delete from test.e where id = 1;
-- 修改表
alter table test.e add column torvalds bigint unsigned after m;
-- 删除表
drop table test.e;
-- 删除库
drop database test;
```

步骤4 观察Maxwell的日志输出，如果没有WARN/ERROR打印，则表示Maxwell安装配置正常。

若要确定数据是否成功上传，可设置config.properties中的log_level为debug，则数据上传成功时会立刻打印如下JSON格式数据，具体字段含义请参考**Maxwell生成的数据格式及常见字段含义**。

```
{"database":"test","table":"e","type":"insert","ts":1541150929,"xid":60556,"commit":true,"data":
{"id":1,"m":4.2341,"c":"2018-11-02 09:28:49.297000","comment":"I am a creature of light."}}
......
```

说明

当整个流程调试通过之后，可以把config.properties文件中的配置项log_level修改为info，减少日志打印量，并重启Maxwell。

```
# log level [debug | info | warn | error]
log_level=info
```

----结束

停止 Maxwell

步骤1 登录Maxwell所在的服务器。

步骤2 执行如下命令，获取Maxwell的进程标识（PID）。输出的第二个字段即为PID。

ps -ef | grep Maxwell | grep -v grep

步骤3 执行如下命令，强制停止Maxwell进程。

kill -9 PID

----结束

Maxwell 生成的数据格式及常见字段含义

Maxwell生成的数据格式为JSON，常见字段含义如下：

- type：操作类型，包含database-create，database-drop，table-create，table-drop，table-alter，insert，update，delete
- database：操作的数据库名称
- ts：操作时间，13位时间戳
- table：操作的表名
- data：数据增加/删除/修改之后的内容
- old：数据修改前的内容或者表修改前的结构定义
- sql：DDL操作的SQL语句
- def：表创建与表修改的结构定义
- xid：事物唯一ID
- commit：数据增加/删除/修改操作是否已提交

12.14.7 创建 Kafka 角色

操作场景

该任务指导系统管理员创建并设置Kafka的角色。

本章节内容适用于MRS 3.x及后续版本。

说明

安全模式支持创建Kafka角色，普通模式不支持创建Kafka角色。

如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Kafka的Ranger访问权限策略**。

前提条件

系统管理员已明确业务需求。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

步骤3 在“配置资源权限”中，选择“*待操作集群的名称* > Kafka”。

步骤4 根据业务需求选择权限，具体配置项，请参见**表12-260**

表 12-260 配置项说明

任务场景	角色授权操作
设置Kafka管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > Kafka > Kafka Manager权限”。 **说明** 设置此权限，拥有Topic的创建、删除等权限，但是不具备任何Topic的生产和消费权限。
设置用户对Topic的生产权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Kafka > Kafka Topic生产和消费权限”。 2. 在指定Topic的“权限”列，勾选“Kafka生产者权限”。
设置用户对Topic的消费权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Kafka > Kafka Topic生产和消费权限”。 2. 在指定Topic的“权限”列，勾选“Kafka消费者权限”。

步骤5 单击“确定”完成，返回“角色”。

----结束

12.14.8 Kafka 常用参数

本章节内容适用于MRS 3.x及后续版本。

参数入口

参数入口，请参考**修改集群服务配置参数**。

常用参数

表 12-261 参数说明

配置参数	说明	缺省值
log.dirs	Kafka数据存储目录列表，以逗号分隔多个目录。	%{@auto.detect.datapart.bk.log.logs}
KAFKA_HEAP_OPTS	Kafka启动Broker时使用的jvm选项。建议根据业务需要进行设置。	-Xmx6G -Xms6G
auto.create.topics.enable	是否自动创建Topic，若参数设置为false，发消息前需要通过命令创建Topic。	true

配置参数	说明	缺省值
default.replication.factor	自动创建Topic时的默认副本数。	2
monitor.preInitDelay	服务启动后，第一次健康检查的延迟时间。如果启动需要较长时间，可以通过调大参数，来完成启动。单位为毫秒。	600000

超时参数

表 12-262 Broker 相关超时参数

参数名称	参数说明	默认值	影响分析
controller.socket.timeout.ms	Controller连接Broker的超时时间。单位：毫秒。	30000	Controller连接Broker的超时时间，一般不需要调整。
group.max.session.timeout.ms	Consumer注册时允许的最大会话超时时间。单位：毫秒。	1800000	允许Consumer配置的session.timeout.ms的最大值（不包含此值）。
group.min.session.timeout.ms	Consumer注册时允许的最小会话超时时间。单位：毫秒。	6000	允许Consumer配置的session.timeout.ms的最小值（不包含此值）。
offsets.commit.timeout.ms	Offset提交请求的超时时间。单位：毫秒。	5000	Offset提交时被延迟处理的最大超时时间。
replica.socket.timeout.ms	副本数据同步请求的超时时间，配置值不得小于replica.fetch.wait.max.ms。单位：毫秒。	30000	同步线程在发送同步请求之前等待通道建立的最大超时时间，要求配置大于replica.fetch.wait.max.ms。
request.timeout.ms	设置客户端发送连接请求后，等待响应的超时时间。如果在超时时间内没有接收到响应，那么客户端重新发送，并在达到重试次数后返回请求失败。单位：毫秒。	30000	Broker节点上的Controller、Replica线程中传入networkclient连接的超时参数。

参数名称	参数说明	默认值	影响分析
transaction.max.timeout.ms	事务允许的最大超时。如果客户端的请求时间超过该值，则Broker将在InitProducerIdRequest中返回一个错误。这样可以防止客户端超时时间过长，而导致消费者无法接收topic。单位：毫秒。	900000	事务最大超时时间。
user.group.cache.timeout.sec	指定缓存中保存用户对应组信息的时间。单位：秒。	300	缓存中用户和组对应关系缓存时间，超过此时间用户信息才会再次通过id -Gn命令查询，在此期间，仅使用缓存中的用户和组对应关系。
zookeeper.connection.timeout.ms	连接ZooKeeper的超时时间。单位：毫秒。	45000	ZooKeeper连接超时时间，这个时间决定了zkclient中初次连接建立过程时允许消耗的时间，超过该时间，zkclient会主动断开。
zookeeper.session.timeout.ms	ZooKeeper会话超时时间。如果Broker在此时间内未向ZooKeeper上报心跳，则被认为失效。单位：毫秒。	45000	ZooKeeper会话超时时间。 作用一：这个时间结合传入的ZKURL中ZooKeeper的地址个数，ZooKeeper客户端以（sessionTimeout/传入ZooKeeper地址个数）为连接一个节点的超时时间，超过此时间未连接成功，则尝试连接下一个节点。 作用二：连接建立后，一个会话的超时时间，如ZooKeeper上注册的临时节点BrokerId，当Broker被停止，则该BrokerId，会经过一个sessionTimeout才会被ZooKeeper清理。

表 12-263 Producer 相关超时参数

配置名称	说明	默认值	影响分析
request.timeout.ms	指定发送消息请求的请求超时时间。	30000	请求超时时间，出现网络问题时，需调大此参数；配置过小，则容易出现Batch Expire异常。

表 12-264 Consumer 相关超时参数

配置名称	说明	默认值	影响分析
connections.max.idle.ms	空闲连接的保留时间。	600000	空闲连接的保留时间，连接空闲时间大于此时间，则会销毁该连接，有需要时重新创建连接。
request.timeout.ms	消费请求的超时时间。	30000	请求超时时间，请求超时会失败然后不断重试。

12.14.9 Kafka 安全使用说明

本章节内容适用于MRS 3.x及后续版本。

Kafka API 简单说明

- Producer API

  指**org.apache.kafka.clients.producer.KafkaProducer**中定义的接口，在使用“kafka-console-producer.sh”时，默认使用此API。

- Consumer API

  指**org.apache.kafka.clients.consumer.KafkaConsumer**中定义的接口，在使用“kafka-console-consumer.sh”时，默认会调用此API。

说明

MRS 3.x后，Kafka不支持旧Producer API和旧Consumer API。

Kafka 访问协议说明

Kafka当前支持四种协议类型的访问：PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL。

Kafka服务启动时，默认会启动PLAINTEXT和SASL_PLAINTEXT两种协议类型的访问监听。可通过设置Kafka服务配置“ssl.mode.enable”为“true”，来启动SSL和SASL_SSL两种协议类型的访问监听。下表是四中协议类型的简单说明：

协议类型	说明	默认端口
PLAINTEXT	支持无认证的明文访问	9092
SASL_PLAINTEXT	支持Kerberos认证的明文访问	21007
SSL	支持无认证的SSL加密访问	9093
SASL_SSL	支持Kerberos认证的SSL加密访问	21009

Topic 的 ACL 设置

Topic的权限信息，需要在Linux客户端上，使用“kafka-acls.sh”脚本进行查看和设置，具体可参考**管理Kafka用户权限**。

针对不同的 Topic 访问场景，Kafka 中 API 使用说明

- 场景一：访问设置了ACL的Topic

使用的API	用户属组	客户端参数	服务端参数	访问的端口
API	用户需满足以下条件之一即可： • 属于系统管理员组 • 属于kafkaadmin组 • 属于kafkasuperuser组 • 被授权的kafka组的用户	security.inter.broker.protocol=SASL_PLAINTEXT sasl.kerberos.service.name = kafka	-	sasl.port（默认21007）
		security.protocol=SASL_SSL sasl.kerberos.service.name = kafka	“ssl.mode.enable”配置为true	sasl-ssl.port（默认21009）

- 场景二：访问未设置ACL的Topic

使用的API	用户属组	客户端参数	服务端参数	访问的端口
API	用户需满足以下条件之一： • 属于系统管理员组 • 属于kafkaadmin组 • 属于kafkasuperuser组	security.protocol=SASL_PLAINTEXT sasl.kerberos.service.name = kafka	-	sasl.port（默认21007）
	用户属于kafka组		“allow.everyone.if.no.acl.found”配置为true **说明** 普通集群下不涉及服务端参数“allow.everyone.if.no.acl.found”的修改	sasl.port（默认21007）
	用户需满足以下条件之一： • 属于系统管理员组 • 属于kafkaadmin组 • kafkasuperuser组用户	security.protocol=SASL_SSL sasl.kerberos.service.name = kafka	“ssl.mode.enable”配置为“true”	sasl-ssl.port（默认21009）
	用户属于kafka组		1. “allow.everyone.if.no.acl.found”配置为“true” 2. “ssl.mode.enable”配置为“true”	sasl-ssl.port（默认21009）
	-	security.protocol=PLAINTEXT	“allow.everyone.if.no.acl.found”配置为“true”	port（默认9092）

使用的API	用户属组	客户端参数	服务端参数	访问的端口
	-	security.protocol=SSL	1. “allow.everyone.if.no.acl.found”配置为“true” 2. “ssl.mode.enable”配置为“true”	ssl.port（默认9063）

12.14.10 Kafka 业务规格说明

本章节内容适用于MRS 3.x及后续版本。

支持的 Topic 上限

支持Topic的个数，受限于进程整体打开的文件句柄数（现场环境一般主要是数据文件和索引文件占用比较多）。

1. 可通过**ulimit -n**命令查看进程最多打开的文件句柄数；
2. 执行**lsof -p** *<Kafka PID>*命令，查看当前单节点上Kafka进程打开的文件句柄（会继续增加）；
3. 权衡当前需要创建的Topic创建完成后，会不会达到文件句柄上限，每个Partition文件夹下会最多保存多大的数据，会产生多少个数据文件（*.log文件，默认配置为1GB，可通过修改**log.segment.bytes**来调整大小）和索引文件（*.index文件，默认配置为10MB，可通过修改**log.index.size.max.bytes**来调整大小），是否会影响Kafka正常运行。

Consumer 的并发量

在一个应用中，同一个Group的Consumer并发量建议与Topic的Partition个数保持一致，保证每个Consumer对应消费一个Partition上的数据。若Consumer的并发量多于Partition个数，那么多余的Consumer将消费不到数据。

Topic 和 Partition 的划分关系说明

- 假设集群中部署了K个Kafka节点，每个节点上配置的磁盘个数为N，每块磁盘大小为M，集群中共有n个Topic（T1,T2…Tn），并且其中第m个Topic的每秒输入数据总流量为X(Tm) MB/s，配置的副本数为R(Tm)，配置数据保存时间为Y(Tm)小时，那么整体必须满足：

$$M \times N \times K > \sum_{i=T1}^{Tn} (X(i)R(i)Y(i) \times 3600)$$

- 假设单个磁盘大小为M，该磁盘上有n个Partition（P0,P1……Pn），并且其中第m个Partition的每秒写入数据流量为Q(Pm) MB/s（计算方法：所属Topic的数据流量除以Partition数）、数据保存时间为T(Pm)小时，那么单个磁盘必须满足：

$$M > \sum_{i=P0}^{Pn} (Q(i)T(i) \times 3600)$$

- 根据吞吐量粗略计算，假设生产者可以达到的吞吐量为P，消费者可以达到的吞吐量为C，预期Kafka吞吐量为T，那么建议该Topic的Partition数目设置为Max(T/P，T/C)。

说明

- 在Kafka集群中，分区越多吞吐量越高，但是分区过多也存在潜在影响，例如文件句柄增加、不可用性增加（如：某个节点故障后，部分Partition重选Leader后时间窗口会比较大）及端到端时延增加等。
- 建议：单个Partition的磁盘占用最大不超过100GB；单节点上Partition数目不超过3000；整个集群的分区总数不超过10000。

12.14.11 使用 Kafka 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Kafka客户端。

本章节适用于MRS 3.x及后续版本。

前提条件

- 已安装客户端。例如安装目录为“/opt/client”。
- 各组件业务用户由系统管理员根据业务需要创建。“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）
- 在修改集群域名后，需要重新下载客户端，以保证客户端配置文件中kerberos.domain.name配置为正确的服务端域名。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤5 执行以下命令切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤6 执行以下命令使用客户端工具查看帮助并使用。

- **./kafka-console-consumer.sh**：Kafka消息读取工具
- **./kafka-console-producer.sh**：Kafka消息发布工具

- **./kafka-topics.sh**：Kafka Topic管理工具

----结束

12.14.12 配置 Kafka 高可用和高可靠参数

操作场景

Kafka消息传输保障机制，可以通过配置不同的参数来保障消息传输，进而满足不同的性能和可靠性要求。本章节介绍如何配置Kafka高可用和高可靠参数。

本章节内容适用于MRS 3.x及后续版本。

对系统的影响

- 配置高可用、高性能的影响：

  须知

  配置高可用、高性能模式后，数据可靠性会降低。在磁盘故障、节点故障等场景下存在数据丢失风险。

- 配置高可靠性的影响：
  - 性能降低：

    在生产数据时，配置了高可靠参数ack=-1之后，需要多个副本均写入成功之后才认为是写入成功。这样会导致单条消息时延增加，客户端处理能力下降。具体性能以现场实际测试数据为准。
  - 可用性降低：

    不允许不在ISR中的副本被选举为Leader。如果Leader下线时，其他副本均不在ISR列表中，那么该分区将保持不可用，直到Leader节点恢复。当分区的一个副本所在节点故障时，无法满足最小写入成功的副本数，那么将会导致业务写入失败。
- 参数配置项为服务级配置需要重启Kafka，建议在变更窗口做服务级配置修改。

参数描述

- 如果业务需要保证高可用和高性能。

  在服务端配置如**表12-265**中参数，参数配置入口请参考**修改集群服务配置参数**。

  表 12-265 服务端高可用性和高性能参数说明

参数	默认值	说明
unclean.leader.election.enable	true	是否允许不在ISR中的副本被选举为Leader，若设置为true，可能会造成数据丢失。

参数	默认值	说明
auto.leader.rebalance.enable	true	是否使用Leader自动均衡功能。 如果设为true，Controller会周期性的为所有节点的每个分区均衡Leader，将Leader分配给更优先的副本。
min.insync.replicas	1	当Producer设置acks为-1时，指定需要写入成功的副本的最小数目。

在客户端配置文件producer.properties中配置如表12-266中参数，producer.properties存放路径为：/opt/client/Kafka/kafka/config/producer.properties，其中/opt/client为Kafka客户端安装目录。

表 12-266 客户端高可用性和高性能参数说明

参数	默认值	说明
acks	1	需要Leader确认消息是否已经接收并认为已经处理完成。该参数会影响消息的可靠性和性能。 ● acks=0 ： Producer将不会等待服务端任何响应。消息将会被认为成功。 ● acks=1 ： 当副本所在Leader确认数据已写入，但是其不会等待所有的副本完全写入即返回响应。在这种情况下，如果Leader确认后但是副本未同步完成时Leader异常，那么数据就会丢失。 ● acks=-1 ： 意味着等待所有的同步副本确认后才认为成功，配合“min.insync.replicas”可以确保多副本写入成功，只要有一个副本保持活跃状态，记录将不会丢失。

- 如果业务需要保证数据高可靠性。

在服务端配置如表12-267参数，参数配置入口请参考修改集群服务配置参数。

表 12-267 服务端高可靠性参数说明

参数	建议值	说明
unclean.leader.election.enable	false	不允许不在ISR中的副本被选举为Leader。
min.insync.replicas	2	当Producer设置acks为-1时，指定需要写入成功的副本的最小数目。 需要满足min.insync.replicas <= replication.factor。

在客户端配置文件producer.properties中配置如表12-268中参数，producer.properties存放路径为：/opt/client/Kafka/kafka/config/producer.properties，其中/opt/client为Kafka客户端安装目录。

表 12-268 客户端高可靠性参数说明

参数	建议值	说明
acks	-1	Producer需要Leader确认消息是否已经接收并认为已经处理完成。 acks=-1需要等待在ISR列表的副本都确认接收到消息并处理完成才表示消息成功。配合“min.insync.replicas”可以确保多副本写入成功，只要有一个副本保持活跃状态，记录将不会丢失，此参数配置为-1时，会降低生产性能，请权衡后配置。

配置建议

请根据以下业务场景对可靠性和性能要求进行评估，采用合理参数配置。

- 对于价值数据，这两种场景下建议Kafka数据目录磁盘配置raid1或者raid5，从而提高单个磁盘故障情况下数据可靠性。
- 参数配置项均为Topic级别可修改的参数，默认采用服务级配置。

  可针对不同Topic可靠性要求对Topic进行单独配置。以root用户登录Kafka客户端节点，在客户端安装目录下配置Topic名称为test的可靠性参数命令：

  cd Kafka/kafka/bin

kafka-topics.sh --zookeeper 192.168.1.205:2181/kafka --alter --topic test --config unclean.leader.election.enable=false --config min.insync.replicas=2

其中192.168.1.205为ZooKeeper业务IP地址。

- 参数配置项为服务级配置需要重启Kafka，建议在变更窗口做服务级配置修改。

12.14.13 更改 Broker 的存储目录

操作场景

本章节内容适用于MRS 3.x及后续版本。

增加Broker的存储目录时，系统管理员需要在FusionInsight Manager中修改Broker的存储目录，以保证Kafka正常工作，新创建的主题分区将在分区最少的目录中生成。适用于以下场景：

说明

由于Kafka不感知磁盘容量，建议各Broker实例配置的磁盘个数和容量保持一致。

- 更改Broker角色的存储目录，所有Broker实例的存储目录将同步修改。
- 更改Broker单个实例的存储目录，只对单个实例生效，其他节点Broker实例存储目录不变。

对系统的影响

- 更改Broker角色的存储目录需要重新启动服务，服务重启时无法访问。
- 更改Broker单个实例的存储目录需要重新启动实例，该节点Broker实例重启时无法提供服务。
- 服务参数配置如果使用旧的存储目录，需要更新为新目录。

前提条件

- 在各个数据节点准备并安装好新磁盘，并格式化磁盘。
- 已安装好Kafka客户端。
- 更改Broker单个实例的存储目录时，保持活动的Broker实例数必须大于创建主题时指定的备份数。

操作步骤

更改Kafka角色的存储目录

步骤1 以**root**用户登录到安装Kafka服务的各个数据节点中，执行如下操作。

1. 创建目标目录。

   例如目标目录为“${BIGDATA_DATA_HOME}/kafka/data2”：

   执行**mkdir ${BIGDATA_DATA_HOME}/kafka/data2**。

2. 挂载目录到新磁盘。例如挂载“${BIGDATA_DATA_HOME}/kafka/data2”到新磁盘。
3. 修改新目录的权限。

   例如新目录路径为“${BIGDATA_DATA_HOME}/kafka/data2”：

执行**chmod 700 ${BIGDATA_DATA_HOME}/kafka/data2 -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/kafka/data2 -R**。

步骤2 MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > 服务 > Kafka > 配置”。

步骤3 添加新目录到“log.dirs”的默认值后面。

在搜索框中输入“log.dirs”进行搜索，将新目录添加到配置项“log.dirs”的默认值后面，多个目录使用逗号分隔。例如“

${BIGDATA_DATA_HOME}/kafka/data1/kafka-logs,${BIGDATA_DATA_HOME}/kafka/data2/kafka-logs”。

步骤4 单击“保存”，并单击“确定”。界面提示“操作成功”，单击“完成”。

步骤5 选择“集群 > 服务 > Kafka ，右上角选择“更多 > 重启服务”，重启Kafka服务。

更改Kafka单个实例的存储目录

步骤6 以**root**用户登录到Broker节点，执行如下操作。

1. 创建目标目录。

   例如目标目录为“${BIGDATA_DATA_HOME}/kafka/data2”：

   执行**mkdir ${BIGDATA_DATA_HOME}/kafka/data2**。

2. 挂载目录到新磁盘。例如挂载“${BIGDATA_DATA_HOME}/kafka/data2”到新磁盘。
3. 修改新目录的权限。

   例如新目录路径为“${BIGDATA_DATA_HOME}/kafka/data2”：

   执行**chmod 700 ${BIGDATA_DATA_HOME}/kafka/data2 -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/kafka/data2 -R**。

步骤7 MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > 服务 > Kafka > 实例”。

步骤8 单击指定的Broker实例并切换到“实例配置”。

在搜索框中输入“log.dirs”进行搜索，将新目录添加到配置项“log.dirs”的默认值后面，多个目录使用逗号分隔。例如“${BIGDATA_DATA_HOME}/kafka/data1/kafka-logs,${BIGDATA_DATA_HOME}/kafka/data2/kafka-logs”。

步骤9 单击“保存”，并单击“确定”，界面提示“操作成功”，单击“完成”。

步骤10 在Broker实例页面选择“更多 > 重启实例”，重启Broker实例。

----结束

12.14.14 查看 Consumer Group 消费情况

操作场景

该任务指导系统管理员根据业务需求，在客户端中查看当前消费情况。

本章节内容适用于MRS 3.x及后续版本。

前提条件

- 系统管理员已明确业务需求，并准备一个系统用户。
- 已安装Kafka客户端。

操作步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。

cd /opt/kafkaclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤5 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤6 使用**kafka-consumer-groups.sh**查看当前消费情况。

- 查看Offset保存在Kafka上的Consumer Group列表：

  ./kafka-consumer-groups.sh --list --bootstrap-server <*Broker的任意一个节点的业务IP:21007*> **--command-config ../config/consumer.properties**

  eg:./kafka-consumer-groups.sh --bootstrap-server 192.168.1.1:21007 --list --command-config ../config/consumer.properties

- 查看Offset保存在Kafka上的Consumer Group消费情况：

  ./kafka-consumer-groups.sh --describe --bootstrap-server <*Broker的任意一个节点的业务IP:21007*> *--group 消费组名称* **--command-config ../config/consumer.properties**

  eg:./kafka-consumer-groups.sh --describe --bootstrap-server 192.168.1.1:21007 --group example-group --command-config ../config/consumer.properties

须知

1. 确保当前consumer在线消费。
2. 确保配置文件consumer.properties中的group.id与命令中--group的参数均配置为待查询的group。
3. Kafka集群IP端口号安全模式下是21007，普通模式下是9092。

----结束

12.14.15 Kafka 均衡工具使用说明

操作场景

该任务指导管理员根据业务需求，在客户端中执行Kafka均衡工具来均衡Kafka集群的负载，一般用于节点的退服、入服以及负载均衡的场景。

本章节内容适用于MRS 3.x及后续版本。3.x之前版本请参考**Kafka扩容节点后数据均衡**

前提条件

- 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组，普通模式不需要）。
- 已安装Kafka客户端。

操作步骤

步骤1 以客户端安装用户，登录已安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。

cd /opt/kafkaclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证（普通模式跳过此步骤）。

kinit *组件业务用户*

步骤5 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka

步骤6 使用“kafka-balancer.sh”进行用户集群均衡，常用命令如下：

- 使用--run命令执行集群均衡：

  ./bin/kafka-balancer.sh --run --zookeeper *<ZooKeeper的任意一个节点的业务IP:zkPort/**kafka**>* **--bootstrap-server** *<Kafka集群IP: port>* **--throttle 10000000 --consumer-config config/consumer.properties --enable-az-aware --show-details**

  该命令包含均衡方案的生成和执行两部分，其中--show-details为可选参数，表示是否打印方案明细，--throttle表示均衡方案执行时的带宽限制，单位:bytes/sec，--enable-az-aware为可选参数，表明生成均衡方案时，开启跨AZ特性，使用此参数时，请务必保证集群已开启跨AZ特性。

- 使用--run命令执行节点退服：

  ./bin/kafka-balancer.sh --run --zookeeper *<ZooKeeper的任意一个节点的业务IP:zkPort/**kafka**>* **--bootstrap-server** *<Kafka集群IP: port>* **--throttle 10000000 --consumer-config config/consumer.properties --remove-brokers** *<BrokerId列表>* **--enable-az-aware --force**

  其中--remove-brokers表示要删除的BrokerId列表，多个间用逗号分隔，--force参数为可选参数，表示忽略磁盘使用率告警，强制生成迁移方案， -enable-az-aware为可选参数，表明生成均衡方案时，开启跨AZ特性，使用此参数时，请务必保证集群已开启跨AZ特性。

- 查看执行状态：

  ./bin/kafka-balancer.sh --status --zookeeper *<ZooKeeper的任意一个节点的业务IP:zkPort/**kafka**>*

- 生成均衡方案：

  ./bin/kafka-balancer.sh --generate --zookeeper *<ZooKeeper的任意一个节点的业务IP:zkPort/**kafka**>* **--bootstrap-server** *<Kafka集群IP:port>* **--consumer-config config/consumer.properties --enable-az-aware**

  该命令仅根据集群当前状态生成迁移方案，并打印到控制台, 其中--enable-az-aware为可选参数，表明生成迁移方案时，开启跨AZ特性，使用此参数时，请务必保证集群已开启跨AZ特性。

- 清理中间状态

  ./bin/kafka-balancer.sh --clean --zookeeper *<ZooKeeper的任意一个节点的业务IP:zkPort/**kafka**>*

  一般在迁移没有正常执行完成时用来清理ZooKeeper上的中间状态信息。

须知

Kafka集群IP端口号安全模式下是21007，普通模式下是9092。

----结束

异常情况处理

在使用Kafka均衡工具进行Partition迁移的过程中，如果出现集群中Broker故障导致均衡工具的执行进度阻塞，这时需要人工介入来恢复，分为以下几种场景：

- 存在Broker因为磁盘占有率达到100%导致Broker故障的情况。
  a. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例”，将运行状态为“正在恢复”的Broker实例停止并记录实例所在节点的管理IP地址以及对应的“broker.id”，该值可通过单击角色名称，在“实例配置”页面中选择“全部配置”，搜索“broker.id”参数获取。
  b. 以**root**用户登录记录的管理IP地址，并执行**df -lh**命令，查看磁盘占用率为100%的挂载目录，例如“${BIGDATA_DATA_HOME}/kafka/data1”。
  c. 进入该目录，执行**du -sh ***命令，查看该目录下各文件夹的大小。查看是否存在除“kafka-logs”目录外的其他文件，并判断是否可以删除或者迁移。
     - 是，删除或者迁移相关数据，然后执行**8**。
     - 否，执行**4**。
  d. 进入“kafka-logs”目录，执行**du -sh ***命令，选择一个待移动的Partition文件夹，其名称命名规则为“Topic名称-Partition标识”，记录Topic及Partition。
  e. 修改“kafka-logs”目录下的“recovery-point-offset-checkpoint”和“replication-offset-checkpoint”文件（两个文件做同样的修改）。
     i. 减少文件中第二行的数字（若移出多个目录，则减少的数字为移出的目录个数）。

ii. 删除待移出的Partition所在的行（行结构为“Topic名称 Partition标识 Offset”，删除前先将该行数据保存，后续此内容还要添加到目的目录下的同名文件中）。

f. 修改目的数据目录下（例如：“${BIGDATA_DATA_HOME}/kafka/data2/kafka-logs”）的“recovery-point-offset-checkpoint”和“replication-offset-checkpoint”文件（两个文件做同样的修改）。

- 增加文件中第二行的数字（若移入多个Partition目录，则增加的数字为移入的Partition目录个数）。
- 添加待移入的Partition行到文件末尾（行结构为“Topic名称 Partition标识 Offset”，直接复制**5**中保存的行数据即可）。

g. 移动数据，将待移动的Partition文件夹移动到目的目录下，移动完成后执行**chown omm:wheel -R** *Partition目录*命令修改Partition目录属组。

h. 登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka > 实例”，启动停止的Broker实例。

i. 等待5至10分钟后查看Broker实例的运行状态是否为“良好”。

- 是，修复完成后按照“ALM-38001 Kafka磁盘容量不足”告警指导彻底解决磁盘容量不足问题。
- 否，联系运维人员。

按照上述步骤将故障Broker进行恢复后，阻塞的均衡任务会继续执行，可使用--status命令来查看任务的执行进度。

- 存在由其他原因导致的Broker故障，且问题场景单一明确，短时间内可以恢复Broker的情况。

  a. 根据问题根因指定恢复方案，恢复故障Broker。

  b. 故障Broker恢复后，阻塞的均衡任务会继续执行，可使用--status命令来查看任务的执行进度。

- 存在由其他原因导致的Broker故障，且问题场景复杂，短时间内无法恢复Broker的情况。

  a. 执行**kinit** *Kafka管理员用户*。（普通模式跳过此步骤）

  b. 使用**zkCli.sh -server** *<ZooKeeper集群业务IP:zkPort/**kafka**>*登录ZooKeeper Shell。

  c. 执行**addauth krbgroup**。（普通模式跳过此步骤）

  d. 删除“/admin/reassign_partitions”目录和“/controller”目录。

  e. 通过以上步骤强行终止迁移，待集群恢复后使用**kafka-reassign-partitions.sh**命令手动将中间过程中导致的多余的副本删除。

12.14.16 Kafka 扩容节点后数据均衡

操作场景

该任务指导管理员在Kafka扩容节点后，在客户端中执行Kafka均衡工具来均衡Kafka集群的负载。

本章节内容适用于MRS 3.x之前版本。3.x及之后版本请参考**Kafka均衡工具使用说明**。

前提条件

- 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组，普通模式不需要）。
- 已安装Kafka客户端，客户端安装目录如“/opt/kafkaclient”。
- 本示例需创建两个Topic，可参考步骤7，分别命名为“test_2”和“test_3”，并创建“move-kafka-topic.json”文件，创建路径如“/opt/kafkaclient/Kafka/kafka”，Topic格式内容如下：

```
{
"topics":
[{"topic":"test_2"},{"topic":"test_3"}],
"version":1
}
```

操作步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录。

cd /opt/kafkaclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤5 执行以下命令进入Kafka客户端的bin目录。

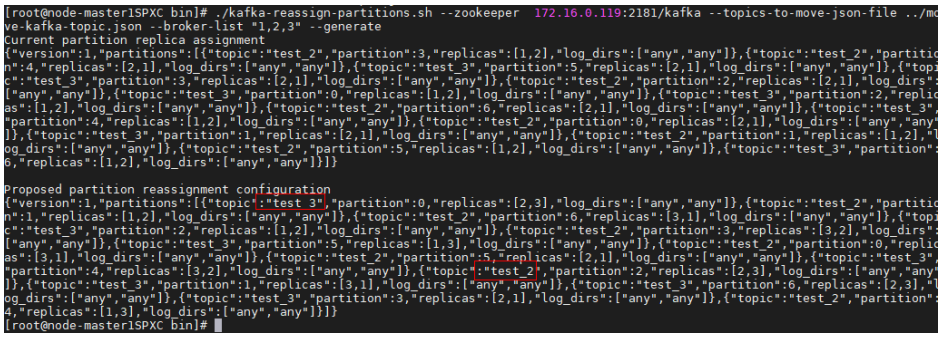
cd Kafka/kafka/bin

步骤6 执行以下命令生成执行计划。

./kafka-reassign-partitions.sh --zookeeper *172.16.0.119*:**2181/kafka --topics-to-move-json-file ../move-kafka-topic.json --broker-list "***1,2,3***" --generate**

说明

- 172.16.0.119：ZooKeeper实例的业务IP。
- --broker-list "1,2,3"：参数中的“1,2,3”为扩容后的所有broker_id。



步骤7 执行**vim ../reassignment.json**创建“reassignment.json”文件并保存，保存路径为“/opt/kafkaclient/Kafka/kafka”。

拷贝**步骤6**中生成的“Proposed partition reassignment configuration”下的内容至“reassignment.json”文件，如下所示：

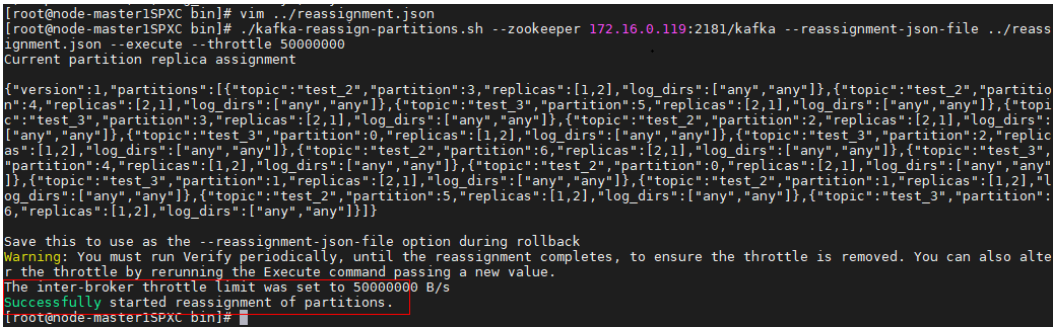
```
{"version":1,"partitions":[{"topic":"test","partition":4,"replicas":[1,2],"log_dirs":["any","any"]},
{"topic":"test","partition":1,"replicas":[1,3],"log_dirs":["any","any"]},{"topic":"test","partition":3,"replicas":
[3,1],"log_dirs":["any","any"]},{"topic":"test","partition":0,"replicas":[3,2],"log_dirs":["any","any"]},
{"topic":"test","partition":2,"replicas":[2,1],"log_dirs":["any","any"]}]}
```

步骤8 执行以下命令进行分区重分布。

./kafka-reassign-partitions.sh --zookeeper *172.16.0.119*:**2181/kafka --reassignment-json-file ../reassignment.json --execute --throttle** *50000000*

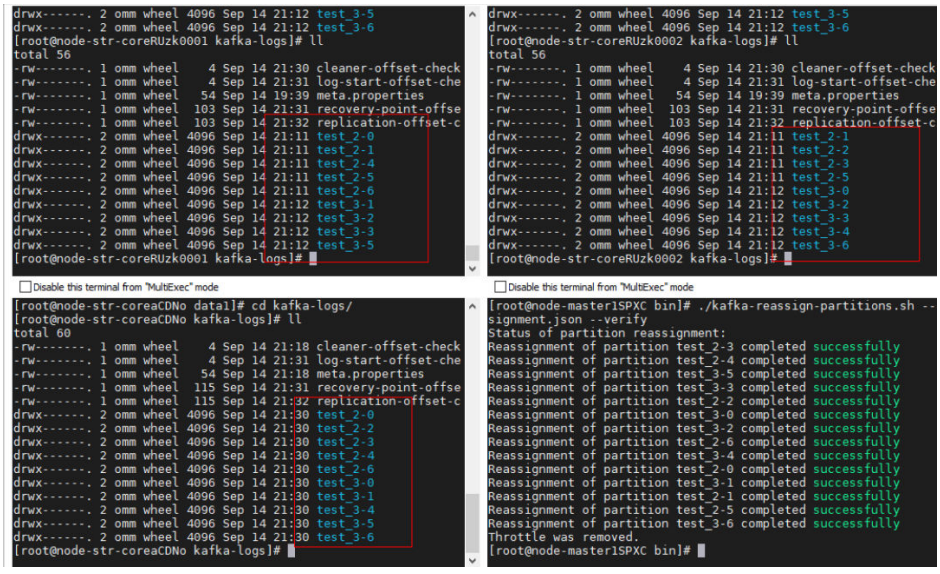
说明

--throttle 50000000：限制网络带宽为50MB。带宽可根据数据量大小及客户对均衡时间的要求进行调整，5TB数据量，使用50MB带宽，均衡时长约8小时。



步骤9 执行以下命令查看迁移状态。

./kafka-reassign-partitions.sh --zookeeper *172.16.0.119*:**2181/kafka --reassignment-json-file ../reassignment.json --verify**



----结束

12.14.17 Kafka Token 认证机制工具使用说明

操作场景

使用Token认证机制时对Token的操作。

本章节内容适用于MRS 3.x及后续版本的安全集群。

前提条件

- 系统管理员已明确业务需求，并准备一个系统用户。
- 已安装Kafka客户端。

操作步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。

cd /opt/kafkaclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证。

kinit *组件业务用户*

步骤5 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤6 使用**kafka-delegation-tokens.sh**对Token进行操作

- 为用户生成Token

  ./kafka-delegation-tokens.sh --create --bootstrap-server <*IP1:PORT, IP2:PORT,...*> **--max-life-time-period** <*Long: max life period in milliseconds*> **--command-config** <*config file*> **--renewer-principal User:**<*user name*>

  例如： **./kafka-delegation-tokens.sh --create --bootstrap-server 192.168.1.1:21007,192.168.1.2:21007,192.168.1.3:21007 --command-config ../config/producer.properties --max-life-time-period -1 --renewer-principal User:username**

- 列出归属在特定用户下的所有Token信息

  ./kafka-delegation-tokens.sh --describe --bootstrap-server <*IP1:PORT, IP2:PORT,...*> **--command-config** <*config file*> **--owner-principal User:**<*user name*>

  例如： **./kafka-delegation-tokens.sh --describe --bootstrap-server 192.168.1.1:21007,192.168.1.2:21007,192.168.1.3:21007 --command-config ../config/producer.properties --owner-principal User:username**

- Token有效期刷新

  ./kafka-delegation-tokens.sh --renew --bootstrap-server <*IP1:PORT, IP2:PORT,...*> **--renew-time-period** <*Long: renew time period in milliseconds*> **--command-config** <*config file*> **--hmac** <*String: HMAC of the delegation token*>

例如： **./kafka-delegation-tokens.sh --renew --bootstrap-server 192.168.1.1:21007,192.168.1.2:21007,192.168.1.3:21007 --renew-time-period -1 --command-config ../config/producer.properties --hmac ABCDEFG**

- 销毁Token

  ./kafka-delegation-tokens.sh --expire --bootstrap-server <*IP1:PORT, IP2:PORT,...*> **--expiry-time-period** <*Long: expiry time period in milliseconds*> **--command-config** <*config file*> **--hmac** <*String: HMAC of the delegation token*>

  例如： **./kafka-delegation-tokens.sh --expire --bootstrap-server 192.168.1.1:21007,192.168.1.2:21007,192.168.1.3:21007 --expiry-time-period -1 --command-config ../config/producer.properties --hmac ABCDEFG**

----结束

12.14.18 Kafka 日志介绍

本章节内容适用于MRS 3.x及后续版本。

日志描述

日志路径：Kafka相关日志的默认存储路径为“/var/log/Bigdata/kafka”，审计日志的默认存储路径为“/var/log/Bigdata/audit/kafka”。

- Broker：“/var/log/Bigdata/kafka/broker”（运行日志）

日志归档规则：Kafka的日志启动了自动压缩归档功能，默认情况下，当日志大小超过30MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。默认最多保留最近的20个压缩文件，压缩文件保留个数和压缩文件阈值可以配置。

表 12-269 Broker 日志列表

日志类型	日志文件名	描述
运行日志	server.log	Broker进程的server运行日志。
	controller.log	Broker进程的controller运行日志。
	kafka-request.log	Broker进程的request运行日志。
	log-cleaner.log	Broker进程的cleaner运行日志。
	state-change.log	Broker进程的state-change运行日志。
	kafkaServer-<SSH_USER>-<DATE>-<PID>-gc.log	Broker进程的GC日志。
	postinstall.log	Broker安装后的工作日志。

日志类型	日志文件名	描述
	prestart.log	Broker启动前的工作日志。
	checkService.log	Broker启动是否成功的检查日志。
	start.log	Broker进程启动日志。
	stop.log	Broker进程停止日志。
	checkavailable.log	Kafka服务健康状态检查日志。
	checkInstanceHealth.log	Broker实例健康状态检测日志。
	kafka-authorizer.log	Broker鉴权日志。
	kafka-root.log	Broker基础日志。
	cleanup.log	Broker卸载的清理日志。
	metadata-backup-recovery.log	Broker备份恢复日志。
	ranger-kafka-plugin-enable.log	Broker启动Ranger插件日志。
	server.out	Broker jvm日志。
	audit.log	Ranger鉴权插件鉴权日志。此日志统一归档在“/var/log/Bigdata/audit/kafka”目录下。

日志级别

Kafka提供了如表12-270所示的日志级别。

运行日志的级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-270 日志级别

级别	描述
ERROR	ERROR表示系统运行的错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 请参考**修改集群服务配置参数**，进入Kafka的“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

----结束

日志格式

Kafka的日志格式如下所示

表 12-271 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件调用类全名>(<日志打印文件>:<行号>)	2015-08-08 11:09:53,483 | INFO | [main] | Loading logs. | kafka.log.LogManager (Logging.scala:68)
	<yyyy-MM-dd HH:mm:ss><HostName><组件名><logLevel><Message>	2015-08-08 11:09:51 10-165-0-83 Kafka INFO Running kafka-start.sh.

12.14.19 性能调优

12.14.19.1 Kafka 性能调优

操作场景

通过调整Kafka服务端参数，可以提升特定业务场景下Kafka的处理能力。

参数调优

修改服务配置参数，请参考**修改集群服务配置参数**。调优参数请参考**表12-272**。

表 12-272 调优参数

配置参数	缺省值	调优场景
num.recovery.threads.per.data.dir	10	在Kafka启动过程中，数据量较大情况下，可调大此参数，可以提升启动速度。

配置参数	缺省值	调优场景
background.threads	10	Broker后台任务处理的线程数目。数据量较大的情况下，可适当调大此参数，以提升Broker处理能力。
num.replica.fetchers	1	副本向Leader请求同步数据的线程数，增大这个数值会增加副本的I/O并发度。
num.io.threads	8	Broker用来处理磁盘I/O的线程数目，这个线程数目建议至少等于硬盘的个数。
KAFKA_HEAP_OPTS	-Xmx6G -Xms6G	Kafka JVM堆内存设置。当Broker上数据量较大时，应适当调整堆内存大小。

12.14.20 Kafka 特性说明

Kafka Idempotent 特性

特性说明：Kafka从0.11.0.0版本引入了创建幂等性Producer的功能，开启此特性后，Producer自动升级成幂等性Producer，当Producer发送了相同字段值的消息后，Broker会自动感知消息是否重复，继而避免数据重复。需要注意的是，这个特性只能保证单分区上的幂等性，即一个幂等性Producer能够保证某个主题的一个分区内不出现重复消息；只能实现单会话上的幂等性，这里的会话指的是Producer进程的一次运行，即重启Producer进程后，幂等性不保证。

开启方法：

1. 二次开发代码中添加 “props.put(“enable.idempotence”，true)”。
2. 客户端配置文件中添加 “enable.idempotence = true”。

Kafka Transaction 特性

特性说明：Kafka在0.11版本中，引入了事务特性，Kafka事务特性指的是一系列的生产者生产消息和消费者提交偏移量的操作在一个事务中，或者说是一个原子操作，生产消息和提交偏移量同时成功或者失败，此特性提供的是read committed隔离级别的事务，保证多条消息原子性的写入到目标分区，同时也能保证Consumer只能看到成功提交的事务消息。Kafka中的事务特性主要用于以下两种场景：

1. 生产者发送多条数据可以封装在一个事务中，形成一个原子操作。多条消息要么都发送成功，要么都发送失败。
2. read-process-write模式：将消息消费和生产封装在一个事务中，形成一个原子操作。在一个流式处理的应用中，常常一个服务需要从上游接收消息，然后经过处理后送达到下游，这就对应着消息的消费和生产。

二次开发代码样例如下：

```
// 初始化配置,开启事务特性
Properties props = new Properties();
props.put("enable.idempotence", true);
```

```
props.put("transactional.id", "transaction1");
...

KafkaProducer producer = new KafkaProducer<String, String>(props);

// init 事务
producer.initTransactions();
try {
    // 开启事务
    producer.beginTransaction();
    producer.send(record1);
    producer.send(record2);
    // 结束事务
    producer.commitTransaction();
} catch (KafkaException e) {
    // 事务 abort
    producer.abortTransaction();
}
```

就近消费特性

特性说明：Kafka 2.4.0之前版本，客户端的生产、消费都是面向各个partition的leader副本，follower副本仅用来做数据冗余，不对外提供服务，常会导致leader副本压力较大，且在跨机房、机架的消费场景下，常会导致大量的机房、机架间的数据传输；Kafka 2.4.0及之后版本，Kafka内核支持从follower副本消费数据，在跨机房、机架的场景中，会大大降低数据传输量，减轻网络带宽压力。社区开放了ReplicaSelector接口来支持此特性，MRS Kafka中默认提供两种实现此接口的方式。

1. RackAwareReplicaSelector：优先从相同机架的副本进行消费（机架内就近消费特性）。
2. AzAwareReplicaSelector：优先从相同AZ内的节点上的副本进行消费（AZ内就近消费特性）。

以RackAwareReplicaSelector为例，描述实现就近消费副本的选取：

```
public class RackAwareReplicaSelector implements ReplicaSelector {

    @Override
    public Optional<ReplicaView> select(TopicPartition topicPartition,
                                        ClientMetadata clientMetadata,
                                        PartitionView partitionView) {
        if (clientMetadata.rackId() != null && !clientMetadata.rackId().isEmpty()) {
            Set<ReplicaView> sameRackReplicas = partitionView.replicas().stream()
                    // 过滤与客户端处于相同Rack的副本
                    .filter(replicaInfo -> clientMetadata.rackId().equals(replicaInfo.endpoint().rack()))
                    .collect(Collectors.toSet());
            if (sameRackReplicas.isEmpty()) {
                // 如果没有副本与客户端处于相同Rack，则返回leader副本
                return Optional.of(partitionView.leader());
            } else {
                // 到这里说明存在与客户端位于同一Rack的副本
                if (sameRackReplicas.contains(partitionView.leader())) {
                    // 如果客户端和leader在同一个机架，则优先返回leader副本
                    return Optional.of(partitionView.leader());
                } else {
                    // 否则，返回和leader同步最新的副本
                    return sameRackReplicas.stream().max(ReplicaView.comparator());
                }
            }
        } else {
            // 如果客户端请求中不包含机架信息，则默认返回leader副本
            return Optional.of(partitionView.leader());
        }
    }
}
```

开启方法：

1. 服务端：根据不同特性更新“replica.selector.class”配置项：
   - 开启“机架内就近消费特性”，配置为“org.apache.kafka.common.replica.RackAwareReplicaSelector”。
   - 开启“AZ内就近消费特性”，配置为“org.apache.kafka.common.replica.AzAwareReplicaSelector”。
2. 客户端：在“{*客户端安装目录*}/Kafka/kafka/config”目录中的“consumer.properties”消费配置文件里添加“client.rack”配置项：
   - 若服务端开启“机架内就近消费特性”，添加客户端所处的机架信息，如client.rack = /default0/rack1。
   - 若服务端开启“AZ内就近消费特性”，添加客户端所处的机架信息，如client.rack = /AZ1/rack1。

Ranger 统一鉴权特性

特性说明：在Kafka 2.4.0之前版本，Kafka组件仅支持社区自带的SimpleAclAuthorizer鉴权插件，Kafka 2.4.0及之后版本，MRS Kafka同时支持Ranger鉴权插件和社区自带鉴权插件。默认使用Ranger鉴权，基于Ranger鉴权插件，可进行细粒度的Kafka Acl管理。

说明

服务端使用Ranger鉴权插件时，若“allow.everyone.if.no.acl.found”配置为“true”，使用非安全端口访问时，所有行为将直接放行。建议使用Ranger鉴权插件的安全集群，不要开启“allow.everyone.if.no.acl.found”。

12.14.21 Kafka 节点内数据迁移

操作场景

该任务指导管理员根据业务需求，通过Kafka客户端命令，在不停止服务的情况下，进行节点内磁盘间的分区数据迁移。

前提条件

- 系统管理员已明确业务需求，并准备一个Kafka用户（属于kafkaadmin组，普通模式不需要）。
- 已安装Kafka客户端。
- Kafka实例状态和磁盘状态均正常。
- 根据待迁移分区当前的磁盘空间占用情况，评估迁移后，不会导致新迁移后的磁盘空间不足。

操作步骤

步骤1 以客户端安装用户，登录已安装Kafka客户端的节点。

步骤2 执行以下命令，切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。

cd /opt/kafkaclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证（普通模式跳过此步骤）。

kinit *组件业务用户*

步骤5 执行以下命令，切换到Kafka客户端目录。

cd Kafka/kafka/bin

步骤6 执行以下命令，查看待迁移的Partition对应的Topic的详细信息。

安全模式：

./kafka-topics.sh --describe --bootstrap-server *Kafka集群IP:21007* **--command-config ../config/client.properties --topic** *主题名称*

普通模式：

./kafka-topics.sh --describe --bootstrap-server *Kafka集群IP:21005* **--command-config ../config/client.properties --topic** *主题名称*

```
Topic:testws    PartitionCount:24       ReplicationFactor:2     Configs:
        Topic: testws   Partition: 0    Leader: 4       Replicas: 4,3   Isr: 4,3
        Topic: testws   Partition: 1    Leader: 5       Replicas: 5,4   Isr: 5,4
        Topic: testws   Partition: 2    Leader: 6       Replicas: 6,5   Isr: 6,5
        Topic: testws   Partition: 3    Leader: 3       Replicas: 3,6   Isr: 3,6
        Topic: testws   Partition: 4    Leader: 4       Replicas: 4,5   Isr: 4,5
        Topic: testws   Partition: 5    Leader: 5       Replicas: 5,4   Isr: 5,4
        Topic: testws   Partition: 6    Leader: 6       Replicas: 6,3   Isr: 6,3
        Topic: testws   Partition: 7    Leader: 3       Replicas: 3,4   Isr: 3,4
        Topic: testws   Partition: 8    Leader: 4       Replicas: 4,6   Isr: 4,6
        Topic: testws   Partition: 9    Leader: 5       Replicas: 5,3   Isr: 5,3
        Topic: testws   Partition: 10   Leader: 6       Replicas: 6,4   Isr: 6,4
        Topic: testws   Partition: 11   Leader: 3       Replicas: 3,5   Isr: 3,5
        Topic: testws   Partition: 12   Leader: 4       Replicas: 4,3   Isr: 4,3
        Topic: testws   Partition: 13   Leader: 5       Replicas: 5,4   Isr: 5,4
        Topic: testws   Partition: 14   Leader: 6       Replicas: 6,5   Isr: 6,5
        Topic: testws   Partition: 15   Leader: 3       Replicas: 3,6   Isr: 3,6
        Topic: testws   Partition: 16   Leader: 4       Replicas: 4,5   Isr: 4,5
        Topic: testws   Partition: 17   Leader: 5       Replicas: 5,6   Isr: 5,6
        Topic: testws   Partition: 18   Leader: 6       Replicas: 6,3   Isr: 6,3
        Topic: testws   Partition: 19   Leader: 3       Replicas: 3,4   Isr: 3,4
        Topic: testws   Partition: 20   Leader: 4       Replicas: 4,6   Isr: 4,6
        Topic: testws   Partition: 21   Leader: 5       Replicas: 5,3   Isr: 5,3
        Topic: testws   Partition: 22   Leader: 6       Replicas: 6,4   Isr: 6,4
```

步骤7 执行以下命令，查询Broker_ID和IP对应关系。

./kafka-broker-info.sh --zookeeper *ZooKeeper的quorumpeer实例业务IP.ZooKeeper客户端端口号***/kafka**

```
Broker_ID     IP_Address
---------------------------
4         192.168.0.100
5         192.168.0.101
6         192.168.0.102
```

说明

- ZooKeeper的quorumpeer实例业务IP：

  ZooKeeper服务所有quorumpeer实例业务IP。登录FusionInsight Manager，选择“集群 > 服务 > ZooKeeper > 实例”，可查看所有quorumpeer实例所在主机业务IP地址。

- ZooKeeper客户端端口号：

  登录FusionInsight Manager，选择“集群 > 服务 > ZooKeeper”，在“配置”页签查看“clientPort”的值。默认为24002。

步骤8 从**步骤6**和**步骤7**回显中获取分区的分布信息和节点信息，在当前目录下创建执行重新分配的json文件。

以迁移的是Broker_ID为6的节点的分区为例，迁移到"/srv/BigData/hadoop/data1/kafka-logs”，完成迁移所需的json配置文件，内容如下。

```
{"partitions":[{"topic": "testws","partition": 2,"replicas": [6,5],"log_dirs": ["/srv/BigData/hadoop/data1/kafka-logs","any"]}],"version":1}
```

说明

- topic为Topic名称，此处以testws为例，具体以实际为准。
- partition为Topic分区。
- replicas中的数字对应Broker_ID。
- log_dirs为需要迁移的磁盘路径。此样例迁移的是Broker_ID为6的节点，Broker_ID为5的节点对应的log_dirs可设置为“any”，Broker_ID为6的节点对应的log_dirs设置为“/srv/BigData/hadoop/data1/kafka-logs”。**注意路径需与节点对应**。

步骤9 使用如下命令，执行重分配操作。

安全模式：

./kafka-reassign-partitions.sh --bootstrap-server *Broker业务IP*:**21007 --command-config ../config/client.properties --zookeeper** *{zk_host}:{port}*/**kafka --reassignment-json-file** ***步骤8****中编写的json文件路径* **--execute**

普通模式：

./kafka-reassign-partitions.sh --bootstrap-server *Broker业务IP*:**21005 --command-config ../config/client.properties --zookeeper** *{zk_host}:{port}*/**kafka --reassignment-json-file** ***步骤8****中编写的json文件路径* **--execute**

提示“Successfully started reassignment of partitions”表示执行成功。

----结束

12.14.22 Kafka 常见问题

12.14.22.1 如何解决 Kafka topic 无法删除的问题

问题

删除Kafka topic后发现未成功删除，如何正常删除?

回答

- 可能原因一：配置项“delete.topic.enable”未配置为“true”，只有配置为“true”才能执行真正删除。
- 可能原因二：“auto.create.topics.enable”配置为“true”，其他应用程序有使用该Topic，并且一直在后台运行。

解决方法：

- 针对原因一：配置页面上将“delete.topic.enable”设置为“true”。
- 针对原因二：先停掉后台使用该Topic的应用程序，或者“auto.create.topics.enable”配置为“false”（需要重启Kafka服务），然后再做删除操作。

12.15 使用 KafkaManager

12.15.1 KafkaManager 介绍

KafkaManager是Apache Kafka的管理工具，提供Kafka集群界面化的Metric监控和集群管理。

通过KafkaManager可以：

- 支持管理多个Kafka集群
- 支持界面检查集群状态（主题，消费者，偏移量，分区，副本，节点）
- 支持界面执行副本的leader选举
- 使用选择生成分区分配以选择要使用的分区方案
- 支持界面执行分区重新分配（基于生成的分区方案）
- 支持界面选择配置创建主题（支持多种Kafka版本集群）
- 支持界面删除主题（仅支持0.8.2+并设置了delete.topic.enable = true）
- 支持批量生成多个主题的分区分配，并可选择要使用的分区方案
- 支持批量运行重新分配多个主题的分区
- 支持为已有主题增加分区
- 支持更新现有主题的配置
- 可以为分区级别和主题级别度量标准启用JMX查询
- 可以过滤掉zookeeper中没有ids / owner /＆offsets /目录的使用者。

12.15.2 访问 KafkaManager 的 WebUI

用户可以通过KafkaManager的WebUI，在图形化界面监控管理Kafka集群。

前提条件

- 已安装KafkaManager服务的集群。
- 获取用户“admin”帐号密码。“admin”密码在创建MRS集群时由用户指定。

访问 KafkaManager 的 WebUI

步骤1 登录集群详情页面，选择“组件管理 > KafkaManager”。

说明

若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 在KafkaManager概述的“KafkaManager WebUI”中单击任意一个UI链接，打开KafkaManager的WebUI页面。

KafkaManager的WebUI支持查看以下信息：

- Kafka集群列表
- Kafka集群Broker节点列表和Metric监控
- Kafka集群副本监控
- Kafka集群Consumer监控

说明

在KafkaManager的任何子页面单击左上角KafkaManager的Logo都可以回到KafkaManager的WebUI主界面，显示集群列表信息。

----结束

12.15.3 管理 Kafka 集群

管理Kafka集群包含以下内容：

- **添加集群到KafkaManager的WebUI界面**
- **更新集群参数**
- **删除KafkaManager的WebUI界面的集群**

添加集群到 KafkaManager 的 WebUI 界面

首次创建Kafka集群后会在KafkaManager的WebUI界面创建名为my-cluster的默认Kafka集群，用户也可以在KafkaManager的WebUI界面自行添加已经通过MRS控制台创建的Kafka集群，用于管理多个Kafka集群。

步骤1 登录KafkaManager的WebUI界面。

步骤2 在页面上方选择“Cluster > Add Cluster” 。

步骤3 设置待添加集群的参数，如下参数请参考样例，其他参数默认不需要修改。

表 12-273 需修改的集群参数

参数名称	取值样例	说明
Cluster Name	mrs-demo	待添加集群在KafkaManager的WebUI界面中显示的名称。
Cluster Zookeeper Hosts	zk1_ip:zk1_port,zk2_ip:zk2_port/kafka	待添加集群的Zookeeper地址。
Kafka Version	1.1.0	待添加集群的Kafka版本，默认1.1.0。
Enable JMX Polling (Set JMX_PORT env variable before starting kafka server)	勾选	-
Poll consumer information (Not recommended for large # of consumers)	勾选	-
Enable Active OffsetCache (Not recommended for large # of consumers)	勾选	-

参数名称	取值样例	说明
Display Broker and Topic Size (only works after applying this patch)	勾选	-
Security Protocol	PLAINTEXT	• 开启Kerberos的Kafka集群选择SASL_PLAINTEXT • 未开启Kerberos集群选择PLAINTEXT

步骤4 单击“Save”完成添加集群。

----结束

更新集群参数

步骤1 登录KafkaManager的WebUI界面。

步骤2 在对应集群的“Operations”列单击“Modify”。

步骤3 进入集群配置参数页面，修改集群参数。

----结束

删除 KafkaManager 的 WebUI 界面的集群

步骤1 登录KafkaManager的WebUI界面。

步骤2 在对应集群的“Operations”列单击“Disable”。

步骤3 等待集群列表页面的“Operations”列出现“Delete”或“Enable”时，单击“Delete”删除集群。也可以单击“Enable”启用集群。

----结束

12.15.4 Kafka 集群监控管理

Kafka集群监控管理包含以下内容：

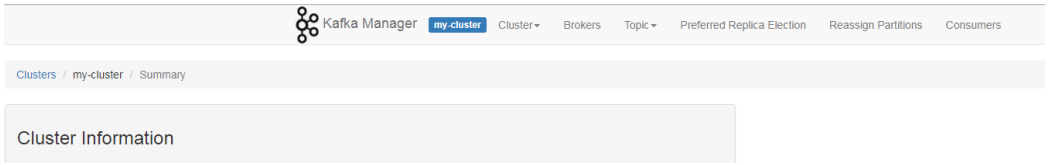
- **查看Broker信息**
- **查看Topic信息**
- **查看Consumers信息**
- **通过KafkaManager修改Topic的partition**

查看 Broker 信息

步骤1 登录KafkaManager的WebUI界面。

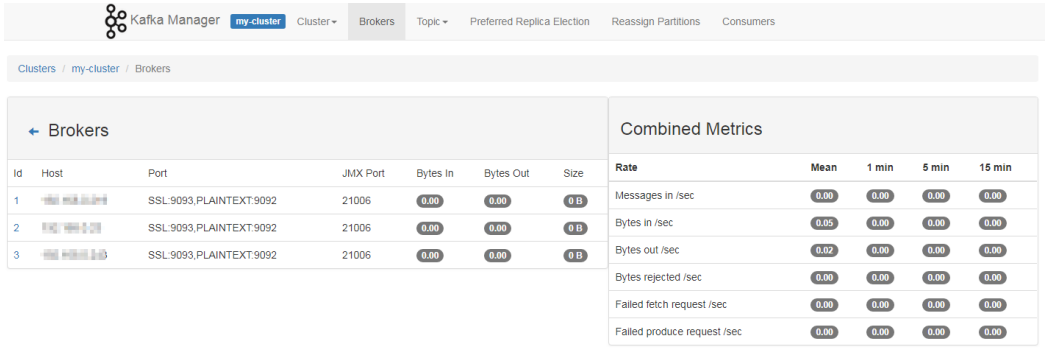
步骤2 在集群列表页面单击对应集群名称进入集群Summary页面。

图 12-24 集群 Summary 页面



步骤3 单击“Brokers”进入Broker监控页面，该页面包括Broker列表和Broker节点的IO统计信息。

图 12-25 Broker 监控页面



----结束

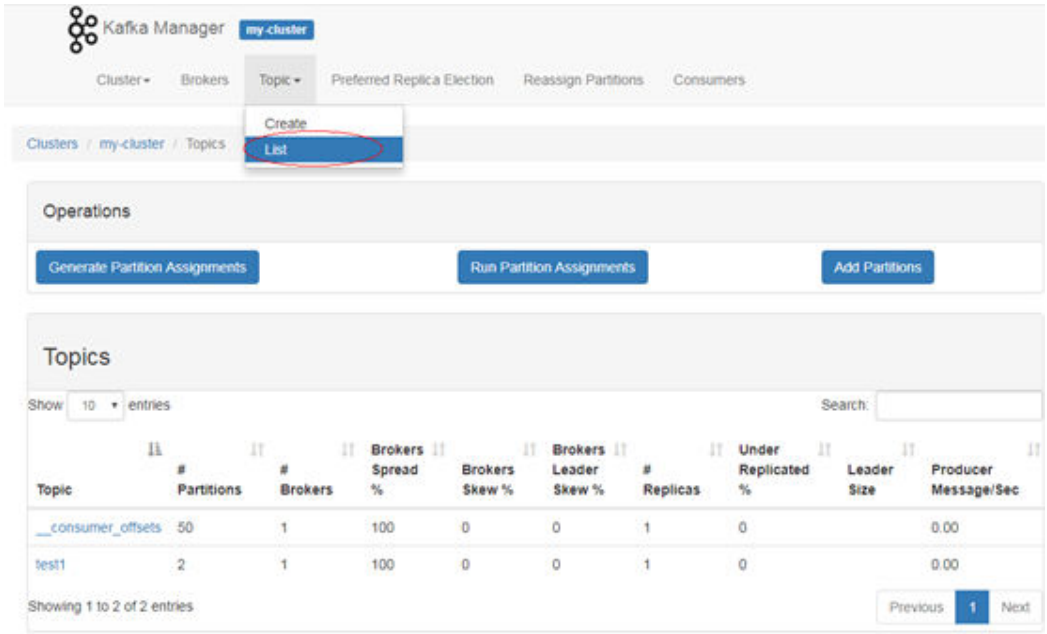
查看 Topic 信息

步骤1 登录KafkaManager的WebUI界面。

步骤2 在集群列表页面单击对应集群名称进入集群Summary页面。

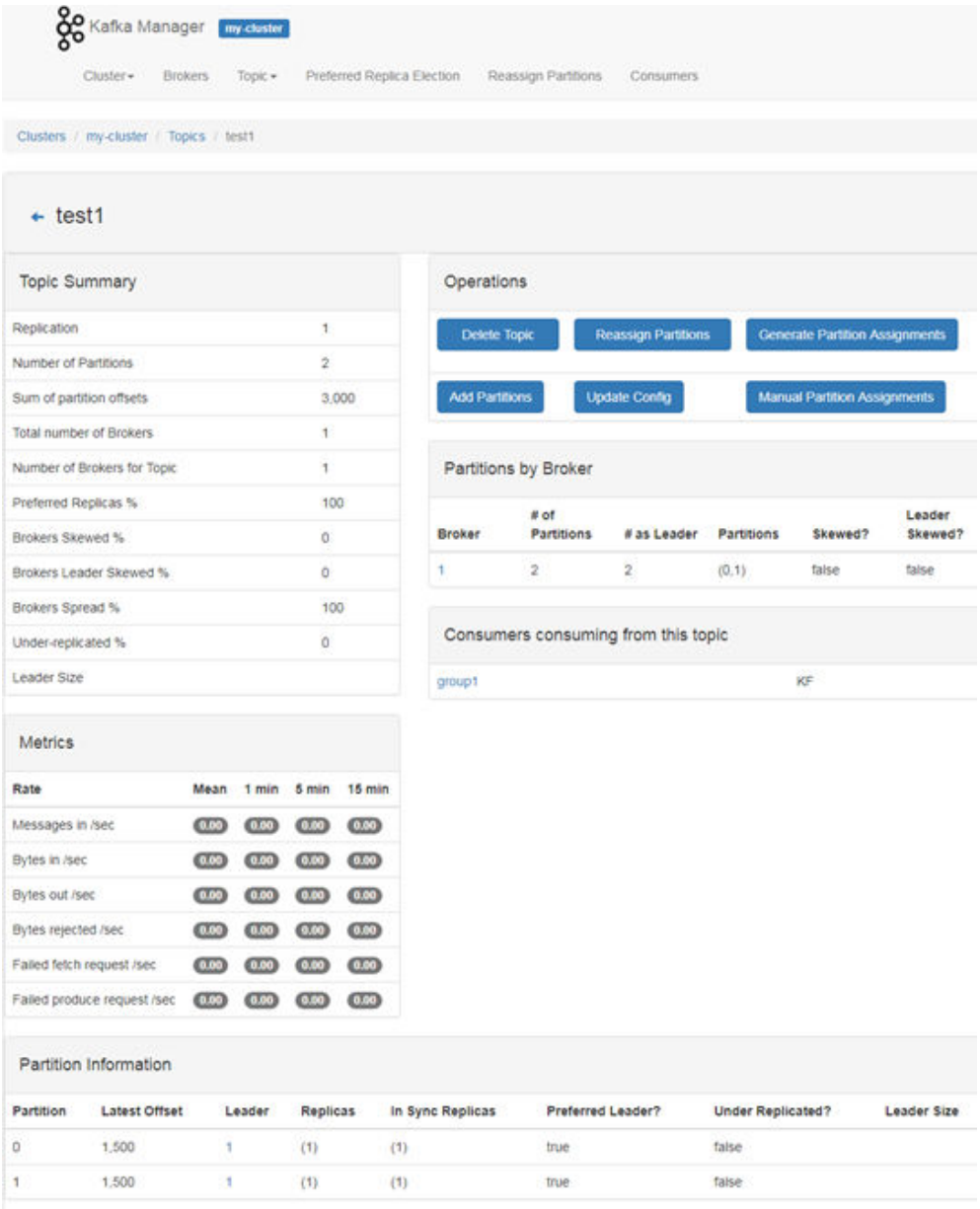
步骤3 单击“Topic > List”查看当前集群的Topic列表及每个Topic的相关信息。

图 12-26 Topic 列表



步骤4 单击具体的Topic名称查看该Topic的详细信息。

图 12-27 Topic 的详细信息



----结束

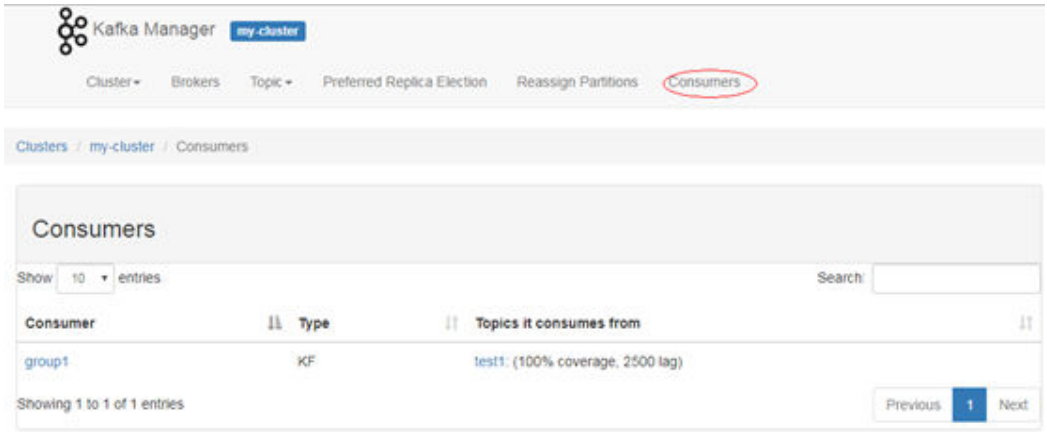
查看 Consumers 信息

步骤1 登录KafkaManager的WebUI界面。

步骤2 在集群列表页面单击对应集群名称进入集群Summary页面。

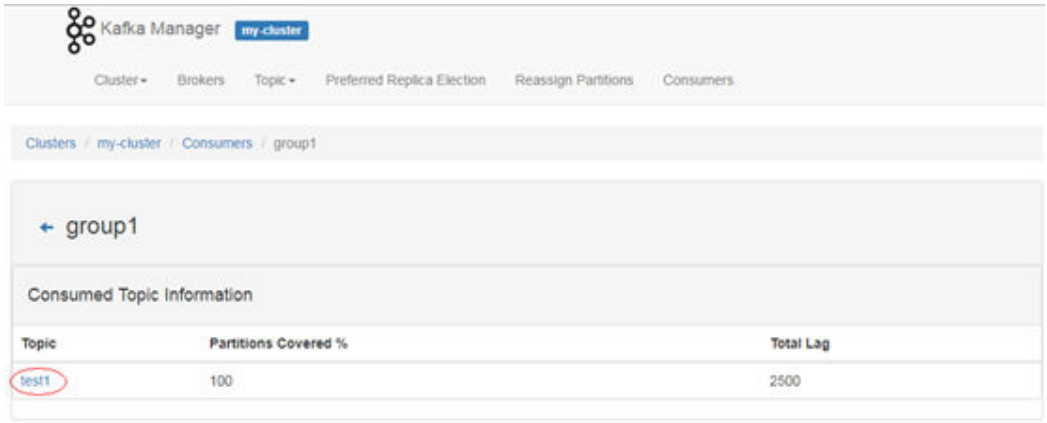
步骤3 单击“Consumers”查看当前集群的Consumers列表及每个Consumer的消费信息。

图 12-28 Consumers 列表



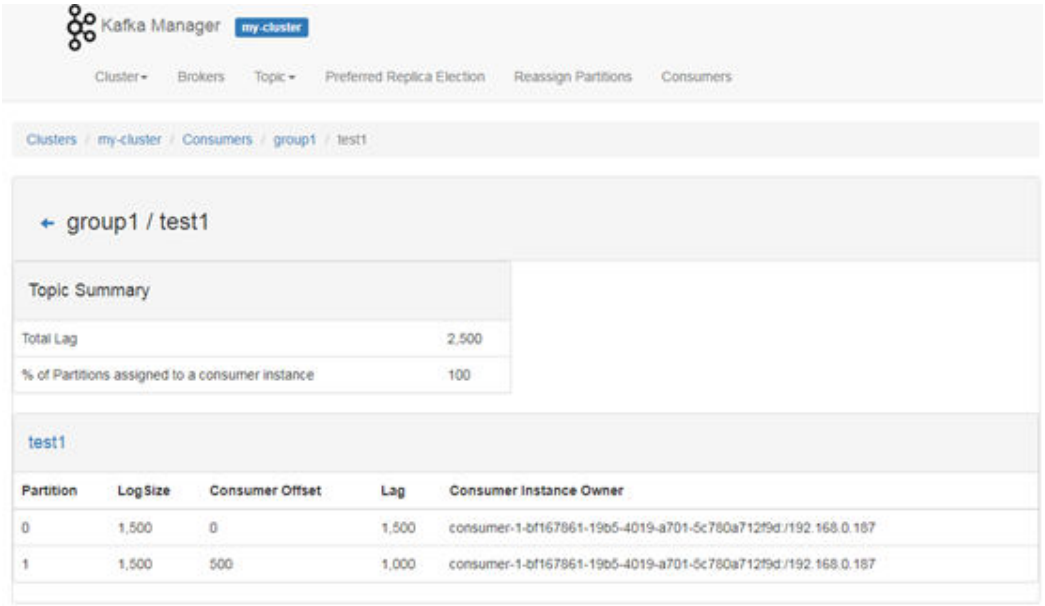
步骤4 单击Consumer的名称查看消费的Topic列表。

图 12-29 Consumer 消费的 Topic 列表



步骤5 单击Consumer下Topic列表中的Topic名称，查看该Consumer对Topic的具体消费情况。

图 12-30 Consumer 对 Topic 的具体消费情况



----结束

通过 KafkaManager 修改 Topic 的 partition

步骤1 登录KafkaManager的WebUI界面。

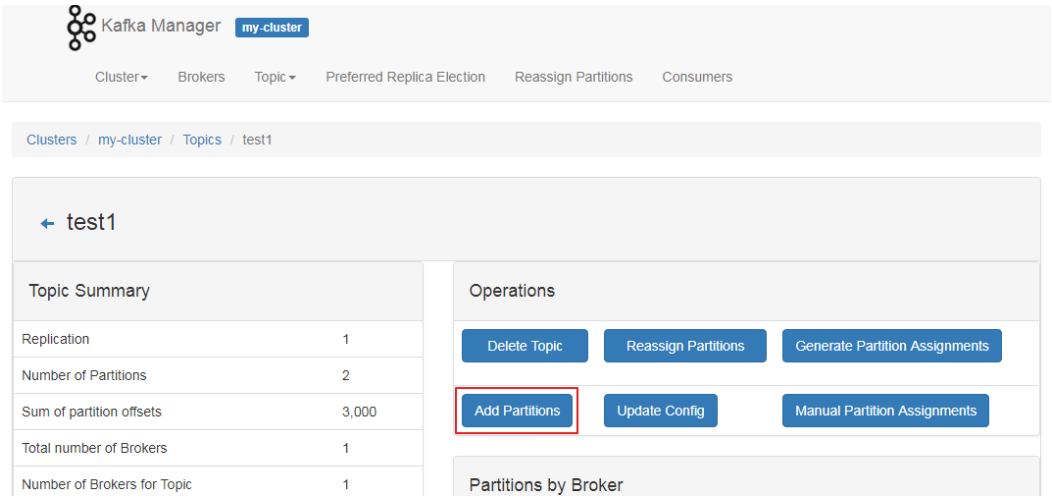
步骤2 在集群列表页面单击对应集群名称进入集群Summary页面。

步骤3 单击“Topic > List”进入当前集群的Topic列表页面。

步骤4 单击具体的Topic名称进入Topic Summary页面。

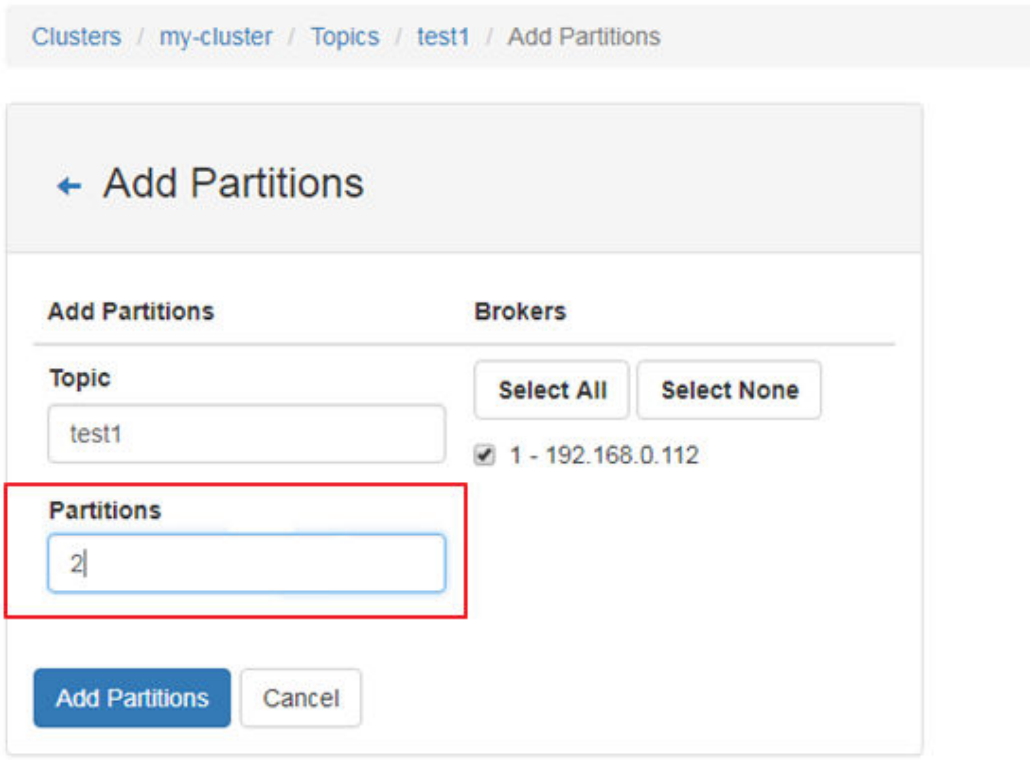
步骤5 单击“add partitions”，进入添加分区页面。

图 12-31 添加分区



步骤6 确认Topic名称并修改“Partitions”数量，单击“Add Partitions”进行分区添加。

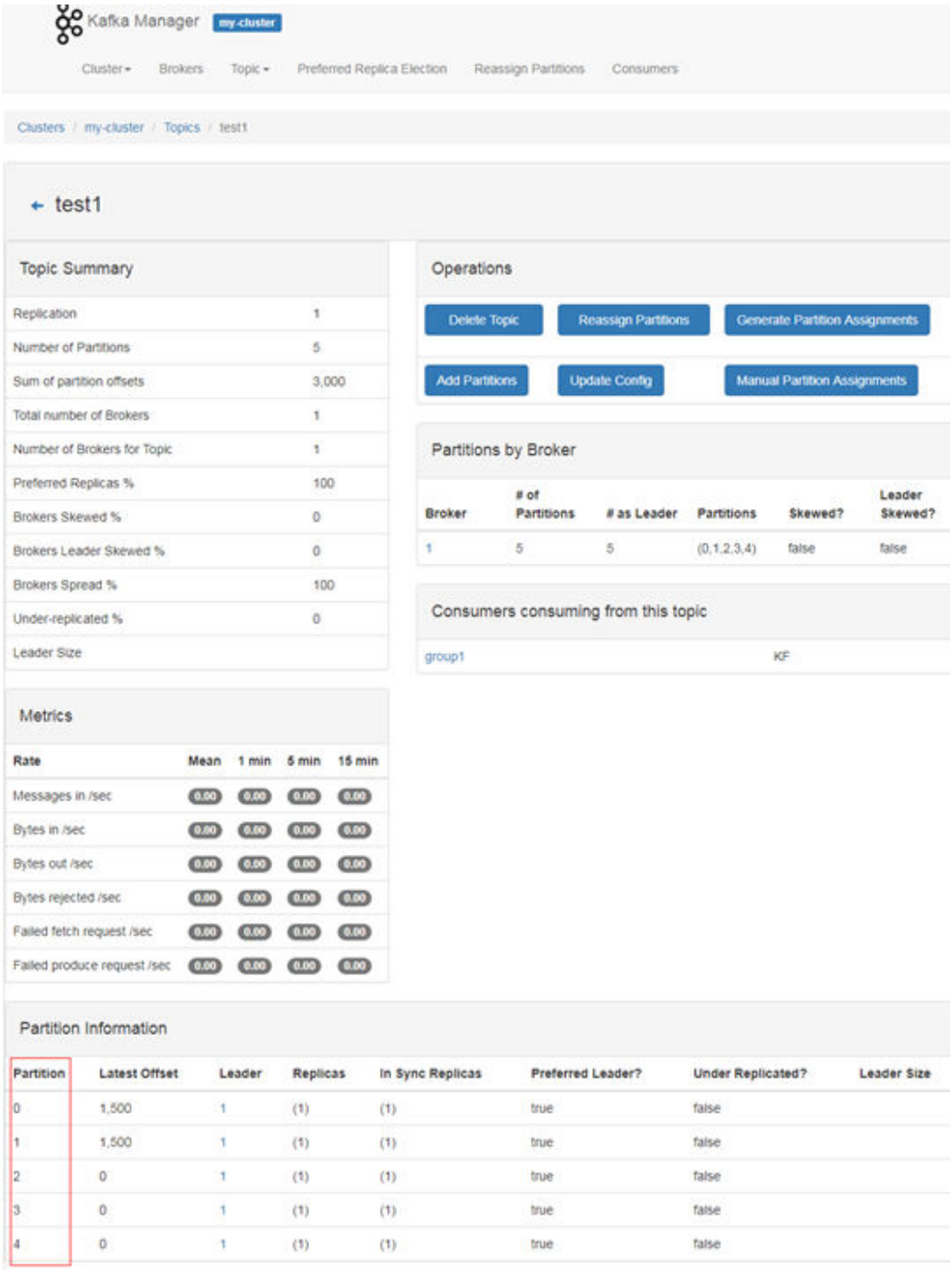
图 12-32 修改 Partitions 数量



步骤7 分区添加成功后，单击“Go to topic view.”返回Topic Summary页面。

步骤8 在Topic Summary页面的下方“Partition Information”中确认partition数量。

图 12-33 Partition Information



步骤9 （可选）若对分配的分区不满意，可以执行Partition的重新分配功能来重新自动分配分区。

1. 在Topic Summary页面单击“Generate Partition Assignments”。
2. 勾选broker实例，单击“Generate Partition Assignments”生成分区。
3. 分区生成完成，单击“Go to topic view.”返回Topic Summary页面。
4. 在Topic Summary页面单击“Reassign Partitions”可以在集群的broker实例上重新自动分配分区。

5. 单击“Go to reassign partitions.”查看重新分配的分区详情。

步骤10 （可选）若对自动分配的分区不满意，可以执行手动分配来重新分配分区。

1. 在Topic Summary页面单击“Manual Partition Assignments”进入手动分配分区页面。
2. 手动为每个分区的副本分配Broker id，然后单击“Save Partition Assignment”保存修改。
3. 单击“Go to topic view.”返回Topic Summary页面，查看分区详情。

----结束

12.16 使用 Kudu

12.16.1 从零开始使用 Kudu

Kudu是专为Apache Hadoop平台开发的列式存储管理器。Kudu具有Hadoop生态系统应用程序的共同技术特性：可水平扩展，并支持高可用性操作。

前提条件

已安装集群客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 运行Kudu命令行工具。

直接执行Kudu组件的命令行工具，查看帮助。

kudu -h

回显信息如下：

```
Usage: kudu <command> [<args>]

<command> can be one of the following:
        cluster   Operate on a Kudu cluster
       diagnose   Diagnostic tools for Kudu servers and clusters
             fs   Operate on a local Kudu filesystem
           hms    Operate on remote Hive Metastores
  local_replica   Operate on local tablet replicas via the local filesystem
         master   Operate on a Kudu Master
            pbc   Operate on PBC (protobuf container) files
           perf   Measure the performance of a Kudu cluster
 remote_replica   Operate on remote tablet replicas on a Kudu Tablet Server
          table   Operate on Kudu tables
         tablet   Operate on remote Kudu tablets
           test   Various test actions
```

```
   tserver   Operate on a Kudu Tablet Server
       wal   Operate on WAL (write-ahead log) files
```

说明

kudu命令行工具不提供DDL、DML等操作，但提供针对cluster、master、tserver、fs、table等的细化查询功能。

常用操作：

- 查看当前集群下有哪些表。

  kudu table list *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051*

- 查询Kudu服务KuduMaster实例的配置信息。

  kudu master get_flags *KuduMaster实例IP:7051*

- 查询表的schema。

  kudu table describe *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051 tablename*

- 删除表。

  kudu table delete *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051 tablename*

  说明

  KuduMaster实例IP获取方式：在集群详情页面，选择“组件管理 > Kudu > 实例”，获取角色KuduMaster的IP地址。

----结束

12.16.2 访问 Kudu 的 WebUI

用户可以通过Kudu的WebUI，在图形化界面查看Kudu作业的相关信息。

前提条件

已安装Kudu服务的集群。

访问 KuduMaster WebUI（MRS 3.x 及之后版本）

步骤1 登录Manager页面，请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤2 选择“集群 > 服务 > Kudu”。

步骤3 在“Kudu 概览”的“KuduMaster WebUI”中单击“KuduMaster(KuduMaster)”，打开KuduMaster的WebUI页面。

图 12-34 KuduMaster WebUI

----结束

访问 KuduMaster WebUI（MRS 3.x 之前版本）

步骤1 登录Manager页面，请参见**访问MRS Manager（MRS 3.x之前版本）**。

步骤2 选择“服务管理 > Kudu”。

步骤3 在“Kudu 概述”的“KuduMaster WebUI”中单击“KuduMaster(KuduMaster)”，打开KuduMaster的WebUI页面。

----结束

12.17 使用 Loader

12.17.1 从零开始使用 Loader

用户可以使用Loader将数据从SFTP服务器导入到HDFS。

本章节适用于MRS 3.x之前版本。

前提条件

- 已准备业务数据。
- 已创建分析集群。

操作步骤

步骤1 访问Loader页面。

1. 登录集群详情页面，选择“服务管理”。

2. 选择“Hue”，在“Hue概述”的“Hue WebUI”，单击“Hue (主)”，打开Hue的WebUI。
3. 选择“Data Browsers > Sqoop”。
   默认显示Loader页面中的作业管理界面。

步骤2 在Loader页面，单击“管理连接”。

步骤3 单击“新建连接”，参考**文件服务器连接**，创建sftp-connector。

步骤4 单击“新建连接”，输入连接名称，选择连接器为hdfs-connector，创建hdfs-connector。

步骤5 访问Loader页面，单击“管理作业”。

步骤6 单击“新建作业”。

步骤7 在“基本信息”填写参数。

1. 在“名称”填写一个作业的名称。
2. 选择**步骤3**创建的“源连接”和**步骤4**创建的“目的连接”。

步骤8 在“自”填写源连接的作业配置。

具体请参见**ftp-connector或sftp-connector**。

步骤9 在“至”填写目的连接的作业配置。

具体请参见**hdfs-connector**。

步骤10 在“任务配置”填写作业的运行参数。

表 12-274 Loader 作业运行属性

参数	说明
抽取并发数	设置map任务的个数。
加载(写入)并发数	设置reduce任务的个数。 该参数只有在目的字段为Hbase和Hive时才会显示。
单个分片的最大错误记录数	设置一个错误阈值，如果单个map任务的错误记录超过设置阈值则任务自动结束，已经获取的数据不回退。 **说明** “generic-jdbc-connector”的“MYSQL”和“MPPDB”默认批量读写数据，每一批次数据最多只记录一次错误记录。
脏数据目录	设置一个脏数据目录，在出现脏数据的场景中在该目录保存脏数据。如果不设置则不保存。

步骤11 单击“保存”。

----结束

12.17.2 Loader 使用简介

本章节适用于MRS 3.x之前版本。

使用流程

通过Loader迁移用户数据时，基本流程如下所示。

1. 访问Hue WebUI的Loader页面。
2. 管理Loader连接。
3. 创建作业，选择数据源的连接以及保存数据的连接。
4. 运行作业，完成数据迁移。

Loader 页面介绍

Loader页面是基于开源Sqoop WebUI的图形化数据迁移管理工具，该页面托管在Hue的WebUI中。进入Loader页面请执行以下操作：

1. 访问Hue WebUI，参见**访问Hue的WebUI**。
2. 选择“Data Browsers > Sqoop”。
   默认显示Loader页面中的作业管理界面。

Loader 连接介绍

Loader连接保存了数据具体位置的相关信息，Loader使用连接来访问数据，或将数据保存到指定的位置。进入Loader连接管理页面请执行以下操作：

1. 进入Loader页面。
2. 单击“管理连接”。
   显示Loader连接管理页面。
   可单击“管理作业”回到作业管理页面。
3. 单击“新建连接”，进入配置页面，并填写参数创建一个Loader连接。

Loader 作业介绍

Loader作业用于管理数据迁移任务，每个作业包含一个源数据的连接，和一个目的数据的连接，通过从源连接读取数据，再将数据保存到目的连接，完成数据迁移任务。

12.17.3 Loader 连接配置说明

本章节适用于MRS 3.x之前版本。

基本介绍

Loader支持以下多种连接，每种连接的配置介绍可根据本章节内容了解。

- obs-connector
- generic-jdbc-connector
- ftp-connector或sftp-connector
- hbase-connector、hdfs-connector或hive-connector

OBS 连接

OBS连接是Loader与OBS进行数据交换的通道，配置参数如**表12-275**所示。

表 12-275 obs-connector 配置

参数	说明
名称	指定一个Loader连接的名称。
OBS服务器	输入OBS endpoint地址，一般格式为 **OBS.*Region*.*DomainName***。 例如执行如下命令查看OBS endpoint地址： **cat /opt/Bigdata/apache-tomcat-7.0.78/webapps/web/WEB-INF/classes/cloud-obs.properties**
端口	访问OBS数据的端口。默认值为“443”。
访问标识(AK)	表示访问OBS的用户的访问密钥AK。
密钥(SK)	表示访问密钥对应的SK。

关系型数据库连接

关系型数据库连接是Loader与关系型数据库进行数据交换的通道，配置参数如表12-276所示。

说明

部分参数需要单击“显示高级属性”后展开，否则默认隐藏。

表 12-276 generic-jdbc-connector 配置

参数	说明
名称	指定一个Loader连接的名称。
数据库类型	表示Loader连接支持的数据，可以选择“ORACLE”、“MYSQL”和“MPPDB”。
数据库服务器	表示数据库的访问地址，可以是IP地址或者域名。
端口	表示数据库的访问端口。
数据库名称	表示保存数据的具体数据库名。
用户名	表示连接数据库使用的用户名称。
密码	表示此用户对应的密码。需要与实际密码保持一致。

表 12-277 高级属性配置

参数	说明
一次请求行数	表示每次连接数据库时，最多可获取的数据量。

参数	说明
连接属性	不同类型数据库支持该数据库连接特有的驱动属性，例如MYSQL的“autoReconnect”。如果需要定义驱动属性，单击“添加”。
引用符号	表示数据库的SQL中保留关键字的定界符，不同类型数据库定义的定界符不完全相同。

文件服务器连接

文件服务器连接包含FTP连接和SFTP连接，是Loader与文件服务器进行数据交换的通道，配置参数如表12-278所示。

表 12-278 ftp-connector 或 sftp-connector 配置

参数	说明
名称	指定一个Loader连接的名称。
主机名或IP	输入文件服务器的访问地址，可以是服务器的主机名或者IP地址。
端口	访问文件服务器的端口。 • FTP协议请使用端口“21”。 • SFTP协议请使用端口“22”。
用户名	表示文件服务器的用户名称。
密码	表示此用户对应的密码。

MRS 集群连接

MRS集群连接包含HBase连接、HDFS连接和Hive连接，是Loader与对应各数据进行数据交换的通道。

配置MRS集群连接时，需要设置名称、选择对应的连接器“hbase-connector”、“hdfs-connector”或“hive-connector”，然后保存即可。

12.17.4 管理 Loader 连接（MRS 3.x 之前版本）

操作场景

Loader页面支持创建、查看、编辑和删除连接。

本章节适用于MRS 3.x之前版本。

前提条件

已访问Loader页面，参见Loader页面介绍。

创建连接

步骤1 在Loader页面，单击“管理连接”。

步骤2 单击“新建连接”，配置连接参数。

参数介绍具体可参见**Loader连接配置说明**。

步骤3 单击“保存”。

如果连接配置，例如IP地址、端口、访问用户等信息不正确，将导致验证连接失败无法保存。另外VPC相关设置，也可能影响网络连通性。

说明

用户可以直接单击“测试”立即检测连接是否可用。

----结束

查看连接

步骤1 在Loader页面，单击“管理连接”。

- 如果集群启用了Kerberos认证，则默认显示所有当前用户创建的连接，不支持显示其他用户创建的连接。
- 如果集群未启用Kerberos认证，则显示集群中全部的Loader连接。

步骤2 在“Sqoop连接”中输入指定连接的名称，可以筛选该连接。

----结束

编辑连接

步骤1 在Loader页面，单击“管理连接”。

步骤2 单击指定连接的名称，进入编辑页面。

步骤3 根据业务需要，修改连接配置参数。

步骤4 单击“测试”。

如果显示测试成功，则执行**步骤5**；如果显示不能连接至OBS Server，则需要重复**步骤3**。

步骤5 单击“保存”。

如果某个Loader作业已集成一个Loader连接，那么编辑连接参数后可能导致Loader作业运行效果也产生变化。

----结束

删除连接

步骤1 在Loader页面，单击“管理连接”。

步骤2 在指定连接所在行，单击“删除”。

步骤3 在弹出的对话框窗口，单击“是，将其删除”。

如果某个Loader作业已集成一个Loader连接，那么该连接不可以被删除。

----结束

12.17.5 Loader 作业源连接配置说明

基本介绍

Loader作业需要从不同数据源获取数据时，应该选择对应类型的连接，每种连接在该场景中需要配置连接的属性。

本章节适用于MRS 3.x之前版本。

obs-connector

表 12-279 obs-connector 数据源连接属性

参数	说明
桶名	保存源数据的OBS文件系统。
源目录或文件	源数据实际存储的形态，可能是文件系统包含一个目录中的全部数据文件，或者是文件系统包含的单个数据文件。
文件格式	Loader支持OBS中存储数据的文件格式，默认支持以下两种： ● CSV_FILE：表示文本格式文件。目的连接为数据库型连接时，只支持文本格式。 ● BINARY_FILE：表示文本格式以外的二进制文件。
换行符	源数据的每行结束标识字符。
字段分割符	源数据的每个字段分割标识字符。
编码类型	源数据的文本编码类型。只对文本类型文件有效。
文件分割方式	支持以下两种： ● File：按总文件个数分配map任务处理的文件数量，计算规则为“文件总个数/抽取并发数”。 ● Size：按文件总大小分配map任务处理的文件大小，计算规则为“文件总大小/抽取并发数”。

generic-jdbc-connector

表 12-280 generic-jdbc-connector 数据源连接属性

参数	说明
模式或表空间	表示源数据对应的数据库名称，支持通过界面查询并选择。
表名	存储源数据的数据表，支持通过界面查询并选择。

参数	说明
抽取分区字段	分区字段，如果需读取多个字段，使用该字段分割结果并获取数据。
Where子句	表示读取数据库时使用的查询语句。

ftp-connector 或 sftp-connector

表 12-281 ftp-connector 或 sftp-connector 数据源连接属性

参数	说明
源目录或文件	源数据实际存储的形态，可能是文件服务器包含一个目录中的全部数据文件，或者是单个数据文件。
文件格式	Loader支持文件服务器中存储数据的文件格式，默认支持以下两种： ● CSV_FILE：表示文本格式文件。目的连接为数据库型连接时，只支持文本格式。 ● BINARY_FILE：表示文本格式以外的二进制文件。
换行符	源数据的每行结束标识字符。 **说明** ftp或sftp作为源连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“换行符”配置无效
字段分割符	源数据的每个字段分割标识字符。 **说明** ftp或sftp作为源连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“字段分割符”配置无效
编码类型	源数据的文本编码类型。只对文本类型文件有效。
文件分割方式	支持以下两种： ● File：按总文件个数分配map任务处理的文件数量，计算规则为“文件总个数/抽取并发数”。 ● Size：按文件总大小分配map任务处理的文件大小，计算规则为“文件总大小/抽取并发数”。

hbase-connector

表 12-282 hbase-connector 数据源连接属性

参数	说明
表名	源数据实际存储的HBase表。

hdfs-connector

表 12-283 hdfs-connector 数据源连接属性

参数	说明
源目录或文件	源数据实际存储的形态，可能是HDFS包含一个目录中的全部数据文件，或者是单个数据文件。
文件格式	Loader支持HDFS中存储数据的文件格式，默认支持以下两种： ● CSV_FILE：表示文本格式文件。目的连接为数据库型连接时，只支持文本格式。 ● BINARY_FILE：表示文本格式以外的二进制文件。
换行符	源数据的每行结束标识字符。 **说明** hdfs作为源连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“换行符”配置无效。
字段分割符	源数据的每个字段分割标识字符。 **说明** hdfs作为源连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“字段分割符”配置无效。
文件分割方式	支持以下两种： ● File：按总文件个数分配map任务处理的文件数量，计算规则为“文件总个数/抽取并发数”。 ● Size：按文件总大小分配map任务处理的文件大小，计算规则为“文件总大小/抽取并发数”。

hive-connector

表 12-284 hive-connector 数据源连接属性

参数	说明
数据库名称	数据源的Hive数据库名称，支持通过界面查询并选择。
表名	数据源的Hive表名称，支持通过界面查询并选择。

12.17.6 Loader 作业目的连接配置说明

基本介绍

Loader作业需要将数据保存到不同目的存储位置时，应该选择对应类型的目的连接，每种连接在该场景中需要配置连接的属性。

obs-connector

表 12-285 obs-connector 目的连接属性

参数	说明
桶名	保存最终数据的OBS文件系统。
写入目录	最终数据在文件系统保存时的具体目录。必须指定一个目录。
文件格式	Loader支持OBS中存储数据的文件格式，默认支持以下两种： ● CSV_FILE：表示文本格式文件。目的连接为数据库型连接时，只支持文本格式。 ● BINARY_FILE：表示文本格式以外的二进制文件。
换行符	最终数据的每行结束标识字符。
字段分割符	最终数据的每个字段分割标识字符。
编码类型	最终数据的文本编码类型。只对文本类型文件有效。

generic-jdbc-connector

表 12-286 generic-jdbc-connector 目的连接属性

参数	说明
模式名称	保存最终数据的数据库名称。
表名	保存最终数据的数据表名称。

ftp-connector 或 sftp-connector

表 12-287 ftp-connector 或 sftp-connector 目的连接属性

参数	说明
写入目录	最终数据在文件服务器保存时的具体目录。必须指定一个目录。
文件格式	Loader支持文件服务器中存储数据的文件格式，默认支持以下两种： ● CSV_FILE：表示文本格式文件。目的连接为数据库型连接时，只支持文本格式。 ● BINARY_FILE：表示文本格式以外的二进制文件。

参数	说明
换行符	最终数据的每行结束标识字符。 **说明** ftp或sftp作为目的连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“换行符”配置无效。
字段分割符	最终数据的每个字段分割标识字符。 **说明** ftp或sftp作为目的连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“字段分割符”配置无效
编码类型	最终数据的文本编码类型。只对文本类型文件有效。

hbase-connector

表 12-288 hbase-connector 目的连接属性

参数	说明
表名	保存最终数据的HBase表名称，支持通过界面查询并选择。
导入方式	支持BULKLOAD、PUTLIST两种方式导入数据到HBase表。
导入前清空数据	标识是否需要清空目标HBase表中的数据，支持以下两种类型： ● True：清空表中的数据。 ● False：不清空表中的数据，选择False时如果表中存在数据，则作业运行会报错。

hdfs-connector

表 12-289 hdfs-connector 目的连接属性

参数	说明
写入目录	最终数据在HDFS保存时的具体目录。必须指定一个目录。
文件格式	Loader支持HDFS中存储数据的文件格式，默认支持以下两种： ● CSV_FILE：表示文本格式文件。目的连接为数据库型连接时，只支持文本格式。 ● BINARY_FILE：表示文本格式以外的二进制文件。
压缩格式	文件在HDFS保存时的压缩行为。支持NONE、DEFLATE、GZIP、BZIP2、LZ4和SNAPPY。

参数	说明
是否覆盖	文件在导入HDFS时对写入目录中原有文件的处理行为，支持以下两种： ● True：默认清空目录中的文件并导入新文件。 ● False：不清空文件。如果写入目录中有文件，则作业运行失败。
换行符	最终数据的每行结束标识字符。 **说明** hdfs作为目的连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“换行符”配置无效。
字段分割符	最终数据的每个字段分割标识字符。 **说明** hdfs作为目的连接时，当“文件格式”配置为BINARY_FILE时，高级属性中的“字段分割符”配置无效

hive-connector

表 12-290 hive-connector 目的连接属性

参数	说明
数据库名称	保存最终数据的Hive数据库名称，支持通过界面查询并选择。
表名	保存最终数据的Hive表名称，支持通过界面查询并选择。

12.17.7 管理 Loader 作业

操作场景

Loader页面支持创建、查看、编辑和删除作业。

本章节适用于MRS 3.x之前版本。

前提条件

已访问Loader页面，参见**Loader页面介绍**。

创建作业

步骤1 访问Loader页面，单击“新建作业”。

步骤2 在“基本信息”填写参数。

1. 在“名称”填写一个作业的名称。
2. 在“源连接”和“目的连接”选择对应的连接。
   选择某个类型的连接，表示从指定的源获取数据，并保存到目的位置。

说明

如果没有需要的连接，可单击“添加新连接”。

步骤3 在“自”填写源连接的作业配置。

具体请参见**Loader作业源连接配置说明**。

步骤4 在“至”填写目的连接的作业配置。

具体请参见**Loader作业目的连接配置说明**。

步骤5 在“目的连接”是否选择了数据库类型的连接?

数据库类型的连接包含以下几种：

- generic-jdbc-connector
- hbase-connector
- hive-connector

“目的连接”选择数据库类型的连接时，还需要配置业务数据与数据库表字段的对应关系：

- 是，请执行**步骤6**。
- 否，请执行**步骤7**。

步骤6 在“字段映射”填写字段对应关系。然后执行**步骤7**。

“字段映射”的对应关系，表示用户数据中每一列与数据库的表字段的匹配关系。

表 12-291 “字段映射”属性

参数	说明
列号	表示业务数据的字段顺序。
样本	表示业务数据的第一行值样例。
列族	“目的连接”为hbase-connector类型时，支持定义保存数据的具体列族。
目的字段	配置保存数据的具体字段。
类型	显示用户选择字段的类型。
行键	“目的连接”为hbase-connector类型时，需要勾选作为行键的“目的字段”。

说明

如果From是sftp/ftp/obs/hdfs等文件类型连接器，Field Mapping 样值取自文件第一行数据，需要保证第一行数据是完整的，Loader作业不会抽取没有Mapping上的列。

步骤7 在“任务配置”填写作业的运行参数。

表 12-292 Loader 作业运行属性

参数	说明
抽取并发数	设置map任务的个数。
加载(写入)并发数	设置reduce任务的个数。 该参数只有在目的字段为Hbase和Hive时才会显示。
单个分片的最大错误记录数	设置一个错误阈值，如果单个map任务的错误记录超过设置阈值则任务自动结束，已经获取的数据不回退。 **说明** “generic-jdbc-connector”的“MYSQL”和“MPPDB”默认批量读写数据，每一批次数据最多只记录一次错误记录。
脏数据目录	设置一个脏数据目录，在出现脏数据的场景中在该目录保存脏数据。如果不设置则不保存。

步骤8 单击“保存”。

----结束

查看作业

步骤1 访问Loader页面，默认显示Loader作业管理页面。

- 如果集群启用了Kerberos认证，则默认显示所有当前用户创建的作业，不支持显示其他用户的作业。
- 如果集群未启用Kerberos认证，则显示集群中全部的作业。

步骤2 在“Sqoop作业”中输入指定作业的名称或连接类型，可以筛选该作业。

步骤3 单击“刷新列表”，可以获取作业的最新状态。

----结束

编辑作业

步骤1 访问Loader页面，默认显示Loader作业管理页面。

步骤2 单击指定作业的名称，进入编辑页面。

步骤3 根据业务需要，修改作业配置参数。

步骤4 单击“保存”。

说明

左侧导航栏支持作业的基本操作，包含“运行”、“复制”、“删除”、“激活”、“历史记录”和“显示作业JSON定义”。

----结束

删除作业

步骤1 访问Loader页面。

步骤2 在指定作业所在行，单击。

您还可以勾选一个或多个作业，单击作业列表右上方的“删除作业”。

步骤3 在弹出的对话框窗口，单击“是，将其删除”。

如果某个Loader作业正处于“运行中”的状态，则无法删除作业。

----结束

12.17.8 准备 MySQL 数据库连接的驱动

操作场景

Loader作为批量数据导出的组件，可以通过关系型数据库导入、导出数据。

前提条件

已准备业务数据。

操作步骤

MRS 3.x之前版本：

步骤1 从MySQL官网下载MySQL jdbc驱动程序“mysql-connector-java-5.1.21.jar”，具体MySQL jdbc驱动程序选择参见下表。

表 12-293 版本信息

jdbc驱动程序版本	MySQL版本
Connector/J 5.1	MySQL 4.1、MySQL 5.0、MySQL 5.1、MySQL 6.0 alpha
Connector/J 5.0	MySQL 4.1、MySQL 5.0 servers、distributed transaction (XA)
Connector/J 3.1	MySQL 4.1、MySQL 5.0 servers、MySQL 5.0 except distributed transaction (XA)
Connector/J 3.0	MySQL 3.x、MySQL 4.1

步骤2 将“mysql-connector-java-5.1.21.jar”上传至MRS master 主备节点loader安装目录

- 针对MRS 3.x之前版本，上传至“/opt/Bigdata/MRS_XXX/install/FusionInsight-Sqoop-1.99.7/FusionInsight-Sqoop-1.99.7/server/jdbc/”

  其中“XXX”为MRS版本号，请根据实际情况修改。

步骤3 修改“mysql-connector-java-5.1.21.jar”包属主为“omm:wheel”。

步骤4 修改配置文件“jdbc.properties”。

将“MYSQL”的键值修改为上传的jdbc驱动包名“mysql-connector-java-5.1.21.jar”，例如：MYSQL=mysql-connector-java-5.1.21.jar。

步骤5 重启Loader服务。

----结束

MRS 3.x及之后版本：

修改关系型数据库对应的驱动jar包文件权限。

步骤1 登录Loader服务的主备管理节点，获取关系型数据库对应的驱动jar包保存在Loader服务主备节点的lib路径：“${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Sqoop-1.99.3/FusionInsight-Sqoop-1.99.3/server/webapps/loader/WEB-INF/ext-lib”。

说明

此处版本号8.1.0.1为示例，具体以实际环境的版本号为准。

步骤2 使用**root**用户在Loader服务主备节点分别执行以下命令修改权限：

cd ${BIGDATA_HOME}/FusionInsight_Porter_8.1.0.1/install/FusionInsight-Sqoop-1.99.3/FusionInsight-Sqoop-1.99.3/server/webapps/loader/WEB-INF/ext-lib

chown omm:wheel *jar包文件名*

chmod 600 *jar包文件名*

步骤3 登录FusionInsight Manager系统，选择“集群 > *待操作集群名称* > 服务 > Loader > 更多 > 重启服务”输入管理员密码重启Loader服务。

----结束

12.17.9 Loader 日志介绍

日志描述

日志存储路径：Loader相关日志的默认存储路径为“/var/log/Bigdata/loader/日志分类”。

- runlog：“/var/log/Bigdata/loader/runlog”（运行日志）
- scriptlog：“/var/log/Bigdata/loader/scriptlog/”（脚本的执行日志）
- catalina：“/var/log/Bigdata/loader/catalina”（tomcat的启停日志）
- audit：“/var/log/Bigdata/loader/audit”（审计日志）

日志归档规则：

Loader的运行日志和审计日志，启动了自动压缩归档功能，默认情况下，当日志大小超过10MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件，压缩文件保留个数可以在Manager界面中配置。

表 12-294 Loader 日志列表

日志类型	日志文件名	描述
运行日志	loader.log	Loader运行日志，记录Loader系统运行时候所产生的大部分日志。
	loader-omm-***-pid***-gc.log.*.current	Loader进程gc日志
	sqoopInstanceCheck.log	Loader实例健康检查日志
审计日志	default.audit	Loader操作审计日志（例如：作业的增删改查、用户的登录）。
tomcat日志	catalina.out	tomcat的运行日志
	catalina. <yyyy-mm-dd >.log	tomcat的运行日志
	host-manager. <yyyy-mm-dd >.log	tomcat的运行日志
	localhost_access_log. <yyyy-mm-dd >.txt	tomcat的运行日志
	manager <yyyy-mm-dd >.log	tomcat的运行日志
	localhost. <yyyy-mm-dd >.log	tomcat的运行日志
脚本日志	postInstall.log	Loader安装脚本日志。 执行loader安装脚本（postInstall.sh）时产生的日志。
	preStart.log	Loader服务的预启动脚本日志。Loader服务启动时，需要先执行一系列的准备操作（preStart.sh），例如生成keytab文件等，该日志正是记录了这些操作信息。
	loader_ctl.log	Loader执行服务启停脚本（sqoop.sh）的日志。

日志级别

Loader中提供了如表12-295所示的日志级别，日志级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-295 日志级别

级别	描述
ERROR	ERROR表示错误日志输出。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示系统及各事件正常运行状态信息。
DEBUG	DEBUG表示系统及系统调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 请参考**修改集群服务配置参数**，进入Loader的“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后即生效，不需要重启服务。

----结束

日志格式

Loader的日志格式如下所示：

表 12-296 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2015-06-29 14:54:35,553 | INFO | [localhost-startStop-1] | ConnectionRequestHandler initialized | org.apache.sqoop.handler.ConnectionRequestHandler.<init>(ConnectionRequestHandler.java:100)

日志类型	格式	示例
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|default|<log中的message>|<日志事件的发生位置>	2015-06-29 15:35:40,969 INFO default: UserName=admin, UserIP=10.52.0.111, Time=2015-06-29 15:35:40,969, Operation=submit, Resource=submission@21, Result=Failure, Detail={[reason:GET_SFTP_SESSION_FAILED:Failed to get sftp session - 10.162.0.35 (caused by: Auth cancel)]; [config:null]}

12.17.10 样例：通过 Loader 将数据从 OBS 导入 HDFS

操作场景

用户需要将大量数据从集群外导入集群内的时候，可以选择从OBS导入到HDFS的方式。

前提条件

- 已准备业务数据。
- 已创建分析集群。

操作步骤

步骤1 将业务数据上传到用户的OBS文件系统。

步骤2 获取用户的AK/SK信息，然后创建一个OBS连接和一个HDFS连接。

具体可参见**Loader连接配置说明**。

步骤3 访问Loader页面。

如果是启用了Kerberos认证的分析集群，可参见**访问Hue的WebUI**。

步骤4 单击“新建作业”。

步骤5 在“基本信息”填写参数。

1. 在“名称”填写一个作业的名称。例如“obs2hdfs”。
2. 在“源连接”选择已创建的OBS连接。
3. “目的连接”选择已创建的HDFS连接。

步骤6 在“自”填写源连接参数。

1. 在“桶名”填写业务数据所保存的OBS文件系统名称。

2. 在“源目录或文件”填写业务数据在文件系统的具体位置。
   如果是单个文件，需要填写包含文件名的完整路径。如果是目录，填写目录的完整路径
3. “文件格式”填写业务数据文件的类型。

可参见**obs-connector**。

步骤7 在“至”填写目的连接参数。

1. 在“定入目录”填写业务数据在HDFS要保存的目录名称。
   如果是启用Kerberos认证的集群，当前访问Loader的用户对保存数据的目录需要有写入权限。
2. 在“文件格式”填写业务数据文件的类型。
   需要与**步骤6.3**的类型对应。
3. 在“压缩格式”填写一种压缩的算法。例如选择不压缩“NONE”。
4. 在“是否覆盖”选择已有文件的处理方式，选择“True”。
5. 单击“显示高级属性”，在“换行符”填写业务数据保存时，系统填充的换行字符。
6. 在“字段分割符”填写业务数据保存时，系统填充的分割字符。

可参见**hdfs-connector**。

步骤8 在“任务配置”填写作业的运行参数。

1. 在“抽取并发数”填写map任务的个数。
2. 在“加载(写入)并发数”填写reduce任务的个数。
   目的连接为HDFS连接时，不显示“加载(写入)并发数”参数。
3. “单个分片的最大错误记录数”填写错误记录阈值。
4. 在“脏数据目录”填写一个脏数据的保存位置，例如“/user/sqoop/obs2hdfs-dd”。

步骤9 单击“保存并运行”。

在“管理作业界面”，查看作业运行结果。可以单击“刷新列表”获取作业的最新状态。

----结束

12.17.11 Loader 常见问题

12.17.11.1 IE 10&IE 11 浏览器无法保存数据

问题

通过IE 10&IE 11浏览器访问Loader界面，提交数据后，会报错。

回答

- 现象
  保存提交数据，出现类似报错：Invalid query parameter jobgroup id. cause: [jobgroup]。

- 原因

  IE 11浏览器的某些版本在接收到HTTP 307响应时，会将POST请求转化为GET请求，从而使得POST数据无法下发到服务端。

- 解决建议

  使用Google Chrome浏览器。

12.17.11.2 将 Oracle 数据库中的数据导入 HDFS 时各连接器的区别

问题

使用Loader将Oracle数据库中的数据导入到HDFS中时，可选择的连接器有generic-jdbc-connector、oracle-connector、oracle-partition-connector三种，要怎么选？有什么区别？

答案

- generic-jdbc-connector

  使用JDBC方式从Oracle数据库读取数据，适用于支持JDBC的数据库。

  在这种方式下，Loader加载数据的性能受限于分区列的数据分布是否均匀。当分区列的数据偏斜（数据集中在一个或者几个值）时，个别Map需要处理绝大部分数据，进而导致索引失效，造成SQL查询性能急剧下降。

  generic-jdbc-connector支持视图的导入导出，而oracle-partition-connector和oracle-connector暂不支持，因此导入视图只能选择该连接器。

- oracle-partition-connector和oracle-connector

  这两种连接器都支持按照Oracle的ROWID进行分区（oracle-partition-connector是自研，oracle-connector是社区开源版本），二者的性能较为接近。

  oracle-connector需要的系统表权限较多，下面是各自需要的系统表，需要赋予读权限。

  - oracle-connector：dba_tab_partitions、dba_constraints、dba_tables、dba_segments、v$version、dba_objects、v$instance、SYS_CONTEXT函数、dba_extents、dba_tab_subpartitions。
  - oracle-partition-connector：DBA_OBJECTS、DBA_EXTENTS。

  相比于generic-jdbc-connector，oracle-partition-connector和oracle-connector具有以下优点：

  a. 负载均匀，数据分片的个数和范围与源表的数据无关，而是由源表的存储结构（数据块）确定，颗粒度可以达到“每个数据块一个分区”。
  b. 性能稳定，完全消除“数据偏斜”和“绑定变量窥探”导致的“索引失效”。
  c. 查询速度快，数据分片的查询速度比用索引快。
  d. 水平扩展性好，如果数据量越大，产生的分片就越多，所以只要增加任务的并发数，就可以获得较理想的性能；反之，减少任务并发数，就可以节省资源。
  e. 简化数据分片逻辑，不需要考虑“精度丢失”、“类型兼容”和“绑定变量”等问题。
  f. 易用性得到增强，用户不需要专门为Loader创建分区列、分区表。

12.18 使用 Mapreduce

12.18.1 配置日志归档和清理机制

配置场景

执行一个MapReduce应用会产生两种类型日志文件：作业日志和任务日志。

- 作业日志由MRApplicationMaster产生，详细记录了作业启动时间、运行时间，每个任务启动时间、运行时间、Counter值等信息。此日志内容被HistoryServer解析以后用于查看作业执行的详细信息。
- 任务日志记录了每个运行在Container中的任务输出的日志信息。默认情况下，任务日志只会存放在各NodeManager的本地磁盘上。打开日志聚合功能后，NodeManager会在作业运行完成后将本地的任务日志进行合并，写入到HDFS中。

由于MapReduce的作业日志和任务日志（聚合功能开启的情况下）都保存在HDFS上。对于计算任务量大的集群，如果不进行合理的配置对日志文件进行定期归档和删除，日志文件将占用HDFS大量内存空间，增加集群负载。

日志归档是通过Hadoop Archives功能实现的，Hadoop Archives启动的并行归档任务数（Map数）与待归档的日志文件总大小有关。计算公式为：并行归档任务数=待归档的日志文件总大小/归档文件大小。

配置描述

进入Mapreduce服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**章节。

在搜索框中输入参数名称。同时需要在Mapreduce客户端节点的“*客户端安装目录*/HDFS/hadoop/etc/hadoop/”路径下的“mapred-site.xml”配置文件中进行如下配置。

表 12-297 参数说明

参数	描述	默认值
mapreduce.jobhistory.cleaner.enable	是否开启作业日志文件清理功能。	true
mapreduce.jobhistory.cleaner.interval-ms	作业日志文件清理启动周期。只有保留时间比“mapreduce.jobhistory.max-age-ms”更长的日志文件才会被清除。	86400000（1天）
mapreduce.jobhistory.max-age-ms	比此项设置的毫秒数保留时间更长的作业日志文件将被清理。	1296000000（15天）

您可以在ResourceManager、NodeManager、MapReduce的JobHistoryServer各节点的“yarn-site.xml”配置文件中进行如下配置，其中yarn.nodemanager.remote-app-log-dir和yarn.nodemanager.remote-app-log-archive-dir这两个参数还需要在YARN的

客户端进行配置，且在ResourceManager、NodeManager和MapReduce HistoryServer各节点的配置与在YARN的客户端的配置必须一致。

表 12-298 参数说明

参数	描述	默认值
yarn.nodemanager.remote-app-log-dir	设置Mapreduce任务日志在HDFS上的聚合路径。	/tmp/logs
yarn.nodemanager.remote-app-log-archive-dir	设置Mapreduce任务日志在HDFS上的归档路径。	/tmp/archived
yarn.log-aggregation.archive.files.minimum	设置Mapreduce任务日志归档最小文件数。当“yarn.nodemanager.remote-app-log-dir”文件夹下文件数大于等于该设置的值时归档任务启动。 该参数适用于MRS 3.x版本。	5000
yarn.log-aggregation.archive-check-interval-seconds	设置Mapreduce任务日志归档任务启动周期（秒）。只有日志文件数达到“yarn.log-aggregation.archive.files.minimum”设置值时日志文件才会被归档。周期设置为“0”或“-1”时归档功能禁用。 该参数适用于MRS 3.x版本。	-1
yarn.log-aggregation.retain-seconds	设置Mapreduce任务日志在HDFS上的保留时间。设置为“-1”时日志文件永久保存。	1296000
yarn.log-aggregation.retain-check-interval-seconds	设置Mapreduce任务日志清理任务的检查周期（秒）。设置为“-1”时检查周期为日志保留时间的十分之一。	86400

12.18.2 降低客户端应用的失败率

配置场景

当网络不稳定或者集群IO、CPU负载过高的情况下，通过调整如下参数值，降低客户端应用的失败率，保证应用的正常运行。

配置描述

在客户端的“mapred-site.xml”配置文件中调整如下参数。

说明

“mapred-site.xml”配置文件在客户端安装路径的conf目录下，例如“/opt/client/Yarn/config”。

表 12-299 参数说明

参数	描述	默认值
mapreduce.reduce.shuffle.max-host-failures	MR任务在reduce过程中读取远端shuffle数据允许失败的次数。当设置次数大于5时，可以降低客户端应用的失败率。该参数适用于MRS 3.x版本。	5
mapreduce.client.submit.file.replication	MR任务在运行时依赖的相关job文件在HDFS上的备份。当备份数大于10时，可以降低客户端应用的失败率。	10

12.18.3 将 MR 任务从 Windows 上提交到 Linux 上运行

配置场景

用户将MapReduce任务从Windows上提交到Linux上运行，则“mapreduce.app-submission.cross-platform”参数值需配置为“true”。若集群无此参数，或参数值为“false”，则表示集群不支持此功能，需要按照如下操作增加该参数或修改参数值进行开启。

说明

本章节操作适用于MRS 3.x及之后版本。

配置描述

在客户端的“mapred-site.xml”配置文件中进行如下配置。“mapred-site.xml”配置文件在客户端安装路径的config目录下，例如“/opt/client/Yarn/config”。

表 12-300 参数说明

参数	描述	默认值
mapreduce.app-submission.cross-platform	支持在Windows上提交到Linux上运行MR任务的配置项。当该参数的值设为“true”时，表示支持。当该参数的值设为“false”时，表示不支持。	true

12.18.4 配置使用分布式缓存

配置场景

说明

本章节操作适用于MRS 3.x及之后版本。

分布式缓存在两种情况下非常有用。

滚动升级

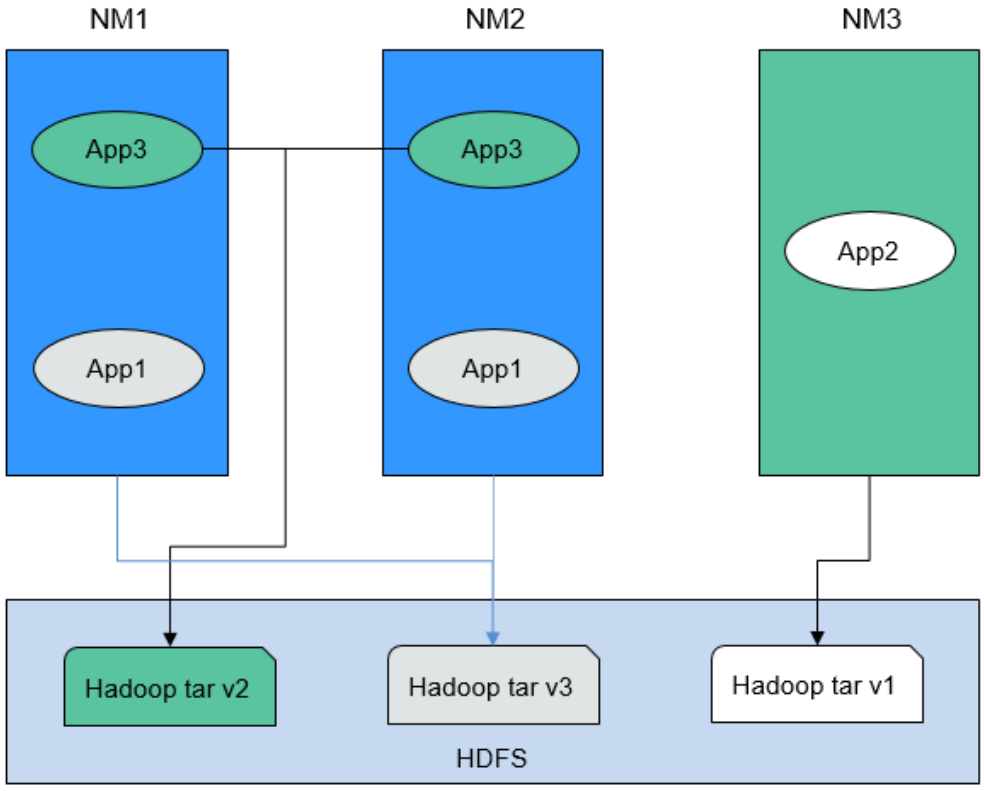
在升级过程中，应用程序必须保持文字内容（jar文件或配置文件）不变。而这些内容并非基于当前版本的YARN，而是要基于其提交时的版本。这是一个具有挑战性的问题。一般情况下，应用程序（例如MapReduce、Hive、Tez等）需要进行完整的本地安装，将库安装至所有的集群机器（客户端及服务器端机器）中。当集群内开始进行滚动升级或降级时，本地安装的库的版本必然会在应用运行过程时发生改变。在滚动升级过程中，首先只会对少数NodeManager进行升级，这些NodeManager会获得新版本的软件。这导致了行为的不一致，并可能发生运行时错误。

同时存在多个YARN版本

集群管理员可能会在一个集群内运行使用多个版本YARN及Hadoop jars的任务。这在当前很难实现，因为jars已被本地化且只有一个版本。

MapReduce应用框架可以通过分布式缓存进行部署，且无需依赖安装中复制的静态版本。因此，可以在HDFS中存放多版本的Hadoop，并通过配置“mapred-site.xml”文件指定任务默认使用的版本。只需设置适当的配置属性，用户就可以运行不同版本的MapReduce，而无需使用部署在集群中的版本。

图 12-35 具有多个版本 NodeManagers 及 Applications 的集群



在图12-35中：可以看出，应用程序可以使用HDFS中的Hadoop jars，而无需使用本地版本。因此在滚动升级中，即使NodeManager已经升级，应用程序仍然可以运行旧版本的Hadoop。

配置描述

步骤1 首先，需要将指定版本的MapReduce tar包存放至HDFS中应用程序可以访问的目录下，如下所示：

$HADOOP_HOME/**bin/hdfs dfs -put** *hadoop-x.tar.gz /mapred/framework/*

步骤2 根据表12-301，对“mapred-site.xml”文件中的参数进行设置。

表 12-301 分布式缓存相关参数

参数	说明	默认值
mapreduce.application.framework.path	此参数值为指向存档位置的URL。 **说明** 如果对URL片段标示名称进行如下指定，该属性还可以为存档创建别名。作为示例，这里将别名设为了mr-framework。 <property> <name>mapreduce.application.framework.path</name> <value>hdfs:/mapred/framework/hadoop-x.tar.gz#mr-framework</value> </property>	NA
mapreduce.application.classpath	设定属性mapreduce.application.classpath，使其可以包含类目录中相关的MR jars。 **说明** 例如，此处利用在框架路径中使用过的别名“mr-framework”对目录进行匹配。 <property> <name>mapreduce.application.classpath</name> <value>$PWD/mr-framework/hadoop/share/hadoop/mapreduce/*:$PWD/mr-framework/hadoop/share/hadoop/mapreduce/lib/*:$PWD/mr-framework/hadoop/share/hadoop/common/*:$PWD/mr-framework/hadoop/share/hadoop/common/lib/*:$PWD/mr-framework/hadoop/share/hadoop/yarn/*:$PWD/mr-framework/hadoop/share/hadoop/yarn/lib/*:$PWD/mr-framework/hadoop/share/hadoop/hdfs/*:$PWD/mr-framework/hadoop/share/hadoop/hdfs/lib/*:/etc/hadoop/conf/secure</value></property>	NA

可以将多个版本的MR tarball上传至HDFS。不同的“mapred-site.xml”文件可以指向不同的位置。用户在此之后可以针对特定的“mapred-site.xml”文件运行任务。以下是一个针对x版本的MR tarball运行MR任务的例子：

hadoop jar *share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar pi* **-conf etc/hadoop-x/mapred-site.xml** *10 10*

----结束

12.18.5 配置 MapReduce shuffle address

配置场景

当MapReduce shuffle服务启动时，它尝试基于localhost绑定IP。如果需要MapReduce shuffle服务去连接特定IP，那么没有可用的配置。下面的描述允许您配置连接到特定的IP。

配置描述

当需要MapReduce shuffle服务绑定特定IP时，需要在NodeManager实例所在节点的配置文件“mapred-site.xml”中设置下面的参数。

表 12-302 参数描述

参数	描述	默认值
mapreduce.shuffle.address	指定地址来运行shuffle服务，格式是IP:PORT，参数的默认值为空。当参数值为空时，将绑定localhost，默认端口为13562。 **说明** 如果涉及到的PORT值和配置的mapreduce.shuffle.port值不一样时，mapreduce.shuffle.port将不会生效。	-

12.18.6 配置集群管理员列表

配置场景

该功能主要用于指定MapReduce集群管理员。

其中，管理员列表由参数“mapreduce.cluster.administrators”指定，集群管理员admin具有所有可以操作的权限。

配置描述

进入Mapreduce服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。

表 12-303 参数描述

参数	描述	默认值
mapreduce.cluster.acls.enabled	是否开启对Job History Server 权限控制的开关。	true
mapreduce.cluster.administrators	用于指定MapReduce集群的管理员列表，可以配置用户和用户组，用户或者用户组之间用逗号间隔，用户和用户组之间用空格间隔，举例：userA,userB groupA,groupB。当配置为*时表示所有用户或用户组。	MRS 3.x之前版本：mapred MRS 3.x及之后版本： mapred supergroup,System_administrator_186

12.18.7 MapReduce 日志介绍

日志描述

日志默认存储路径：

- JobhistoryServer：“/var/log/Bigdata/mapreduce/jobhistory”（运行日志），“/var/log/Bigdata/audit/mapreduce/jobhistory”（审计日志）
- Container：“/srv/BigData/hadoop/data1/nm/containerlogs/application_${appid}/container_{$contid}”

说明

运行中的任务日志存储在以上路径中，运行结束后会基于YARN的配置是否汇聚到HDFS目录中，详情请参见**Yarn常用参数**。

日志归档规则：

MapReduce的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过50MB的时候，会自动压缩，压缩后的日志文件名规则为：“*<原有日志名>*-<yyyy-mm-dd_hh-mm-ss>.[*编号*].log.zip”。最多保留最近的100个压缩文件，压缩文件保留个数可以在参数配置界面中配置。

在MapReduce服务中，JobhistoryServer会定时去清理HDFS上存储的旧的日志文件（默认目录为HDFS文件系统中的“/mr-history/done”），具体清理的时间间隔参数配置为mapreduce.jobhistory.max-age-ms，默认值为1296000000，即15天。

表 12-304 MR 日志列表

日志类型	日志文件名	描述
运行日志	jhs-daemon-start-stop.log	守护进程（Daemon）的启动日志。
	hadoop-<SSH_USER>-jhshadaemon-<hostname>.log	守护进程（Daemon）的运行日志。
	hadoop-<SSH_USER>-<process_name>-<hostname>.out	MR运行环境信息日志。
	historyserver-<SSH_USER>-<DATE>-<PID>-gc.log	MR服务垃圾回收日志。
	jhs-haCheck.log	MR实例主备状态检查日志。
	yarn-start-stop.log	MR服务启停操作日志。
	yarn-prestart.log	MR服务启动前集群操作的记录日志。
	yarn-postinstall.log	MR服务安装后启动前的工作日志。
	yarn-cleanup.log	MR服务卸载时候的清理日志。
	mapred-service-check.log	MR服务健康状态检测日志。
	container_{$contid}	Container日志。

日志类型	日志文件名	描述
	hadoop-<SSH_USER>-<process_name>-<hostname>.log	MR运行日志。
	mapred-switch-jhs.log	MR主备倒换日志。
	env.log	实例启停前的环境信息日志。
审计日志	mapred-audit-jobhistory.log	MR操作审计日志。
	SecurityAuth.audit	MR安全审计日志。

日志级别

MapReduce中提供了如表12-305所示的日志级别。其中日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-305 日志级别

级别	描述
FATAL	FATAL表示当前事件处理存在严重错误信息。
ERROR	ERROR表示当前事件处理存在错误信息。
WARN	WARN表示当前事件处理存在异常告警信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 进入MapReduce服务参数“全部配置”界面，具体操作请参考**修改集群服务配置参数**。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后立即生效，不需要重启服务。

----结束

日志格式

MapReduce日志格式如下所示：

表 12-306 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<*Log Level*>|<*产生该日志的线程名字*>|<*log中的message*>|<*日志事件的发生位置*>	2020-01-26 14:18:59,109 | INFO | main | Client environment:java.compiler=<NA> | org.apache.zookeeper.Environment.logEnv(Environment.java:100)
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<*Log Level*>|<*产生该日志的线程名字*>|<*log中的message*>|<*日志事件的发生位置*>	2020-01-26 14:24:43,605 | INFO | main-EventThread | USER=omm OPERATION=refreshAdminAcls TARGET=AdminService RESULT=SUCCESS | org.apache.hadoop.yarn.server.resourcemanager.RMAuditLogger$LogLevel $6.printLog(RMAuditLogger.java:91)

12.18.8 MapReduce 性能调优

12.18.8.1 多 CPU 内核下的调优配置

操作场景

当CPU内核数很多时，如CPU内核为磁盘数的3倍时的调优配置。

操作步骤

以下参数有如下两个配置入口：

- 服务器端配置
  进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。
- 客户端配置
  直接在客户端中修改相应的配置文件。

  说明

  - HDFS客户端配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/hdfs-site.xml。
  - Yarn客户端配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/yarn-site.xml。
  - MapReduce客户端配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/mapred-site.xml。

表 12-307 多 CPU 内核设置

配置	描述	参数	默认值	Server/Client	影响	备注
节点容器槽位数	如下配置组合决定了每节点任务(map、reduce)的并发数。 ● “yarn.nodemanager.resource.memory-mb” ● “mapreduce.map.memory.mb” ● “mapreduce.reduce.memory.mb”	yarn.nodemanager.resource.memory-mb **说明** MRS 3.x之前版本：需要在MRS控制台上进行配置。 MRS 3.x及之后版本：需要在FusionInsight Manager系统进行配置。	MRS 3.x之前版本： 8192 MRS 3.x及之后版本： 16384	Server	如果所有的任务(map/reduce)需要读写数据至磁盘，多个进程将会同时访问一个磁盘。这将会导致磁盘的IO性能非常的低下。为了改善磁盘的性能，请确保客户端并发访问磁盘的数不大于3。	最大并发的container数量应该为[2.5 * Hadoop中磁盘配置数]。
		mapreduce.map.memory.mb **说明** 需要在客户端进行配置，配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/mapred-site.xml。	4096	Client		
		mapreduce.reduce.memory.mb **说明** 需要在客户端进行配置，配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/mapred-site.xml。	4096	Client		

配置	描述	参数	默认值	Server/Client	影响	备注
Map输出与压缩	Map任务所产生的输出可以在写入磁盘之前被压缩，这样可以节约磁盘空间并得到更快的写盘速度，同时可以减少至Reducer的数据传输量。需要在客户端进行配置。 ● mapreduce.map.output.compress指定了Map任务输出结果可以在网络传输前被压缩。这是一个per-job的配置。 ● mapreduce.map.output.compress.codec指定用于压缩的编解码器。	mapreduce.map.output.compress **说明** 需要在客户端进行配置，配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/mapred-site.xml。	true	Client	在这种情况下，磁盘的IO是主要瓶颈。所以可以选择一种压缩率非常高的压缩算法。	编解码器可配置为Snappy，Benchmark测试结果显示Snappy是非常平衡以及高效的编码器。
		mapreduce.map.output.compress.codec **说明** 需要在客户端进行配置，配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/mapred-site.xml。	org.apache.hadoop.io.compress.Lz4Codec	Client		

配置	描述	参数	默认值	Server/Client	影响	备注
Spills	mapreduce.map.sort.spill.percent	mapreduce.map.sort.spill.percent **说明** 需要在客户端进行配置，配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/mapred-site.xml。	0.8	Client	磁盘IO是主要瓶颈，合理配置“mapreduce.task.io.sort.mb”可以使溢出至磁盘的内容最小化。	-
数据包大小	当HDFS客户端写数据至数据节点时，数据会被累积，直到形成一个包。然后这个数据包会通过网络传输。dfs.client-write-packet-size配置项可以指定该数据包的大小。这个可以通过每个job进行指定。	dfs.client-write-packet-size **说明** 需要在客户端进行配置，配置文件路径：*客户端安装目录*/HDFS/hadoop/etc/hadoop/hdfs-site.xml。	262144	Client	数据节点从HDFS客户端接收数据包，然后将数据包里的数据单线程写入磁盘。当磁盘处于并发写入状态时，增加数据包的大小可以减少磁盘寻道时间，从而提升IO性能。	dfs.client-write-packet-size = 262144

12.18.8.2 确定 Job 基线

操作场景

确定Job基线是调优的基础，一切调优项效果的检查，都是通过和基线数据做对比来获得。

Job基线的确定有如下三个原则：

- 充分利用集群资源
- reduce阶段尽量放在一轮
- 每个task的执行时间要合理

操作步骤

- **原则一：充分利用集群资源。**

  Job运行时，会让所有的节点都有任务处理，且处于繁忙状态，这样才能保证资源充分利用，任务的并发度达到最大。可以通过调整处理的数据量大小，以及调整map和reduce个数来实现。

  Reduce个数的控制使用“mapreduce.job.reduces”。

  Map个数取决于使用了哪种InputFormat，以及待处理的数据文件是否可分割。默认的TextFileInputFormat将根据block的个数来分配map数(一个block一个map)。通过如下配置参数进行调整。

  参数入口：

  进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考修改集群服务配置参数章节。

参数	描述	默认值
mapreduce.input.fileinputformat.split.maxsize	map输入信息应被拆分成的数据块的最大大小。 由用户定义的分片大小的设置及每个文件block大小的设置，可以计算分片的大小。计算公式如下： `splitSize = Math.max(minSize, Math.min(maxSize, blockSize))` 如果maxSize设置大于blockSize，那么每个block就是一个分片，否则就会将一个block文件分隔为多个分片，如果block中剩下的一小段数据量小于splitSize，还是认为它是独立的分片。	-
mapreduce.input.fileinputformat.split.minsize	可以设置数据分片的数据最小值。	0

- **原则二：控制reduce阶段在一轮中完成。**

  避免以下两种场景：

  - 大部分的reduce在第一轮运行完后，剩下唯一一个reduce继续运行。这种情况下，这个reduce的执行时间将极大影响这个job的运行时间。因此需要将reduce个数减少。
  - 所有的map运行完后，只有个别节点有reduce在运行。这时候集群资源没有得到充分利用，需要增加reduce的个数以便每个节点都有任务处理。

- **原则三：每个task的执行时间要合理。**

  如果一个job，每个map或reduce的执行时间只有几秒钟，就意味着这个job的大部分时间都消耗在task的调度和进程启停上了，因此需要增加每个task处理的数据大小。建议一个task处理时间为1分钟。

  控制单个task处理时间的大小，可以通过如下配置来调整。

  参数入口：

  进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考修改集群服务配置参数章节。

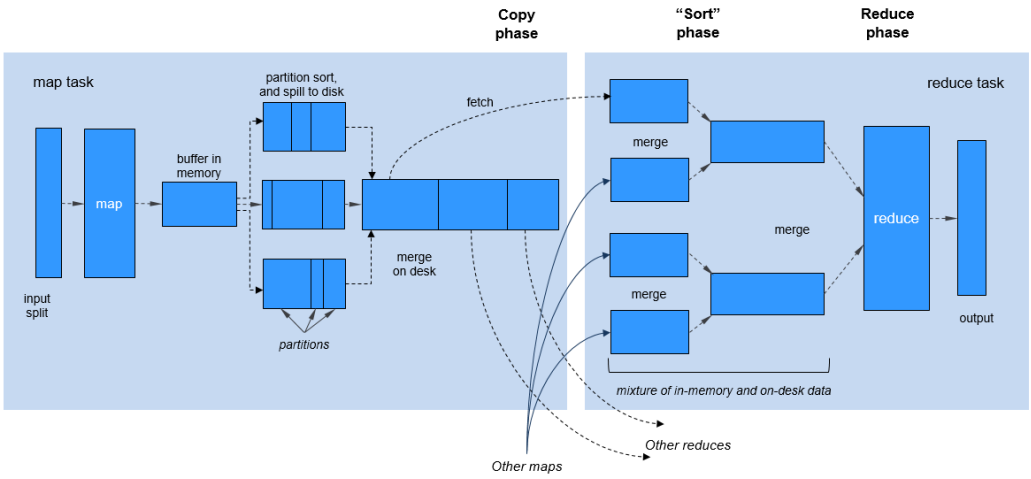
参数	描述	默认值
mapreduce.input.fileinputformat.split.maxsize	map输入信息应被拆分成的数据块的最大大小。 由用户定义的分片大小的设置及每个文件block大小的设置，可以计算分片的大小。计算公式如下： `splitSize = Math.max(minSize, Math.min(maxSize, blockSize))` 如果maxSize设置大于blockSize，那么每个block就是一个分片，否则就会将一个block文件分隔为多个分片，如果block中剩下的一小段数据量小于splitSize，还是认为它是独立的分片。	-
mapreduce.input.fileinputformat.split.minsize	可以设置数据分片的数据最小值。	0

12.18.8.3 Shuffle 调优

操作场景

Shuffle阶段是MapReduce性能的关键部分，包括了从Map task将中间数据写到磁盘一直到Reduce task拷贝数据并最终放到reduce函数的全部过程。这一块Hadoop提供了大量的调优参数。

图 12-36 Shuffle 过程



操作步骤

1. **Map阶段的调优**
   - 判断Map使用的内存大小

     判断Map分配的内存是否足够，一个简单的办法是查看运行完成的job的Counters中，对应的task是否发生过多次GC，以及GC时间占总task运行时间

之比。通常，GC时间不应超过task运行时间的10%，即GC time elapsed (ms)/CPU time spent (ms)<10%。

主要通过如下参数进行调整。

参数入口：

进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。

表 12-308 参数说明

参数	描述	默认值
mapreduce.map.memory.mb	map任务的内存限制。	4096

参数	描述	默认值
mapreduce.map.java.opts	map子任务的JVM参数。如果设置，会替代mapred.child.java.opts参数；如果未设置-Xmx，Xmx值从mapreduce.map.memory.mb*mapreduce.job.heap.memory-mb.ratio计算获取。	MRS 3.x之前版本：-Xmx2048M -Djava.net.preferIPv4Stack=true MRS 3.x及之后版本： • 集群已开启Kerberos认证：-Djava.net.preferIPv4Stack=true -Djava.net.preferIPv6Addresses=false -Djava.security.krb5.conf=*${BIGDATA_HOME}*/common/runtime/krb5.conf -Dbeetle.application.home.path=*${BIGDATA_HOME}*/common/runtime/security/config • 集群未开启Kerberos认证：-Djava.net.preferIPv4Stack=true -Djava.net.preferIPv6Addresses=false -Dbeetle.application.home.path=*${BIGDATA_HOME}*/common/runtime/security/config

建议：配置“mapreduce.map.java.opts”参数中“-Xmx”值为“mapreduce.map.memory.mb”参数值的0.8倍。

- 使用Combiner

  在Map阶段，有一个可选过程，将同一个key值的中间结果合并，叫做combiner。一般将reduce类设置为combiner即可。通过combine，一般情况下可以显著减少Map输出的中间结果，从而减少shuffle过程的网络带宽占用。可通过如下接口为一个任务设置Combiner类。

表 12-309 Combiner 设置接口

类名	接口名	描述
org.apache.hadoop.mapreduce.Job	public void setCombinerClass(Class<? extends Reducer> cls)	为Job设置一个combiner类。

2. **Copy阶段的调优**
   - 数据是否压缩

     对Map的中间结果进行压缩，当数据量大时，会显著减少网络传输的数据量，但是也因为多了压缩和解压，带来了更多的CPU消耗。因此需要做好权衡。当任务属于网络瓶颈类型时，压缩Map中间结果效果明显。针对bulkload调优，压缩中间结果后性能提升60%左右。

     配置方法：将“mapreduce.map.output.compress”参数值设置为“true”，将“mapreduce.map.output.compress.codec”参数值设置为“org.apache.hadoop.io.compress.SnappyCodec”。
3. **Merge阶段的调优**

   通过调整如下参数减少reduce写磁盘的次数。

   参数入口：

   进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。

表 12-310 参数说明

参数	描述	默认值
mapreduce.reduce.merge.inmem.threshold	内存合并进程的文件数阈值。累计文件数达到阈值时会发起内存合并及溢出到磁盘。小于等于0的值表示该阈值不生效且仅基于ramfs的内存使用情况来触发合并。	1000
mapreduce.reduce.shuffle.merge.percent	发起内存合并的使用率阈值，表示为分配给映射输出信息的内存的比例（是由mapreduce.reduce.shuffle.input.buffer.percent设置的）。	0.66
mapreduce.reduce.shuffle.input.buffer.percent	shuffle过程中分配给映射输出信息的内存占最大堆大小的比例。	0.70
mapreduce.reduce.input.buffer.percent	Reduce过程中保存映射输出信息的内存相对于最大堆大小的比例。当shuffle结束时，需保证reduce开始前内存中所有剩余的映射输出信息所使用的内存小于该阈值。	0.0

12.18.8.4 大任务的 AM 调优

操作场景

任务场景：运行的一个大任务（map总数达到了10万的规模），但是一直没有跑成功。经过查询，发现是ApplicationMaster（以下简称AM）反应缓慢，最终超时失败。

此任务的问题是，task数量变多时，AM管理的对象也线性增长，因此就需要更多的内存来管理。AM默认分配的内存堆大小是1GB。

操作步骤

通过调大如下的参数来进行AM调优。

参数入口：

在Yarn客户端的“mapred-site.xml”配置文件中调整如下参数。“mapred-site.xml”配置文件在客户端安装路径的conf目录下，例如“/opt/client/Yarn/config”。

参数	描述	默认值
yarn.app.mapreduce.am.resource.mb	该参数值必须大于下面参数的堆大小。单位：MB	1536
yarn.app.mapreduce.am.command-opts	传递到MapReduce ApplicationMaster的JVM启动参数。	MRS 3.x之前版本：-Xmx1024m -XX:CMSFullGCsBeforeCompaction=1 -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:+UseCMSCompactAtFullCollection -verbose:gc MRS 3.x及之后版本：-Xmx1024m -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -verbose:gc -Djava.security.krb5.conf=${KRB5_CONFIG} -Dhadoop.home.dir=*${BIGDATA_HOME}*/FusionInsight_HD_*xxx*/install/FusionInsight-Hadoop-*xxx*/hadoop

12.18.8.5 推测执行

操作场景

当集群规模很大时（如几百上千台节点的集群），个别机器出现软硬件故障的概率就变大了，并且会因此延长整个任务的执行时间（跑完的任务都在等出问题的机器跑结束）。推测执行通过将一个task分给多台机器跑，取先运行完的那个，会很好的解决这个问题。对于小集群，可以将这个功能关闭。

操作步骤

参数入口：

进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。

参数	描述	默认值
mapreduce.map.speculative	设置是否并行执行某些映射任务的多个实例。true表示开启。	false
mapreduce.reduce.speculative	设置是否并行执行某些reduce任务的多个实例。true表示开启。	false

12.18.8.6 通过“Slow Start”调优

操作场景

Slow Start特性指定Map任务完成度为多少时Reduce任务可以启动，过早启动Reduce任务会导致资源占用，影响任务运行效率，但适当的提早启动Reduce任务会提高Shuffle阶段的资源利用率，提高任务运行效率。例如：某集群可启动10个Map任务，MapReduce作业共15个Map任务，那么在一轮Map任务执行完成后只剩5个Map任务，集群还有剩余资源，在这种场景下，配置Slow Start参数值小于1，比如0.8，则Reduce就可以利用集群剩余资源。

操作步骤

参数入口：

进入Mapreduce服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。

参数	描述	默认值
mapreduce.job.reduce.slowstart.completedmaps	为job安排reduce前应完成的映射数的分数形式。默认100%的Map跑完后开始起Reduce。	1.0

12.18.8.7 MR job commit 阶段优化

操作场景

默认情况下，如果一个MR任务会产生大量的输出结果文件，那么该job在最后的commit阶段会耗费较长的时间将每个task的临时输出结果commit到最终的结果输出目录。特别是在大集群中，大Job的commit过程会严重影响任务的性能表现。

针对以上情况，可以通过将以下参数“mapreduce.fileoutputcommitter.algorithm.version”配置为“2”，来提升MR Job commit阶段的性能。

操作步骤

参数入口：

进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。具体操作请参考**修改集群服务配置参数**章节。

表 12-311 参数说明

参数	描述	默认值
mapreduce.fileoutputcommitter.algorithm.version	用于指定Job的最终输出文件提交的算法版本，取值为“1”或“2”。 **说明** 版本2为建议的优化算法版本。该算法通过让任务直接将每个task的输出结果提交到最终的结果输出目录，从而减少大作业的输出提交时间。	2

12.18.9 MapReduce 常见问题

12.18.9.1 ResourceManager 进行主备切换后，任务中断后运行时间过长

问题

在MapReduce任务运行过程中，ResourceManager发生主备切换，切换完成后，MapReduce任务继续执行，此时任务的运行时间过长。

回答

因为ResourceManager HA已启用，但是Work-preserving RM restart功能未启用。

如果Work-preserving RM restart功能未启用，ResouceManager切换时container会被kill，然后导致Application Master超时。Work-preserving RM restart功能介绍请参见：**http://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/ResourceManagerRestart.html**

可以通过如下方式解决此问题：

设置如下参数启用Work-preserving RM restart功能。

“yarn.resourcemanager.work-preserving-recovery.enabled”=“true”

12.18.9.2 MapReduce 任务长时间无进展

问题

MapReduce任务长时间无进展。

回答

一般是因为内存太少导致的。当内存较小时，任务中拷贝map输出的时间将显著增加。

为了减少等待时间，您可以适当增加堆内存空间。

任务的配置可根据mapper的数量和各mapper的数据大小来进行优化。根据输入数据的大小，优化如下参数：

- “mapreduce.reduce.memory.mb”
- “mapreduce.reduce.java.opts”

例如：如果10个mapper的数据大小为5GB，那么理想的堆内存是1.5GB。随着数据大小的增加而增加堆内存大小。

12.18.9.3 运行任务时，客户端不可用

问题

当运行任务时，将MR ApplicationMaster或ResourceManager移动为D状态，为什么此时客户端会不可用?

回答

当运行任务时，将MR ApplicationMaster或ResourceManager移动为D状态（不间断睡眠状态）或T状态（停止状态），客户端会等待返回任务运行的状态，由于AM无返回，客户端会一直处于等待状态。

为避免出现上述场景，使用“core-site.xml”中的“ipc.client.rpc.timeout”配置项设置客户端超时时间。

该参数的参数值为毫秒。默认值为0，表示无超时。客户端超时的取值范围可以为0～2147483647毫秒。

说明

- 如果Hadoop进程已处于D状态，重启该进程所处的节点。
- “core-site.xml”配置文件在客户端安装路径的conf目录下，例如“/opt/hadoopClient/Yarn/config”。

12.18.9.4 在缓存中找不到 HDFS_DELEGATION_TOKEN

问题

安全模式下，为什么在缓存中找不到HDFS_DELEGATION_TOKEN?

回答

在MapReduce中，默认情况下，任务完成之后，HDFS_DELEGATION_TOKEN将会被删除。因此如果在下一个任务中再次使用HDFS_DELEGATION_TOKEN，缓存中将会找不到HDFS_DELEGATION_TOKEN。

为了能够在随后的工作中再次使用同一个Token，为MapReduce任务配置参数。当参数为false时，用户能够再次使用同一个Token。

```
jobConf.setBoolean("mapreduce.job.complete.cancel.delegation.tokens", false);
```

12.18.9.5 如何在提交 MapReduce 任务时设置任务优先级

问题

如何在提交MapReduce任务时设置任务优先级?

回答

当您在客户端提交MapReduce任务时，可以在命令行中增加“-Dmapreduce.job.priority=<priority>”参数来设置任务优先级。格式如下：

yarn jar *<jar> [mainClass]* ***-Dmapreduce.job.priority=****<priority> [path1] [path2]*

命令行中参数含义为：

- <jar>：指定需要运行的jar包名称。
- [mainClass]：指jar包应用工程中的类得main方法。
- <priority>：指定任务的优先级，其取值可为：VERY_HIGH、HIGH、NORMAL、LOW、VERY_LOW。
- [path1]：指数据输入路径。
- [path2]：指数据输出路径。

例如，将“/opt/client/HDFS/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples*.jar”包设置为高优先级任务。

yarn jar /opt/client/HDFS/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples*.jar wordcount -Dmapreduce.job.priority=VERY_HIGH /DATA.txt /out/

12.18.9.6 MapReduce 任务运行失败，ApplicationMaster 出现物理内存溢出异常

问题

HBase bulkload任务有210000个map和10000个reduce，MapReduce任务运行失败，ApplicationMaster出现物理内存溢出异常。

```
For more detailed output, check the application tracking page:https://bigdata-55:8090/cluster/app/application_1449841777199_0003
Then click on links to logs of each attempt.
Diagnostics: Container [pid=21557,containerID=container_1449841777199_0003_02_000001] is running beyond physical memory limits
Current usage: 1.0 GB of 1 GB physical memory used; 3.6 GB of 5 GB virtual memory used. Killing container.
Dump of the process-tree for container_1449841777199_0003_02_000001 :
|- PID PPID PGRPID SESSID CMD_NAME USER_MODE_TIME(MILLIS) SYSTEM_TIME(MILLIS) VMEM_USAGE(BYTES) RSSMEM_USAGE(PAGES) FULL_CMD_LINE
|- 21584 21557 21557 21557 (java) 12342 1627 3871748096 271331 ${BIGDATA_HOME}/jdk1.8.0_51//bin/java
-Djava.io.tmpdir=/srv/BigData/hadoop/data1/nm/localdir/usercache/hbase/appcache/application_1449841777199_0003/container_1449841777199_0003_02_000001/tmp -Dlog4j.configuration=container-log4j.properties
-Dyarn.app.container.log.dir=/srv/BigData/hadoop/data1/nm/containerlogs/application_1449841777199_0003/container_1449841777199_0003_02_000001 -Dyarn.app.container.log.filesize=0 -Dhadoop.root.logger=INFO,CLA
-Dhadoop.root.logfile=syslog -Xmx784m org.apache.hadoop.mapreduce.v2.app.MRAppMaster
|- 21557 21547 21557 21557 (bash) 0 0 13074432 368 /bin/bash -c /opt/xxx/Bigdata/jdk1.8.0_51//bin/java -Djava.io.tmpdir=/srv/BigData/hadoop/data1/nm/localdir/usercache/hbase/appcache/
```

```
application_1449841777199_0003/container_1449841777199_0003_02_000001/tmp -
Dlog4j.configuration=container-log4j.properties
-Dyarn.app.container.log.dir=/srv/BigData/hadoop/data1/nm/containerlogs/
application_1449841777199_0003/container_1449841777199_0003_02_000001 -
Dyarn.app.container.log.filesize=0 -Dhadoop.root.logger=INFO,CLA
-Dhadoop.root.logfile=syslog -Xmx784m org.apache.hadoop.mapreduce.v2.app.MRAppMaster 1>/srv/
BigData/hadoop/data1/nm/containerlogs/application_1449841777199_0003/
container_1449841777199_0003_02_000001/stdout
2>/srv/BigData/hadoop/data1/nm/containerlogs/application_1449841777199_0003/
container_1449841777199_0003_02_000001/stderr
Container killed on request. Exit code is 143
Container exited with a non-zero exit code 143
Failing this attempt. Failing the application.
```

回答

这是性能规格的问题，MapReduce任务运行失败的根本原因是由于ApplicationMaster的内存溢出导致的，即物理内存溢出导致被NodeManager kill。

解决方案：

将ApplicationMaster的内存配置调大，在客户端“mapred-site.xml”配置文件中优化如下参数：

- “yarn.app.mapreduce.am.resource.mb”
- “yarn.app.mapreduce.am.command-opts”，该参数中-Xmx值建议为0.8*“yarn.app.mapreduce.am.resource.mb”

参考规格：

ApplicationMaster配置如下时，可以同时支持并发Container数为2.4万个。

- “yarn.app.mapreduce.am.resource.mb”=2048
- “yarn.app.mapreduce.am.command-opts”该参数中-Xmx=1638m

12.18.9.7 MapReduce JobHistoryServer 服务地址变更后，为什么运行完的 MapReduce 作业信息无法通过 ResourceManager Web UI 页面的 Tracking URL 打开

问题

MapReduce JobHistoryServer服务地址变更后，为什么运行完的MapReduce作业无法通过ResourceManager Web UI页面打开?

回答

JobHistoryServer地址（mapreduce.jobhistory.address / mapreduce.jobhistory.webapp.<https.>address）是MapReduce参数，MapReduce客户端提交作业时，会将此地址随任务一起提交给ResourceManager。ResourceManager在作业完成后，将此参数作为查看作业历史信息的跳转地址保存在RMStateStore中。

JobHistoryServer服务地址变更后，需要将新的服务地址及时更新到MapReduce客户端配置文件中，否则，新运行的作业在查看作业历史信息时，仍然会指向原JobHistoryServer地址，导致无法正常跳转到作业历史信息页面。服务地址变更前运行的MapReduce作业，由于其跳转信息已经保存在RMStateStore中，无法变更，因此从

ResourceManager Web UI页面是无法进行正常跳转的，但可以直接访问新的JobHistoryServer服务地址进行查找，作业信息不会丢失。

12.18.9.8 多个 NameService 环境下，运行 MapReduce 任务失败

问题

多个NameService环境下，运行使用viewFS功能的MapReduce或YARN任务失败。

回答

当使用viewFS时，只有在viewFS中挂载的目录才能被访问到。所以最可能的原因是配置的路径没有在viewFS的挂载点上。例如：

```
<property>
<name>fs.defaultFS</name>
<value>viewfs://ClusterX/</value>
</property>
<property>
<name>fs.viewfs.mounttable.ClusterX.link./folder1</name>
<value>hdfs://NS1/folder1</value>
</property>
<property>
<name>fs.viewfs.mounttable.ClusterX.link./folder2</name>
<value>hdfs://NS2/folder2</value>
</property>
```

对于依赖HDFS的MR配置中，需要使用已挂载的目录。

错误示例：

```
<property>
<name>yarn.app.mapreduce.am.staging-dir</name>
<value>/tmp/hadoop-yarn/staging</value>
</property>
```

根目录（/）在viewFS中是无法访问的。

正确示例：

```
<property>
<name>yarn.app.mapreduce.am.staging-dir</name>
<value>/folder1/tmp/hadoop-yarn/staging</value>
</property>
```

12.18.9.9 基于分区的任务黑名单

问题

Map&Reduce任务失败，并且故障节点数与集群总节点数的比值低于“yarn.resourcemanager.am-scheduling.node-blacklisting-disable-threshold”配置的黑名单阈值，为什么Map&Reduce任务故障节点没有加入黑名单?

回答

当集群中有超过阈值的节点都被加入黑名单时，黑名单会释放这些节点，其中阈值为故障节点数与集群总节点数的比值。现在每个节点都有其标签表达式，黑名单阈值应根据有效节点标签表达式关联的节点数进行计算，其值为故障节点数与有效节点标签表达式关联的节点数的比值。

假设集群中有100个节点，其中有10个节点为有效节点标签表达式关联的节点（labelA）。其中所有有效节点标签表达式关联的节点都已经故障，黑名单节点释放阈值默认值为0.33，按照传统的计算方式，10/100=0.1，远小于该阈值。这就造成这10个节点永远无法得到释放，Map&Reduce任务一直无法获取节点，应用程序无法正常运行。实际需要根据与Map&Reduce任务的有效节点关联的节点总数进行计算，即10/10=1，大于黑名单节点释放阈值，节点被释放。

因此即使故障节点数与集群总节点数的比值没有超过阈值，也存在黑名单将这些节点释放的情况。

12.19 使用 Oozie

12.19.1 从零开始使用 Oozie

Oozie是一个基于工作流引擎的开源框架，能够提供对Hadoop作业的任务调度与协调。

Oozie支持提交多种类型任务，例如Hive、Spark2x、Loader、Mapreduce、Java、DistCp、Shell、HDFS、SSH、SubWorkflow、Streaming、定时任务等。

本章节指导用户通过使用Oozie客户端提交MapReduce任务。

前提条件

已安装客户端。例如安装目录为“/opt/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。假如客户端安装目录为：/opt/client，请根据实际安装目录修改。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 判断集群认证模式。

- 安全模式，执行以下命令进行用户认证。*UserOozie*为提交任务的用户。

  kinit *UserOozie*
- 普通模式，执行**步骤5**。

步骤5 上传Oozie配置文件以及Jar包至HDFS：

hdfs dfs -mkdir /user/*UserOozie*

hdfs dfs -put -f /opt/client/Oozie/oozie-client-*/examples /user/*UserOozie*/

说明

- “/opt/client/”为客户端安装目录，请根据实际安装目录修改。
- UserOozie为提交任务的用户。

步骤6 修改任务执行配置文件：

cd /opt/client/Oozie/oozie-client-*/examples/apps/map-reduce/

vi job.properties

```
nameNode=hdfs://hacluster
resourceManager=10.64.35.161:8032（10.64.35.161为Yarn resourceManager（Active）节点业务平面IP；8032为yarn.resourcemanager.port）
queueName=default
examplesRoot=examples
user.name=admin
oozie.wf.application.path=${nameNode}/user/${user.name}/${examplesRoot}/apps/map-reduce #hdfs上传路径
outputDir=map-reduce
oozie.wf.rerun.failnodes=true
```

步骤7 运行oozie任务：

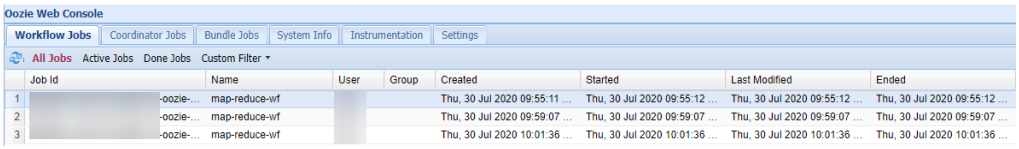
oozie job -oozie https://*oozie角色的主机名*:21003/oozie/ -config job.properties -run

```
[root@kwephispra44947 map-reduce]# oozie job -oozie https://kwephispra44948:21003/oozie/ -config job.properties -run
.....
job: 0000000-200730163829770-oozie-omm-W
```

步骤8 登录FusionInsight Manager。具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。

步骤9 选择“集群 > *待操作集群的名称* > 服务 > Oozie ”，单击“oozie WebUI”后的超链接进入Oozie页面，在Oozie的WebUI上查看任务运行结果。

图 12-37 任务运行结果



----结束

12.19.2 使用 Oozie 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Oozie客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。

使用 Oozie 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录，该操作的客户端目录只是举例，请根据实际安装目录修改。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 判断集群认证模式。

- 安全模式，执行以下命令进行用户认证。*exampleUser*为提交任务的用户名。

  kinit *exampleUser*
- 普通模式，执行**步骤5**。

步骤5 配置Hue。

1. spark2x环境配置（如果不涉及spark2x任务，可以跳过此步骤）：

   hdfs dfs -put /opt/client/Spark2x/spark/jars/*.jar /user/oozie/share/lib/spark2x/

   当HDFS目录“/user/oozie/share”中的Jar包发生变化时，需要重启Oozie服务。
2. 上传Oozie配置文件以及Jar包至HDFS：

   hdfs dfs -mkdir /user/*exampleUser*

   hdfs dfs -put -f /opt/client/Oozie/oozie-client-*/examples /user/*exampleUser*/

   说明

   - *exampleUser*为提交任务的用户名。
   - 在提交任务的用户和非job.properties文件均无变更的前提下，客户端安装目录/Oozie/oozie-client-*/examples目录一经上传HDFS，后续可重复使用，无需多次提交。
   - 解决Spark和Yarn关于jetty的jar冲突。

     hdfs dfs -rm -f /user/oozie/share/lib/spark/jetty-all-9.2.22.v20170606.jar
   - 普通模式下，上传过程如果遇到“Permission denied”的问题，可执行以下命令进行处理。

     su - omm

     source /opt/client/bigdata_env

     hdfs dfs -chmod -R 777 /user/oozie

     exit

----结束

12.19.3 使用 Oozie 客户端提交作业

12.19.3.1 提交 Hive 任务

操作场景

该任务指导用户在使用Oozie客户端提交Hive任务

Hive任务有如下类型：

- Hive作业
  使用JDBC方式连接的Hive作业。
- Hive2作业
  使用Beeline方式连接的Hive作业。

本文以使用Oozie客户端提交Hive作业为例介绍。

说明

- 使用Oozie客户端提交Hive2作业与提交Hive作业操作步骤一致，只需将操作步骤中对应路径的“/Hive”改成“/Hive2”即可。
  例如，Hive作业运行目录“/opt/client/Oozie/oozie-client-*/examples/apps/hive/”，则Hive2对应的运行目录为“/opt/client/Oozie/oozie-client-*/examples/apps/hive2/”。
- 建议下载使用最新版本的客户端。

前提条件

- Hive和Oozie组件及客户端已经安装，并且正常运行。
- 已创建或获取访问Oozie服务的人机用户帐号及密码。

  说明

  - 该用户需要从属于hadoop、supergroup、hive组，同时添加Oozie的角色操作权限。若使用Hive多实例，该用户还需要从属于具体的Hive实例组，如hive3。
  - 用户同时还需要至少有manager_viewer权限的角色。
- 获取运行状态的Oozie服务器（任意实例）URL，如“https://10.1.130.10:21003/oozie”。
- 获取运行状态的Oozie服务器主机名，如“10-1-130-10”。
- 获取Yarn ResourceManager主节点IP，如10.1.130.11。

操作步骤

步骤1 以客户端安装用户，登录安装Oozie客户端的节点。

步骤2 执行以下命令，获取安装环境信息。其中“/opt/client/”为客户端安装路径，该操作的客户端目录只是举例，请根据实际安装目录修改。

source /opt/client/bigdata_env

步骤3 判断集群认证模式。

- 安全模式，执行**kinit**命令进行用户认证。
  例如，使用**oozieuser**用户进行认证。
  kinit oozieuser
- 普通模式，执行**步骤4**。

步骤4 执行以下命令，进入样例目录。

cd /opt/client/Oozie/oozie-client-*/examples/apps/hive/

该目录下需关注文件如**表12-312**所示。

表 12-312 文件说明

文件名称	描述
hive-site.xml	Hive任务的配置文件。
job.properties	工作流的参数变量定义文件。
script.q	Hive任务的SQL脚本。
workflow.xml	工作流的规则定制文件。

步骤5 执行以下命令，编辑“job.properties”文件。

vi job.properties

修改如下内容：

更改“userName”的参数值为提交任务的人机用户名，例如“userName=oozieuser”。

步骤6 执行**oozie job**命令，运行工作流文件。

oozie job -oozie https://*oozie角色的主机名*:21003/oozie/ -config job.properties -run

说明

- 命令参数解释如下：
  -oozie 实际执行任务的Oozie服务器URL
  -config 工作流属性文件
  -run 运行工作流
- 执行完工作流文件，显示job id表示提交成功，例如：job: 0000021-140222101051722-oozie-omm-W。登录Oozie管理页面，查看运行情况。
  使用**oozieuser**用户，登录Oozie WebUI页面：https://*oozie角色的ip地址*:21003/oozie 。
  Oozie的WebUI界面中，可在页面表格根据jobid查看已提交的工作流信息。

----结束

12.19.3.2 提交 Spark2x 任务

操作场景

该任务指导用户在使用Oozie客户端提交Spark2x任务。

说明

请下载使用最新版本的客户端。

前提条件

- Spark2x和Oozie组件安装完成且运行正常，客户端安装成功。
  如果当前客户端为旧版本，需要重新下载和安装客户端。
- 已创建或获取访问Oozie服务的人机用户帐号及密码。

说明

- 该用户需要从属于hadoop、supergroup、hive组，同时添加Oozie的角色操作权限。若使用Hive多实例，该用户还需要从属于具体的Hive实例组，如hive3。
- 用户同时还需要至少有manager_viewer权限的角色。

- 获取运行状态的Oozie服务器（任意实例）URL，如“https://*10.1.130.10:21003*/oozie”。
- 获取运行状态的Oozie服务器主机名，如“*10-1-130-10*”。
- 获取Yarn ResourceManager主节点IP，如“*10.1.130.11*”。

操作步骤

步骤1 以客户端安装用户登录安装Oozie客户端的节点。

步骤2 执行以下命令，获取安装环境信息。其中“/opt/client/”为客户端安装路径，该操作的客户端目录只是举例，请根据实际安装目录修改。

source /opt/client/bigdata_env

步骤3 判断集群认证模式。

- 安全模式，执行**kinit**命令进行用户认证。

  例如，使用**oozieuser**用户进行认证。

  kinit oozieuser
- 普通模式，执行**步骤4**。

步骤4 执行以下命令，进入样例目录。

cd /opt/client/Oozie/oozie-client-*/examples/apps/spark2x/

该目录下需关注文件如**表12-313**所示。

表 12-313 文件说明

文件名称	描述
job.properties	工作流的参数变量定义文件。
workflow.xml	工作流的规则定制文件。
lib	工作流运行依赖的jar包目录。

步骤5 执行以下命令，编辑“job.properties”文件。

vi job.properties

修改如下内容：

更改“userName”的参数值为提交任务的人机用户名，例如“userName=oozieuser”。

步骤6 执行**oozie job**命令，运行工作流文件。

oozie job -oozie https://*oozie角色的主机名*:21003/oozie/ -config job.properties -run

说明

- 命令参数解释如下：
  -oozie 实际执行任务的Oozie服务器URL
  -config 工作流属性文件
  -run 运行工作流
- 执行完工作流文件，显示“job id”表示提交成功，例如“job: 0000021-140222101051722-oozie-omm-W”。登录Oozie管理页面，查看运行情况。
  使用**oozieuser**用户，登录Oozie WebUI页面：https://*oozie角色的ip地址*:21003/oozie 。
  Oozie的WebUI界面中，可在页面表格根据“job id”查看已提交的工作流信息。

----结束

12.19.3.3 提交 Loader 任务

操作场景

该任务指导用户在使用Oozie客户端提交Loader任务。

说明

请下载使用最新版本的客户端。

前提条件

- Loader和Oozie组件及客户端已经安装，并且正常运行。
- 已创建或获取访问Oozie服务的人机用户帐号及密码。

  说明

  - 该用户需要从属于hadoop、supergroup、hive组，同时添加Oozie的角色操作权限。若使用Hive多实例，该用户还需要从属于具体的Hive实例组，如hive3。
  - 用户同时还需要至少有manager_viewer权限的角色。
- 获取运行状态的Oozie服务器（任意实例）URL，如“https://10.1.130.10:21003/oozie”。
- 获取运行状态的Oozie服务器主机名，如“10-1-130-10”。
- 获取Yarn ResourceManager主节点IP，如10.1.130.11。
- 创建需要调度的Loader作业，并获取该作业ID。

操作步骤

步骤1 以客户端安装用户，登录安装Oozie客户端的节点。

步骤2 执行以下命令，获取安装环境信息。其中“/opt/client/”为客户端安装路径，该操作的客户端目录只是举例，请根据实际安装目录修改。

source /opt/client/bigdata_env

步骤3 判断集群认证模式。

- 安全模式，执行**kinit**命令进行用户认证。
  例如，使用**oozieuser**用户进行认证。

kinit oozieuser

- 普通模式，执行**步骤4**。

步骤4 执行以下命令，进入样例目录。

cd /opt/client/Oozie/oozie-client-*/examples/apps/sqoop/

该目录下需关注文件如**表12-314**所示。

表 12-314 文件说明

文件名称	描述
job.properties	工作流的参数变量定义文件。
workflow.xml	工作流的规则定制文件。

步骤5 执行以下命令，编辑“job.properties”文件。

vi job.properties

修改如下内容：

更改“userName”的参数值为提交任务的人机用户名，例如“userName=oozieuser”。

步骤6 执行以下命令，编辑“workflow.xml”文件。

vi workflow.xml

修改如下内容：

“command”的值修改为需要调度的已有Loader作业ID，例如1。

将“workflow.xml”文件上传至 "job.properties" 文件中的HDFS路径。

hdfs dfs -put -f workflow.xml /user/userName/examples/apps/sqoop

步骤7 执行**oozie job**命令，运行工作流文件。

oozie job -oozie https://*oozie角色的主机名*:21003/oozie/ -config job.properties -run

说明

- 命令参数解释如下：
  -oozie 实际执行任务的Oozie服务器URL
  -config 工作流属性文件
  -run 运行工作流
- 执行完工作流文件，显示job id表示提交成功，例如：job: 0000021-140222101051722-oozie-omm-W。登录Oozie管理页面，查看运行情况。
  使用**oozieuser**用户，登录Oozie WebUI页面：https://*oozie角色的ip地址*:21003/oozie 。
  Oozie的WebUI界面中，可在页面表格根据jobid查看已提交的工作流信息。

----结束

12.19.3.4 提交 DistCp 任务

操作场景

该任务指导用户在使用Oozie客户端提交DistCp任务。

说明

请下载使用最新版本的客户端。

前提条件

- HDFS和Oozie组件安装完成且运行正常，客户端安装成功。
  如果当前客户端为旧版本，需要重新下载和安装客户端。
- 已创建或获取访问Oozie服务的人机用户帐号及密码。

  说明

  - 该用户需要从属于hadoop、supergroup、hive组，同时添加Oozie的角色操作权限。若使用Hive多实例，该用户还需要从属于具体的Hive实例组，如hive3。
  - 用户同时还需要至少有manager_viewer权限的角色。
- 已获取运行状态的Oozie服务器（任意实例）URL，如“https://*10.1.130.10:21003*/oozie”。
- 已获取运行状态的Oozie服务器主机名，如“*10-1-130-10*”。
- 已获取Yarn ResourceManager主节点IP，如“*10.1.130.11*”。

操作步骤

步骤1 以客户端安装用户登录安装Oozie客户端的节点。

步骤2 执行以下命令，获取安装环境信息。其中“/opt/client/”为客户端安装路径，该操作的客户端目录只是举例，请根据实际安装目录修改。

source /opt/client/bigdata_env

步骤3 判断集群认证模式。

- 安全模式，执行**kinit**命令进行用户认证。
  例如，使用**oozieuser**用户进行认证。
  kinit oozieuser
- 普通模式，执行**步骤4**。

步骤4 执行以下命令，进入样例目录。

cd /opt/client/Oozie/oozie-client-*/examples/apps/distcp/

该目录下需关注文件如**表12-315**所示。

表 12-315 文件说明

文件名称	描述
job.properties	工作流的参数变量定义文件。

文件名称	描述
workflow.xml	工作流的规则定制文件。

步骤5 执行以下命令，编辑“job.properties”文件。

vi job.properties

修改如下内容：

更改“userName”的参数值为提交任务的人机用户名，例如“userName=oozieuser”。

步骤6 是否是跨安全集群的DistCp。

- 是，执行步骤**步骤7**。
- 否，则执行步骤**步骤9**。

步骤7 对两个集群进行跨Manager集群互信。

步骤8 备份并且修改workflow.xml的文件内容，命令如下：

cp workflow.xml workflow.xml.bak

vi workflow.xml

修改以下内容：

```
<workflow-app xmlns="uri:oozie:workflow:1.0" name="distcp-wf">
   <start to="distcp-node"/>
   <action name="distcp-node">
      <distcp xmlns="uri:oozie:distcp-action:1.0">
         <resource-manager>${resourceManager}</resource-manager>
         <name-node>${nameNode}</name-node>
         <prepare>
            <delete path="hdfs://target_ip:target_port/user/${userName}/${examplesRoot}/output-data/${outputDir}"/>
         </prepare>
         <configuration>
            <property>
               <name>mapred.job.queue.name</name>
               <value>${queueName}</value>
            </property>
            <property>
               <name>oozie.launcher.mapreduce.job.hdfs-servers</name>
               <value>hdfs://source_ip:source_port,hdfs://target_ip:target_port</value>
            </property>
         </configuration>
         <arg>${nameNode}/user/${userName}/${examplesRoot}/input-data/text/data.txt</arg>
         <arg>hdfs://target_ip:target_port/user/${userName}/${examplesRoot}/output-data/${outputDir}/data.txt</arg>
         </distcp>
      <ok to="end"/>
      <error to="fail"/>
   </action>
   <kill name="fail">
      <message>DistCP failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
   </kill>
   <end name="end"/>
</workflow-app>
```

其中“target_ip:target_port”为另一个互信集群的HDFS active namenode地址，例如：10.10.10.233:25000。

“source_ip:source_port ”为源集群的HDFS active namenode地址，例如：10.10.10.223:25000。

两个IP地址和端口都需要根据自身的集群实际情况修改。

步骤9 执行**oozie job**命令，运行工作流文件。

oozie job -oozie https://*oozie角色的主机名*:21003/oozie/ -config job.properties -run

说明

- 命令参数解释如下：
  -oozie 实际执行任务的Oozie服务器URL
  -config 工作流属性文件
  -run 运行工作流
- 执行完工作流文件，显示“job id”表示提交成功，例如“job: 0000021-140222101051722-oozie-omm-W”。登录Oozie管理页面，查看运行情况。
  使用**oozieuser**用户，登录Oozie WebUI页面：https://*oozie角色的ip地址*:21003/oozie 。
  Oozie的WebUI界面中，可在页面表格根据“job id”查看已提交的工作流信息。

----结束

12.19.3.5 提交其它任务

操作场景

除了Hive、Spark2x、Loader任务，也支持使用Oozie客户端提交MapReduce、Java、Shell、HDFS、SSH、SubWorkflow、Streaming、定时等任务。

说明

请下载使用最新版本的客户端。

前提条件

- Oozie组件及客户端已经安装，并且正常运行。
- 已创建或获取访问Oozie服务的人机用户帐号及密码。

  说明

  - Shell任务：
    该用户需要从属于hadoop、supergroup组，添加Oozie的角色操作权限，并确保Shell脚本在每个nodemanager节点都有执行权限。
  - SSH任务：
    该用户需要从属于hadoop、supergroup组，添加Oozie的角色操作权限，并完成互信配置。
  - 其他任务：
    该用户需要从属于hadoop、supergroup组，添加Oozie的角色操作权限，并具备对应任务类型所需的权限。
  - 用户同时还需要至少manager_viewer权限的角色。
- 获取运行状态的Oozie服务器（任意实例）URL，如“https://10.1.130.10:21003/oozie”。

- 获取运行状态的Oozie服务器主机名，如“10-1-130-10”。
- 获取Yarn ResourceManager主节点IP，如10.1.130.11。

操作步骤

步骤1 以客户端安装用户，登录安装Oozie客户端的节点。

步骤2 执行以下命令，获取安装环境信息。其中“/opt/client/”为客户端安装路径，该操作的客户端目录只是举例，请根据实际安装目录修改。

source /opt/client/bigdata_env

步骤3 判断集群认证模式。

- 安全模式，执行**kinit**命令进行用户认证。

  例如，使用**oozieuser**用户进行认证。

  kinit oozieuser
- 普通模式，执行**步骤4**。

步骤4 根据提交任务类型，进入对应样例目录。

表 12-316 样例目录列表

任务类型	样例目录
Mapreduce任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/map-reduce
Java任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/java-main
Shell任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/shell
Streaming任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/streaming
SubWorkflow任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/subwf
SSH任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/ssh
定时任务	*客户端安装目录*/Oozie/oozie-client-*/examples/apps/cron

说明

其他任务样例中已包含HDFS任务样例。

样例目录下需关注文件如**表12-317**所示。

表 12-317 文件说明

文件名称	描述
job.properties	工作流的参数变量定义文件。

文件名称	描述
workflow.xml	工作流的规则定制文件。
lib	工作流运行依赖的jar包目录。
coordinator.xml	“cron”目录下存在，定时任务配置文件，用于设置定时策略。
oozie_shell.sh	“shell”目录下存在，提交Shell任务需要的Shell脚本文件。

步骤5 执行以下命令，编辑“job.properties”文件。

vi job.properties

修改如下内容：

更改“userName”的参数值为提交任务的人机用户名，例如“userName=oozieuser”。

步骤6 执行**oozie job**命令，运行工作流文件。

oozie job -oozie https://*oozie角色的主机名***:21003/oozie -config** *job.properties文件所在路径* **-run**

例如：

oozie job -oozie https://10-1-130-10:21003/oozie -config

/opt/client/Oozie/oozie-client-*/examples/apps/map-reduce/job.properties -run

说明

- 命令参数解释如下：
  -oozie 实际执行任务的Oozie服务器URL
  -config 工作流属性文件
  -run 运行工作流
- 执行完工作流文件，显示job id表示提交成功，例如：job: 0000021-140222101051722-oozie-omm-W。登录Oozie管理页面，查看运行情况。
  使用**oozieuser**用户，登录Oozie WebUI页面：https://*oozie角色的ip地址*:21003/oozie 。
  Oozie的WebUI界面中，可在页面表格根据jobid查看已提交的工作流信息。

----结束

12.19.4 使用 Hue 提交 Oozie 作业

12.19.4.1 创建工作流

操作场景

用户通过Hue管理界面可以进行提交Oozie作业，提交作业之前，首先需要创建一个工作流。

前提条件

使用Hue提交Oozie作业之前，需要提前配置好Oozie客户端，并上传样例配置文件和jar至HDFS指定目录，具体操作请参考**使用Oozie客户端**章节。

操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在界面左侧导航栏单击  ，选择“Workflow”，打开Workflow编辑器。

步骤3 单击“文档”后的下拉框选择“操作”， 在操作列表中选择需要创建的作业类型，将其拖到操作界面中即可。

不同类型作业提交请参考以下章节：

- **提交Hive2作业**
- **提交Spark2x作业**
- **提交Java作业**
- **提交Loader作业**
- **提交Mapreduce作业**
- **提交Sub workflow作业**
- **提交Shell作业**
- **提交HDFS作业**
- **提交Streaming作业**
- **提交Distcp作业**

----结束

12.19.4.2 提交 Workflow 工作流作业

12.19.4.2.1 提交 Hive2 作业

操作场景

该任务指导用户通过Hue界面提交Hive2类型的Oozie作业。

操作步骤

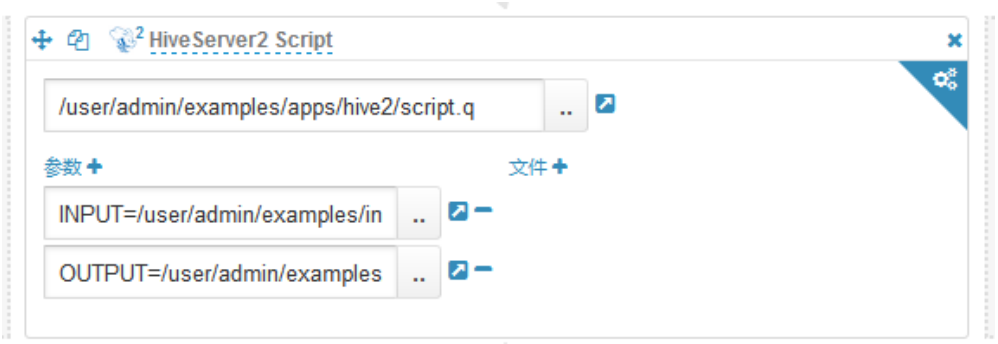
步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“HiveServer2 脚本”按钮  ，将其拖到操作区中。

步骤3 在弹出的“HiveServer2 Script”窗口中配置HDFS上的脚本路径，例如“/user/admin/examples/apps/hive2/script.q”，然后单击“添加”。

步骤4 单击“参数+”，添加输入输出参数。

例如输入参数为“INPUT=/user/admin/examples/input-data/table”，输出参数为“OUTPUT=/user/admin/examples/output-data/hive2_workflow”。



步骤5 单击右上角的配置按钮  。在打开的配置界面中，单击“删除+”，添加删除目录，例如“/user/admin/examples/output-data/hive2_workflow”。

步骤6 配置“作业 XML”，例如配置为hdfs路径“/user/admin/examples/apps/hive2/hive-site.xml”。

说明

若以上的参数和值在使用过程中发生了修改，可在“Oozie客户端安装目录/oozie-client-*/conf/hive-site.xml”文件中查询。

步骤7 单击Oozie编辑器右上角的 。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Hive2-Workflow”。

步骤8 保存完成后，单击 ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.2 提交 Spark2x 作业

操作场景

该任务指导用户通过Hue界面提交Spark2x类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“Spark 程序”按钮 ，将其拖到操作区中。

步骤3 在弹出的“Spark”窗口配置“Files”，例如“hdfs://hacluster/user/admin/examples/apps/spark2x/lib/oozie-examples.jar”。配置“jar/py name”，例如“oozie-examples.jar”，配置完成后单击“添加”。

步骤4 配置“Main class”的值。例如“org.apache.oozie.example.SparkFileCopy”。

步骤5 单击“参数+”，添加输入输出相关参数。

例如添加：

- “hdfs://hacluster/user/admin/examples/input-data/text/data.txt”
- “hdfs://hacluster/user/admin/examples/output-data/spark_workflow”

步骤6 在“Options list”文本框指定spark参数，例如“--conf spark.yarn.archive=hdfs://hacluster/user/spark2x/jars/8.1.0.1/spark-archive-2x.zip --conf spark.eventLog.enabled=true --conf spark.eventLog.dir=hdfs://hacluster/spark2xJobHistory2x”。

说明

此处版本号8.1.0.1为示例，具体以实际环境的版本号为准。

步骤7 单击右上角的配置按钮  。配置“Spark Master”的值，例如“yarn-cluster”。配置“Mode”的值，例如“cluster”。

步骤8 在打开的配置界面中，单击“删除+”，添加删除目录，例如“hdfs://hacluster/user/admin/examples/output-data/spark_workflow”。

步骤9 单击“属性+”，添加oozie使用的sharelib，左边文本框填写属性名称“oozie.action.sharelib.for.spark”，右边文本框填写属性值“spark2x”。

步骤10 单击Oozie编辑器右上角的  。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Spark-Workflow”。

步骤11 保存完成后，单击  ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.3 提交 Java 作业

操作场景

该任务指导用户通过Hue界面提交Java类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“Java 程序”按钮  ，将其拖到操作区中。

步骤3 在弹出的“Java program”窗口中配置“Jar name”的值，例如“/user/admin/examples/apps/java-main/lib/oozie-examples-5.1.0.jar”。配置“Main class”的值，例如“org.apache.oozie.example.DemoJavaMain”。然后单击“添加”。

步骤4 单击Oozie编辑器右上角的 。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Java-Workflow”。

步骤5 保存完成后，单击 ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.4 提交 Loader 作业

操作场景

该任务指导用户通过Hue界面提交Loader类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“Loader”按钮 ，将其拖到操作区中。

步骤3 在弹出的“Loader”窗口中配置“Job id”的值，例如“1”。然后单击“添加”。

说明

“Job id”是需要编排的Loader作业的ID值，可从Loader页面获取。

创建需要调度的Loader作业，并获取该作业ID，具体操作请参见**使用Loader**相关章节。

步骤4 单击Oozie编辑器右上角的 。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Loader-Workflow”。

步骤5 保存完成后，单击 ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.5 提交 Mapreduce 作业

操作场景

该任务指导用户通过Hue界面提交Mapreduce类型的Oozie作业。

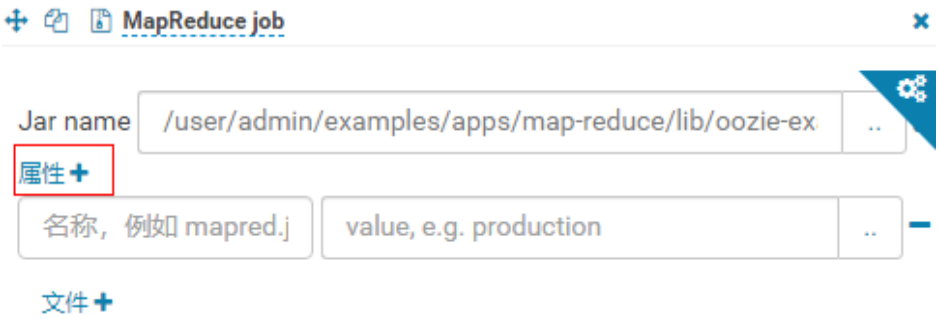
操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“MapReduce 作业”按钮  ，将其拖到操作区中。

步骤3 在弹出的“MapReduce job”窗口中配置“Jar name”的值，例如“/user/admin/examples/apps/map-reduce/lib/oozie-examples-5.1.0.jar”。然后单击“添加”。

步骤4 单击“属性+”，添加输入输出相关属性。



例如配置“mapred.input.dir”的值为“/user/admin/examples/input-data/text”，配置“mapred.output.dir”的值为“/user/admin/examples/output-data/map-reduce_workflow”。

步骤5 单击右上角的配置按钮  。在打开的配置界面中，单击“删除+”，添加删除目录，例如“/user/admin/examples/output-data/map-reduce_workflow”。

步骤6 单击Oozie编辑器右上角的  。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“MapReduce-Workflow”。

步骤7 保存完成后，单击  ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.6 提交 Sub workflow 作业

操作场景

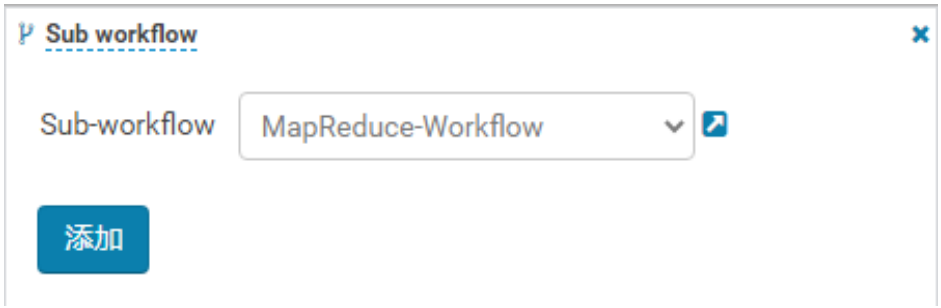
该任务指导用户通过Hue界面提交Sub Workflow类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“子Workflow”按钮  ，将其拖到操作区中。

步骤3 在弹出的“Sub workflow”窗口中配置“Sub-workflow”的值，例如从下拉列表中选取“Java-Workflow”（这个值是已经创建好的工作流之一），然后单击“添加”。



步骤4 单击Oozie编辑器右上角的  。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Subworkflow-Workflow”。

步骤5 保存完成后，单击  ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.7 提交 Shell 作业

操作场景

该任务指导用户通过Hue界面提交Shell类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“Shell”按钮  ，将其拖到操作区中。

步骤3 在弹出的“Shell”窗口中配置“Shell command”的值，例如“oozie_shell.sh”，然后单击“添加”。

步骤4 单击“文件+”，添加Shell命令执行文件和Oozie样例执行文件。例如“/user/admin/examples/apps/shell/oozie_shell.sh”。

步骤5 单击Oozie编辑器右上角的 。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Shell-Workflow”。

步骤6 保存完成后，单击 ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

说明

- 配置Shell命令为Linux指令时，请指定为原始指令，不要使用快捷键指令。例如：**ls -l** ,不要配置成**ll**。可配置成Shell命令**ls** ，参数添加一个 “-l”。
- Windows上传Shell脚本到HDFS时，请保证Shell脚本的格式为Unix，格式不正确会导致Shell作业提交失败。

----结束

12.19.4.2.8 提交 HDFS 作业

操作场景

该任务指导用户通过Hue界面提交HDFS类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“Fs”按钮 ，将其拖到操作区中。

步骤3 在弹出的“Fs”窗口中单击“添加”。

步骤4 单击“CREATE DIRECTORY+”，添加待创建的HDFS目录。例如“/user/admin/examples/output-data/mkdir_workflow”和“/user/admin/examples/output-data/mkdir_workflow1”。

步骤5 单击Oozie编辑器右上角的 。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“HDFS-Workflow”。

步骤6 保存完成后，单击  ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.9 提交 Streaming 作业

操作场景

该任务指导用户通过Hue界面提交Streaming类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“数据流”按钮  ，将其拖到操作区中。

步骤3 在弹出的“Streaming”窗口中配置“Mapper”的值，例如“/bin/cat”。配置“Reducer”的值，例如“/usr/bin/wc”。然后单击“添加”。

步骤4 单击“文件+”，添加运行所需的文件。

例如“/user/oozie/share/lib/mapreduce-streaming/hadoop-streaming-3.1.1.jar”和“/user/oozie/share/lib/mapreduce-streaming/oozie-sharelib-streaming-5.1.0.jar”。

步骤5 单击右上角的配置按钮  。在打开的配置界面中，单击“删除+”，添加删除目录，例如“/user/admin/examples/output-data/streaming_workflow”。

步骤6 单击“属性+”，添加下列属性。

- 左边框填写属性名称“mapred.input.dir”，右边框填写属性值“/user/admin/examples/input-data/text”。
- 左边框填写属性名称“mapred.output.dir”，右边框填写属性值“/user/admin/examples/output-data/streaming_workflow”。

步骤7 单击Oozie编辑器右上角的  。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Streaming-Workflow”。

步骤8 保存完成后，单击  ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.10 提交 Distcp 作业

操作场景

该任务指导用户通过Hue界面提交Distcp类型的Oozie作业。

操作步骤

步骤1 创建工作流，请参考**创建工作流**。

步骤2 在工作流编辑页面，选择“DistCp”按钮 ，将其拖到操作区中。

步骤3 当前DistCp操作是否是跨集群操作。

- 是，执行**步骤4**。
- 否，执行**步骤7**。

步骤4 对两个集群进行跨Manager集群互信。

步骤5 在弹出的“Distcp”窗口中配置“源”的值，例如“hdfs://hacluster/user/admin/examples/input-data/text/data.txt”。配置“目标”的值，例如“hdfs://target_ip:target_port/user/admin/examples/output-data/distcp-workflow/data.txt”。然后单击“添加”。

步骤6 单击右上角的配置按钮 ，在打开的“属性”页签配置界面中，单击“属性+”，在左边文本框中填写属性名称“oozie.launcher.mapreduce.job.hdfs-servers”，在右边文本框中填写属性值“hdfs://*source_ip:source_port*,hdfs://*target_ip:target_port*”，执行**步骤8**。

说明

source_ip：源集群的HDFS的NameNode的业务地址。

source_port：源集群的HDFS的NameNode的端口号。

target_ip：目标集群的HDFS的NameNode的业务地址。

target_port：目标集群的HDFS的NameNode的端口号。

步骤7 在弹出的“Distcp”窗口中配置“源”的值，例如“/user/admin/examples/input-data/text/data.txt”。配置“目标”的值，例如“/user/admin/examples/output-data/distcp-workflow/data.txt”。然后单击“添加”。

步骤8 单击右上角的配置按钮 ，在打开的配置界面中，单击“删除+”，添加删除目录，例如“/user/admin/examples/output-data/distcp-workflow”。

步骤9 单击Oozie编辑器右上角的 。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Distcp-Workflow”。

步骤10 保存完成后，单击 ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.11 互信操作示例

操作场景

在使用Oozie节点通过SSH作业执行外部节点的Shell，需要单向免密互信时，可以参考此示例。

前提条件

已经安装Oozie，而且能与外部节点（SSH连接的节点）通信。

操作步骤

步骤1 在外部节点上确保连接SSH时使用的用户存在，且该用户“~/.ssh”目录存在。

步骤2 在Oozie所在节点上用**omm**用户登录，执行**ssh-keygen -t rsa**，生成公私钥。

步骤3 执行语句**cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys**，把公钥添加到“authorized_keys”里。

步骤4 以**root**用户将**id_rsa.pub**文件传给用户所在外部节点的某个已存在的目录下，例如/“opt/”下。

scp ~/.ssh/id_rsa.pub root@*外部节点ip*:/opt/id_rsa.pub

步骤5 登录Shell所在外部节点，进入步骤4的目录，可以看到“id_rsa.pub”这个文件。

执行**cat id_rsa.pub >> ~/.ssh/authorized_keys**语句，把公钥也添加到Shell所在的用户的“authorized_keys”里。

步骤6 更改目录的权限。

chmod 700 ~/.ssh

chmod 600 /opt/id_rsa.pub

chmod 600 ~/.ssh/authorized_keys

说明

- Shell所在节点（外部节点）的帐户需要有权限执行Shell脚本并对于所有Shell脚本里涉及到的所有目录文件有足够权限。
- 如果Oozie具有多个节点，需要在所有Oozie节点执行**步骤2**~**步骤6**。

----结束

12.19.4.2.12 提交 SSH 作业

操作场景

该任务指导用户通过Hue界面提交SSH类型的Oozie作业。

由于有安全攻击的隐患，所以默认是无法提交SSH作业的，如果想使用SSH功能，需要手动开启。

操作步骤

步骤1 开启SSH功能：

1. 在FusionInsight Manager界面，选择“集群 > 服务 > Oozie > 配置 > 全部配置 > oozie（角色） > 安全”，修改“oozie.job.ssh.enable”的值为“true”，单击“保存”，在弹出的“保存配置”界面单击“确定”，保存配置。
2. 在Oozie的“概览”界面，选择右上角“更多 > 重启服务”，重启Oozie服务。

步骤2 创建工作流，请参考**创建工作流**。

步骤3 添加互信操作，请参考**互信操作示例**。

步骤4 在工作流编辑页面，选择“Ssh”按钮  ，将其拖到操作区中。

步骤5 在弹出的“Ssh”窗口中配置“User and Host”和“Ssh command”的值，然后单击“添加”。

步骤6 单击Oozie编辑器右上角的  。

保存前如果需要修改作业名称（默认为“My Workflow”），可以直接单击该名称进行修改，例如“Ssh-Workflow”。

步骤7 保存完成后，单击  ，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.2.13 提交 Hive 脚本

操作场景

该任务指导用户通过Hue界面提交Hive脚本作业。

操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在界面左侧导航栏选择“ > Workflow”，打开Workflow编辑器。

步骤3 单击“文档”， 在操作列表中选择Hive脚本，将其拖到操作界面中。

步骤4 在弹出的“HiveServer2 Script”框中, 选择之前保存的Hive脚本，关于保存Hive脚本参考**在Hue WebUI使用HiveQL编辑器**章节。选择脚本后单击“添加”。

步骤5 配置“作业 XML”，例如配置为hdfs路径“/user/admin/examples/apps/hive2/hive-site.xml”，配置方式参考**提交Hive2作业**。

步骤6 单击Oozie编辑器右上角的 。

步骤7 保存完成后，单击，提交该作业。

作业提交后，可通过Hue界面查看作业的详细信息、日志、进度等相关内容。

----结束

12.19.4.3 提交 Coordinator 定时调度作业

操作场景

该任务指导用户通过Hue界面提交定时调度类型的作业。

前提条件

提交Coordinator任务之前需要提前配置好相关的workflow作业。

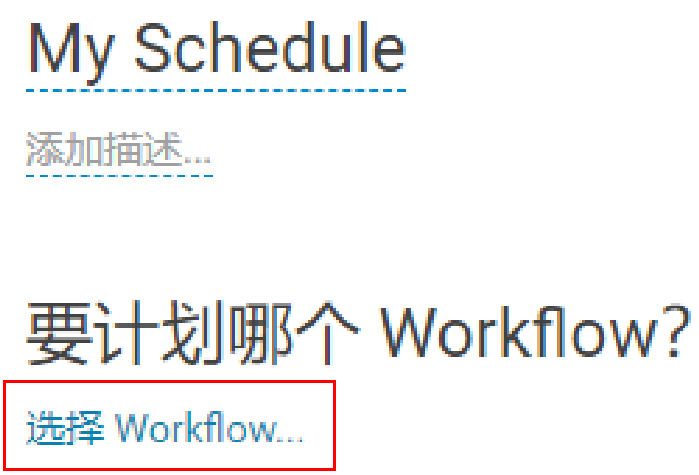
操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 在界面左侧导航栏单击  ，选择“计划”，打开Coordinator编辑器。

步骤3 在作业编辑界面中单击“My Schedule”修改作业的名称。

步骤4 单击“选择Workflow...”选择需要编排的Workflow。



步骤5 选择好Workflow，根据界面提示设置作业执行的频率，然后单击右上角的 保存作业。

说明

因时区转化的原因，此处时间有可能会与当地系统实际时间差异数个小时。

步骤6 单击编辑器右上角的  ，设置定时任务执行的时间范围的起始值与结束值，然后单击“提交”提交作业。

说明

因时区转化的原因，此处时间有可能会与当地系统实际时间差异数个小时。

----结束

12.19.4.4 提交 Bundle 批处理作业

操作场景

当同时存在多个定时任务的情况下，用户可以通过Bundle任务进行批量管理作业。该任务指导用户通过Hue界面提交批量类型的作业。

前提条件

提交Bundle批处理之前需要提前配置好相关的Workflow和Coordinator作业。

操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

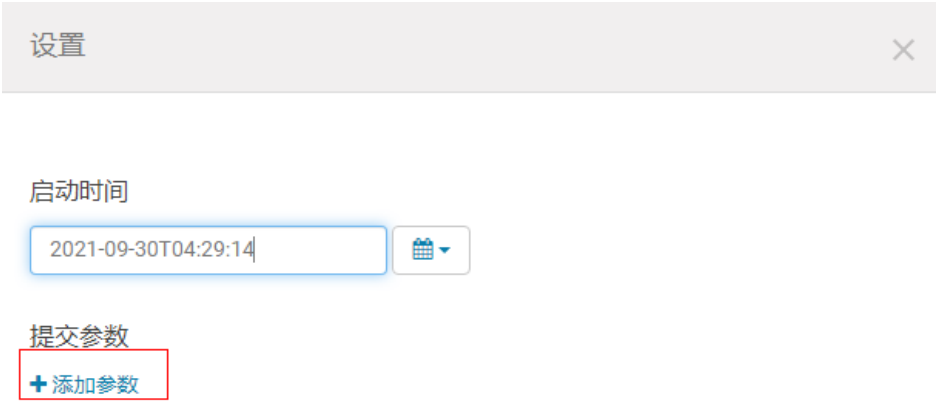
步骤2 在界面左侧导航栏单击  ，选择“Bundle”，打开Bundle编辑器。

步骤3 在作业编辑界面中单击“My Bundle”修改作业的名称。

步骤4 单击“+添加Coordinator”选择需要编排的Coordinator作业。

步骤5 根据界面提示设置Coordinator任务调度的开始、结束时间，然后单击右上角的  保存作业。

步骤6 单击编辑器右上角的  ，在弹出菜单选择  ，设置Bundle任务的启动时间，根据实际需求单击“+添加参数”设置提交参数，然后关闭对话框保存设置。



说明

因时区转化的原因，此处时间有可能会与当地系统实际时间差异数个小时。

步骤7 单击编辑器右上角的  ，在弹出的确认界面中单击“提交”提交作业。

----结束

12.19.4.5 作业结果查询

操作场景

提交作业后，可以通过Hue界面查看具体作业的执行情况。

操作步骤

步骤1 访问Hue WebUI，请参考**访问Hue的WebUI**。

步骤2 单击菜单左侧的  ，在打开的页面中可以查看Workflow、计划、Bundles任务的相关信息。

----结束

12.19.5 Oozie 日志介绍

日志描述

日志路径：Oozie相关日志的默认存储路径为：

- 运行日志：“/var/log/Bigdata/oozie”。
- 审计日志：“/var/log/Bigdata/audit/oozie”。

日志归档规则：Oozie的日志分三类：运行日志、 脚本日志和审计日志。运行日志每个文件最大20M，最多20个。审计日志每个文件最大20M，最多20个。

说明

“oozie.log”日志每小时生成一个日志压缩文件，默认保留720个（一个月的日志）。

表 12-318 Oozie 日志列表

日志类型	日志文件名	描述
运行日志	jetty.log	Oozie内置jetty服务器日志，处理OozieServlet的request/response信息
	jetty.out	Oozie进程启动日志
	oozie_db_temp.log	Oozie数据库连接日志
	oozie-instrumentation.log	Oozie仪表盘日志，主要记录Oozie运行状态，各组件的配置信息
	oozie-jpa.log	openJPa运行日志
	oozie.log	Oozie运行日志
	oozie-<SSH_USER>-<DATE>-<PID>-gc.log	Oozie服务垃圾回收日志
	oozie-ops.log	Oozie操作日志
	check-serviceDetail.log	Oozie健康检查日志
	oozie-error.log	Oozie运行错误日志
	threadDump-<DATE>.log	记录服务进程正常退出时堆栈信息的日志
脚本日志	postinstallDetail.log	安装后启动前的工作日志
	prestartDetail.log	预启动日志
	startDetail.log	服务启动日志
	stopDetail.log	服务停止日志
	upload-sharelib.log	sharelib上传操作日志
审计日志	oozie-audit.log	审计日志

日志级别

Oozie中提供了如表12-319所示的日志级别。

日志级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-319 日志级别

级别	描述
ERROR	ERROR表示错误日志，可能会导致进程异常。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及数据库底层数据传输的信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 登录FusionInsight Manager系统。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Oozie > 配置”。

步骤3 选择“全部配置”。

步骤4 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤5 选择所需修改的日志级别。

步骤6 单击“保存”，单击“确定”，处理结束后生效。

----结束

日志格式

Oozie的日志格式如下所示。

表 12-320 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS><Log Level><日志事件的发生位置>|<log中的message>	2015-05-29 21:01:45,268 INFO StatusTransitService $StatusTransitRunnable:539 - USER[-] GROUP[-] Released lock for [org.apache.oozie.service.StatusTransitService]

日志类型	格式	示例
脚本日志	<yyyy-MM-dd HH:mm:ss,SSS><主机名><Log Level><log中的message>	2015-06-01 17:18:03 001 suse11-192-168-0-111 oozie INFO Running oozie service check script
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<线程名称>|<log中的message>|<日志事件的发生位置>	2015-06-01 22:38:41,323 | INFO | http-bio-21003-exec-8 | IP [192.168.0.111] USER [null], GROUP [null], APP [null], JOBID [null], OPERATION [null], PARAMETER [null], RESULT [SUCCESS], HTTPCODE [200], ERRORCODE [null], ERRORMESSAGE [null] | org.apache.oozie.util.XLog.log(XLog.java:539)

12.19.6 Oozie 常见问题

12.19.6.1 Oozie 定时任务没有准时运行

问题

在Hue或者Oozie客户端设置执行Coordinator定时任务，但没有准时执行。

回答

需要使用UTC时间，例如在“job.properties”中配置“start=2016-12-20T09:00Z”。

12.19.6.2 HDFS 上更新了 oozie 的 share lib 目录但没有生效

问题

在HDFS的“/user/oozie/share/lib”目录上传了新的jar包，但执行任务时仍然报找不到类的错误。

回答

在客户端执行如下命令刷新目录：

oozie admin -oozie https://xxx.xxx.xxx.xxx:21003/oozie -sharelibupdate

12.19.6.3 Oozie 常用排查手段

1. 根据任务在Yarn上的任务日志排查，首先把实际的运行任务，比如Hive SQL通过beeline运行一遍，确认Hive无问题。
2. 出现“classnotfoundException”等报错，排查“/user/oozie/share/lib”路径下各组件有没有报错的类的Jar包，如果没有，添加Jar包并执行**HDFS上更新了oozie**

的share lib目录但没有生效。如果执行了更新“share lib”目录依然报找不到类，那么可以查看执行更新“share lib”的命令打印出来的路径“sharelibDirNew”是否是“ /user/oozie/share/lib”， 一定不能是其它目录。

```
[root@host-                client]#
[root@host-                client]#
[root@host-                client]# oozie admin -oozie https://h                :21003/oozie/ -sharelibupdate
INFO CMD=admin -oozie https://host                21003/oozie/ -sharelibupdate
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/client/Oozie/oozie-client-5.1.0-hw-ei-313001-SNAPSHOT/lib/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/client/Oozie/oozie-client-5.1.0-hw-ei-313001-SNAPSHOT/lib/slf4j-simple-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
[ShareLib update status]
        sharelibDirOld = /user/oozie/share/lib
        host = https://                :21003/oozie
        sharelibDirNew = /user/oozie/share/lib
        status = Successful
```

3. 出现NosuchMethodError，排查“/user/oozie/share/lib”路径下各组件的Jar包是不是有多个版本，注意业务本身上传的Jar包冲突，可通过Oozie在Yarn上的运行日志打印的加载的Jar包排查是否有Jar包冲突。
4. 自研代码运行异常，可以先运行Oozie的自带样例，排除Oozie自身的异常。
5. 寻求技术人员的支持，需要收集Yarn上Oozie任务运行日志、Oozie自身的日志及组件的运行的日志，例如使用Oozie运行Hive报异常，需收集Hive的日志。

12.20 使用 Presto

12.20.1 访问 Presto 的 WebUI

用户可以通过Presto的WebUI，在图形化界面查看Presto的统计信息。Presto的WebUI界面不支持使用IE浏览器访问，建议使用Google浏览器访问。

前提条件

- 已安装Presto服务的集群。
- 已安装集群客户端，例如安装目录为“/opt/client”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

访问 Presto 的 WebUI

- 方法一（适用于MRS 3.x及之后版本）：
  a. 登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 ”。
  b. 选择“Presto”并在“基本信息”的“Coordinator WebUI”中单击“Coordinator(Coordinator)”，打开Presto的WebUI页面。

  图 12-38 Coordinator WebUI

- 方法二（适用于MRS 3.x之前版本）：
  a. 登录MRS Manager页面，选择“服务管理”。
  b. 选择“Presto”并在“Presto 概述”的“Presto WebUI”中单击“Coordinator (主)”，打开Presto的WebUI页面。

  图 12-39 Presto WebUI

  说明

  第一次访问Presto WebUI，需要在浏览器中添加站点信任以继续打开页面。

- 方法三（适用于MRS 1.9.2及之后版本）：
  a. 在集群列表页面，单击集群名称，登录集群详情页面，选择“组件管理”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

  b. 选择“Presto”并在“Presto 概述”的“Presto WebUI”中单击“Coordinator (主)”，打开Presto的WebUI页面。

图 12-40 Presto WebUI

12.20.2 使用客户端执行查询语句

用户可以根据业务需要，在MRS集群的客户端中进行交互式查询。启用Kerberos认证的集群，需要提交拓扑的用户属于“presto”组。

MRS 3.x版本Presto组件暂不支持开启Kerberos认证。

前提条件

- 获取用户“admin”帐号密码。“admin”密码在创建MRS集群时由用户指定。
- 已刷新客户端。
- 3.x版本的集群需要手动安装Presto客户端。

操作步骤

步骤1 启用Kerberos认证的集群，登录MRS Manager页面，创建拥有“Hive Admin Privilege”权限的角色。

步骤2 创建属于“Presto”和“Hive”组的用户，同时为该用户绑定**步骤1**中创建的角色，然后下载用户认证文件。

步骤3 将下载的user.keytab文件和krb5.conf上传到MRS客户端所在节点。

说明

步骤**步骤2**-**步骤3**仅启用Kerberos认证的集群执行，普通集群请请直接从步骤**步骤4**开始执行。

步骤4 根据业务情况，准备好客户端，并登录安装客户端的节点。

例如在Master2节点更新客户端，则登录该节点使用客户端。

步骤5 执行以下命令切换用户。

sudo su - omm

步骤6 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤7 执行以下命令，配置环境变量。

source bigdata_env

步骤8 连接Presto Server。根据客户端的不同，提供如下两种客户端的链接方式。

- 使用MRS提供的客户端。
  - 未启用Kerberos认证的集群，执行以下命令连接本集群的Presto Server。

    presto_cli.sh
  - 未启用Kerberos认证的集群，执行以下命令连接其他集群的Presto Server，其中ip为对应集群的Presto的浮动IP（可通过在Presto配置项中搜索PRESTO_COORDINATOR_FLOAT_IP的值获得），port为Presto Server的端口号，默认为7520。

    presto_cli.sh --server *http://ip:port*
  - 启用Kerberos认证的集群，执行以下命令连接本集群的Presto Server。

    presto_cli.sh --krb5-config-path *krb5.conf文件路径* --krb5-principal *用户principal* --krb5-keytab-path *user.keytab文件路径* --user *presto用户名*
  - 启用Kerberos认证的集群，执行以下命令连接其他集群的Presto Server，其中ip为对应集群的Presto的浮动IP（可通过在Presto配置项中搜索PRESTO_COORDINATOR_FLOAT_IP的值获得），port为Presto Server的端口号，默认为7521。

    presto_cli.sh --krb5-config-path *krb5.conf文件路径* --krb5-principal *用户principal* --krb5-keytab-path *user.keytab文件路径* --server *https://ip:port* --krb5-remote-service-name *Presto Server name*
- 使用原生客户端

  Presto原生客户端为客户端目录下的Presto/presto/bin/presto。

步骤9 执行Query语句，如“show catalogs”。

说明

启用Kerberos认证的集群使用Presto查询Hive Catalog的数据时，运行Presto客户端的用户需要有Hive表的访问权限，并且需要在Hive beeline中执行命令**grant all on table [table_name] to group hive**，给Hive组赋权限。

步骤10 查询结束后，执行以下命令退出客户端。

quit

----结束

12.21 使用 Ranger（MRS 3.x）

12.21.1 登录 Ranger 管理界面

Ranger服务提供了集中式的权限管理框架，可以对HDFS、HBase、Hive、Yarn等组件进行细粒度的权限访问控制，并且提供了Web UI方便管理员进行操作。

Ranger 用户类型

Ranger中的用户可分为Admin、User、Auditor等类型，不同用户具有的Ranger管理界面查看和操作权限不同。

- Admin：安全管理员，可查看所有页面内容，进行服务权限管理插件及权限访问控制策略的管理操作，可查看审计信息内容，可进行用户类型设置。
- Auditor：审计管理员，可查看服务权限管理插件及权限访问控制策略的内容。
- User：普通用户，可以被管理员赋予具体权限。

登录 Ranger 管理界面

安全模式（集群开启了Kerberos认证）

步骤1 使用**admin**用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。选择“集群 > 服务 > Ranger”，进入Ranger服务概览页面。

步骤2 单击“基本信息”区域中的“RangerAdmin”，进入Ranger WebUI界面。

- **admin**用户在Ranger中的用户类型为“User”，只能查看Access Manager和Security Zone页面。
- 如需查看所有管理页面，需要切换至**rangeradmin**用户或者其他具有Ranger管理员权限的用户：

  a. 在Ranger WebUI界面，单击右上角用户名，选择“Log Out”，退出当前用户。

  b. 使用**rangeradmin**用户（默认密码为Rangeradmin@123）或者其他具有Ranger管理员权限用户重新登录。

----结束

普通模式（集群关闭了Kerberos认证）：

步骤1 使用**admin**用户登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。选择“集群 > 服务 > Ranger”，进入Ranger服务概览页面。

步骤2 单击“基本信息”区域中的“RangerAdmin”，进入Ranger WebUI界面。

admin用户在Ranger中的用户类型为“Admin”，能查看Ranager所有管理页面，无需切换至**rangeradmin**用户。

说明

普通模式下使用**rangeradmin**用户登录Ranger WebUI界面，页面报错401。

----结束

在Ranger管理首页可查看当前Ranger已集成的各服务权限管理插件，用户可通过对应插件设置更细粒度的权限，具体主要操作页面功能描述参见**表12-321**。

表 12-321 Ranger 界面操作入口功能描述

入口	功能描述
Access Manager	查看当前Ranger已集成的各服务权限管理插件，用户可通过对应插件设置更细粒度的权限，具体操作请参考**配置组件权限策略**。
Audit	查看Ranger运行及权限管控相关审计日志信息，具体操作请参考**查看Ranger审计信息**。
Security Zone	配置安全区域，管理员可将各组件的资源切分为多个区域，由不同管理员为服务的指定资源设置安全策略，以便更好的管理，具体操作可参考**配置Ranger安全区**。
Settings	查看Ranger相关权限设置信息，例如查看用户、用户组、Role等，具体操作可参考**查看Ranger权限信息**

12.21.2 启用 Ranger 鉴权

操作场景

该章节指导用户如何启用Ranger鉴权。安全模式默认开启Ranger鉴权，普通模式默认关闭Ranger鉴权。

操作步骤

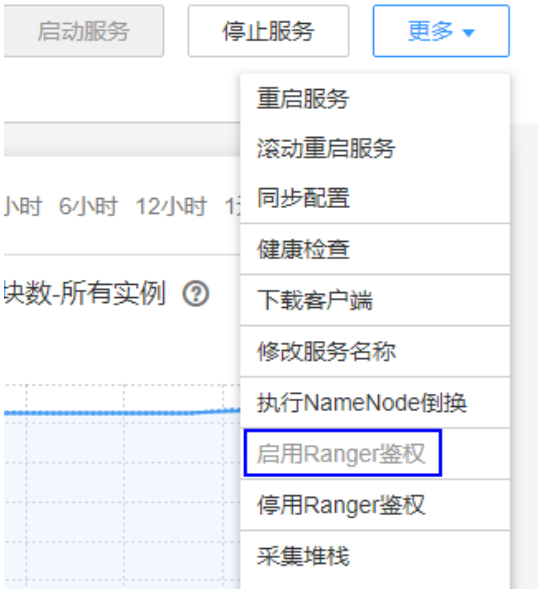
步骤1 登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。选择“集群 > 服务 > *需要启用Ranger鉴权的服务名称*”。

步骤2 在服务“概览”页面右上角单击“更多”，选择“启用Ranger鉴权”。在弹出的对话框中输入密码，单击“确定”，操作成功后单击“完成”。

说明

如果“启用Ranger鉴权”是灰色，表示已开启Ranger鉴权，如**图12-41**所示。

图 12-41 启用 Ranger 鉴权



步骤3 滚动重启服务或者重启服务。

----结束

12.21.3 配置组件权限策略

新安装的MRS集群默认安装Ranger服务并启用了Ranger鉴权模型，管理员可以通过组件权限插件对组件资源的访问设置细粒度的安全访问策略。

目前安全模式集群中支持Ranger的组件包括：HDFS、Yarn、HBase、Hive、Spark2x、Kafka、Storm。

通过 Ranger 配置用户权限策略

步骤1 使用管理员登录Ranger管理页面。

步骤2 在Ranger首页的“Service Manager”区域内，单击组件名称下的权限插件名称，即可进入组件安全访问策略列表页面。

说明

各组件的策略列表中，系统默认会生成若干条目，用于保证集群内的部分默认用户或用户组的权限（例如supergroup用户组），请勿删除，否则系统默认用户或用户组的权限会受影响。

步骤3 单击“Add New Policy”，根据业务场景规划配置相关用户或者用户组的资源访问策略。

不同组件的访问策略配置样例参考：

- **添加HDFS的Ranger访问权限策略**
- **添加HBase的Ranger访问权限策略**
- **添加Hive的Ranger访问权限策略**
- **添加Yarn的Ranger访问权限策略**

- 添加Spark2x的Ranger访问权限策略
- 添加Kafka的Ranger访问权限策略
- 添加Storm的Ranger访问权限策略

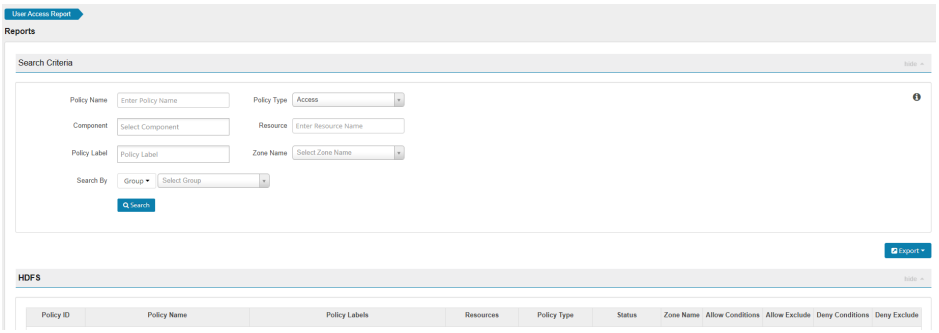
策略添加后，需等待30秒左右，待系统生效。

说明

组件每次启动都会检查组件默认的Ranger Service是否存在，如果不存在则会创建以及为其添加默认Policy。如果用户在使用过程中误删了Service，可以重启或者滚动重启相应组件服务来恢复，若是误删了默认Policy，可先手动删除Service，再重启组件服务。

步骤4 单击“Access Manager > Reports”，可查看各组件所有的安全访问策略。

系统策略较多时，可通过策略名称、类型、组件、资源对象、策略标签、安全区域、用户或用户组等信息进行过滤搜索，也可以单击“Export”导出相关策略内容。



说明

- 对于同一个固定资源对象通常只能配置一条策略，多条策略针对的具体资源对象重复时将无法保存。
- 配置策略时，不同条件的优先级可参考Ranger权限策略条件判断优先级。

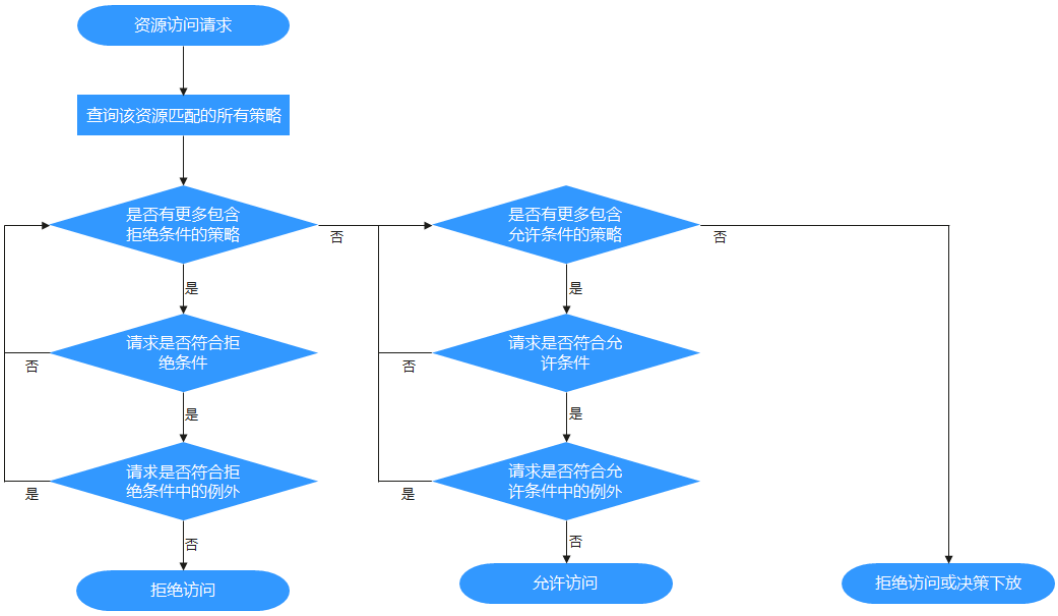
----结束

Ranger 权限策略条件判断优先级

配置资源的权限策略时，可配置针对该资源的允许条件（Allow Conditions）、允许例外条件（Exclude from Allow Conditions）、拒绝条件（Deny Conditions）以及拒绝例外条件（Exclude from Deny Conditions），以满足不同场景下的例外需求。

不同条件的优先级由高到低为：拒绝例外条件 > 拒绝条件 > 允许例外条件 > 允许条件。

系统判断流程可参考下图所示，如果组件资源请求未匹配到Ranger中的权限策略，系统默认将拒绝访问。但是对于HDFS和Yarn，系统会将决策下放给组件自身的访问控制层继续进行判断。



例如要将一个文件夹FileA的读写权限授权给用户组groupA，但是该用户组内某个用户UserA除外，这时可以增加一个允许条件及一个例外条件即可实现。

12.21.4 查看 Ranger 审计信息

管理员可通过Ranger界面查看Ranger运行审计日志及组件使用Ranger鉴权后权限管控审计日志信息。

查看 Ranger 审计信息内容

步骤1 登录Ranger管理页面。

步骤2 单击“Audit”，查看相关审计信息，各页签内容说明请参考**表12-322**，条目较多时，单击搜索框可根据关键字字段进行筛选。

表 12-322 Audit 信息

页签	内容描述
Access	用户通过Ranger鉴权访问组件资源的审计信息。
Admin	Ranger上操作审计信息，例如安全访问策略的创建/更新/删除、组件权限策略的创建/删除、role的创建/更新/删除等。
Login Sessions	登录Ranger的用户会话审计信息。
Plugins	Ranger内组件权限策略信息。
Plugin Status	各组件节点权限策略的同步审计信息。
User Sync	Ranger与LDAP用户同步审计信息。

----结束

12.21.5 配置 Ranger 安全区

Ranger支持配置安全区，管理员可将各组件的资源切分为多个安全区，由对应管理员用户为区域的指定资源设置安全策略，以便更好的细分资源管理。安全区中定义的策略仅适用于区域中的资源，服务的资源被划分到安全区后，非安全区针对该资源的访问权限策略将不再生效。安全区的管理员只能在其作为管理员的安全区中设置策略。

添加安全区

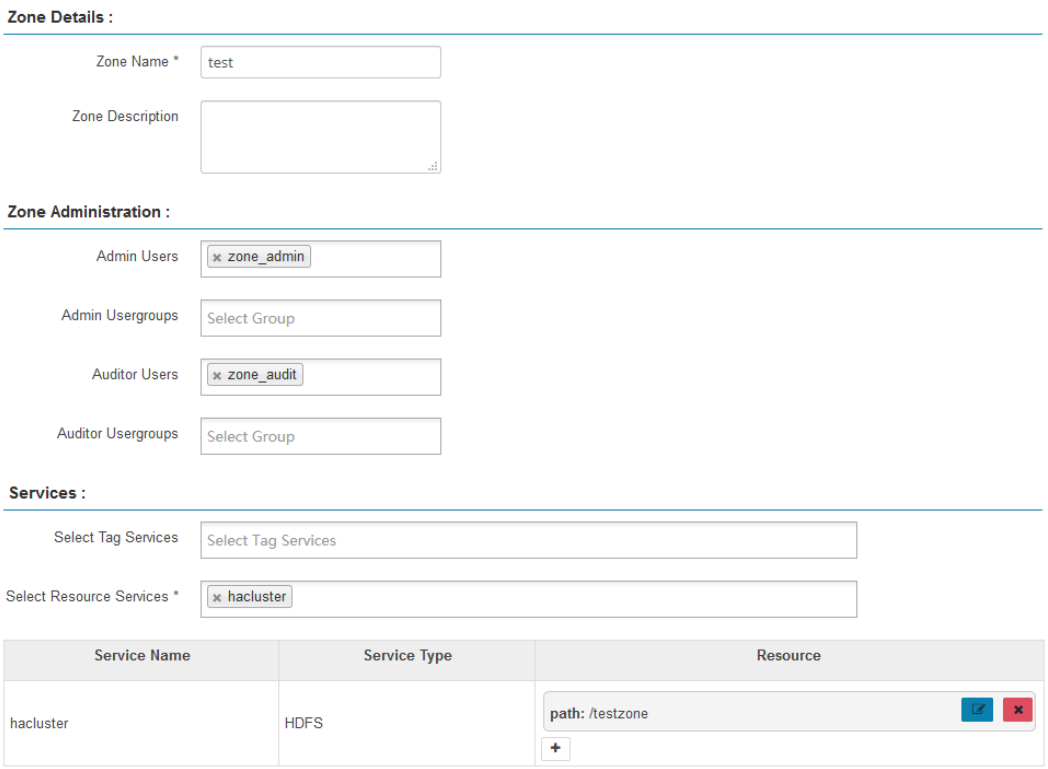
步骤1 使用Ranger管理员登录Ranger管理界面。

步骤2 单击“Security Zone”，在区域列表页面中单击  ，添加安全区。

表 12-323 安全区配置参数

参数名称	描述	示例
Zone Name	配置安全区的名称。	test
Zone Description	配置安全区的描述信息。	-
Admin Users/ Admin Usergroups	配置安全区的管理用户/用户组，可在安全区中添加及修改相关资源的权限策略。 必须至少配置一个用户或用户组。	zone_admin
Auditor Users/ Auditor Usergroups	添加审计用户/用户组，可在安全区中查看相关资源权限策略内容。 必须至少配置一个用户或用户组。	zone_user
Select Tag Services	选择服务的标签信息。	-
Select Resource Services	选择安全区内包含的服务及具体资源。 在“Select Resource Services”中选择服务后，需要在“Resource”列中添加具体的资源对象，例如HDFS服务器的文件目录、Yarn的队列、Hive的数据库及表、HBase的表及列。	/testzone

例如针对HDFS中的“/testzone”目录创建一个安全区，配置如下：



步骤3 单击“Save”，等待安全区添加成功。

Ranger管理员可在“Security Zone”页面查看当前的所有安全区并单击“Edit”修改安全区的属性信息，当相关资源不需要在安全区中进行管理时，可单击“Delete”删除对应安全区。

----结束

在安全区中配置权限策略

步骤1 使用安全区管理员用户登录Ranger管理页面。

步骤2 在Ranger首页右上角的“Security Zone”选项的下拉列表中选择对应的安全区，即可切换至该安全区内的权限视图。

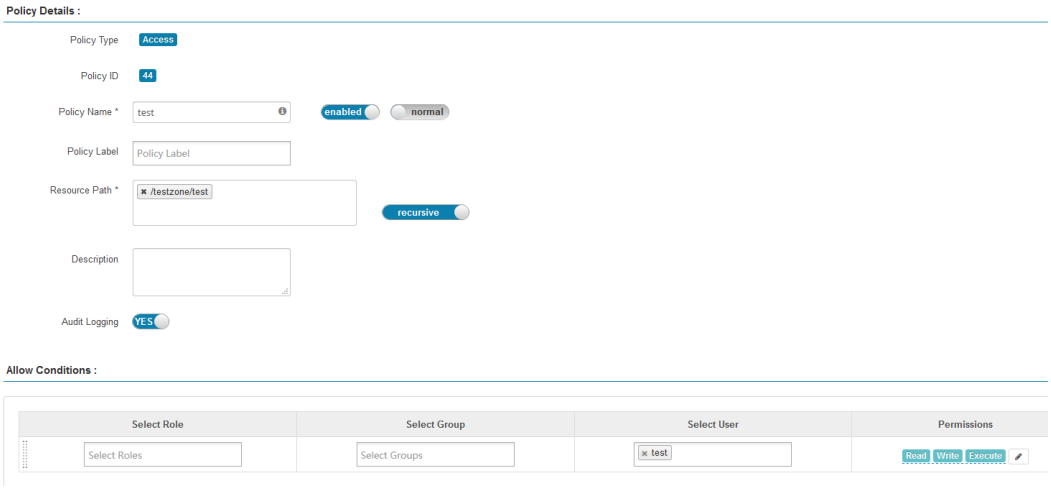
步骤3 单击组件名称下的权限插件名称，即可进入组件安全访问策略列表页面。

说明

各组件的策略列表中，系统默认生成的条目会自动继承至安全区内，用于保证集群内的部分系统默认用户或用户组的权限。

步骤4 单击“Add New Policy”，根据业务场景规划配置相关用户或者用户组的资源访问策略。

例如在本章节样例中，在安全区内配置一条允许“test”用户访问“/testzone/test”目录的策略：



其他不同组件的完整访问策略配置样例参考：

- 添加HDFS的Ranger访问权限策略
- 添加HBase的Ranger访问权限策略
- 添加Hive的Ranger访问权限策略
- 添加Yarn的Ranger访问权限策略
- 添加Spark2x的Ranger访问权限策略
- 添加Kafka的Ranger访问权限策略
- 添加Storm的Ranger访问权限策略

策略添加后，需等待30秒左右，待系统生效。

说明

- 安全区中定义的策略仅适用于区域中的资源，服务的资源被划分到安全区后，非安全区针对该资源的访问权限策略将不再生效。
- 如需配置针对当前安全区之外其他资源的访问策略，需在Ranger首页右上角的“Security Zone”选项中退出当前安全区后进行配置。

----结束

12.21.6 普通集群修改 Ranger 数据源为 Ldap

安全集群Ranger数据源默认为FusionInsight Manager Ldap用户。普通集群Ranger数据源默认为集群Unix用户。

前提条件

- 集群模式为普通模式。
- 已安装Ranger组件。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 单击“节点管理”页签，选择“节点类型”为“Master”的节点组。

步骤4 进入**Master**节点弹性云服务器页面，单击“远程登录”按钮。

步骤5 使用**root**用户登录**Master**节点，进入“ /opt/Bigdata/components/FusionInsight_HD_8.1.0.1/Ranger”目录，修改configurations.xml文件中参数“ranger.usersync.sync.source”值为“ldap”。

```
ranger.usersync.sync.source
<value model="NoSec">ldap</value>
```

说明

所有Master节点都需要修改该参数。

步骤6 在**主Master**节点执行如下命令，重启controller进程。

su - omm

sh /opt/Bigdata/om-server_8.1.0.1/om/sbin/restart-controller.sh

说明

重启controller进程会出现短暂的MRS Manager页面不可访问现象，属于正常现象，待controller启动后，Manager页面即可访问。

步骤7 登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。选择“集群 > 服务 > Ranger”。在服务“概览”页面右上角单击“更多”，选择“同步配置”。

步骤8 在Ranger实例页面，勾选“UserSync”实例，选择“更多 > 重启实例”。

步骤9 在Ranger服务“概览”页面，单击“RangerAdmin”，查看“Settings > Users/Groups/Roles”页面是否有ldap用户。

----结束

12.21.7 查看 Ranger 权限信息

查看Ranger相关权限设置信息，例如查看用户、用户组、Role。

查看 Ranger 权限信息

步骤1 使用管理员登录Ranger管理页面。

步骤2 选择“Settings > Users/Groups/Roles”，可查看系统中的用户、用户组、Roles信息。

- Users：显示Ranger从LDAP或者OS同步的所有用户信息。
- Groups：显示Ranger从LDAP或者OS同步的所有用户组、角色信息。
- Roles：显示Ranger中创建的Role信息。

说明

- 在FusionInsight Manager上创建的用户、角色、用户组会定期自动同步至Ranger，默认周期为300000毫秒（5分钟）。FusionInsight Manager中的角色和用户组在同步至Ranger后都变为用户组（Group）。只有被用户关联了的角色和用户组才会自动同步至Ranger。
- Ranger界面中创建的Role为用户或用户组的集合，用于灵活设置组件的权限访问策略，与FusionInsight Manager中的“角色”不同，请注意区分。

----结束

调整 Ranger 用户类型

步骤1 登录Ranger管理页面。

调整Ranger用户类型须使用Admin类型的用户（例如**admin**）进行操作，具体用户类型请参考**Ranger用户类型**。

步骤2 选择“Settings > Users/Groups/Roles”，在“Users”用户列表中，单击待修改类型的用户名。

步骤3 设置“Select Role”配置项为待修改的类型。

步骤4 单击“Save”。

----结束

创建 Ranger Role

管理员在设置组件的权限访问策略时，可基于用户、用户组或者Role灵活配置，其中用户与用户组信息从LDAP中自动同步，Role可手动添加。

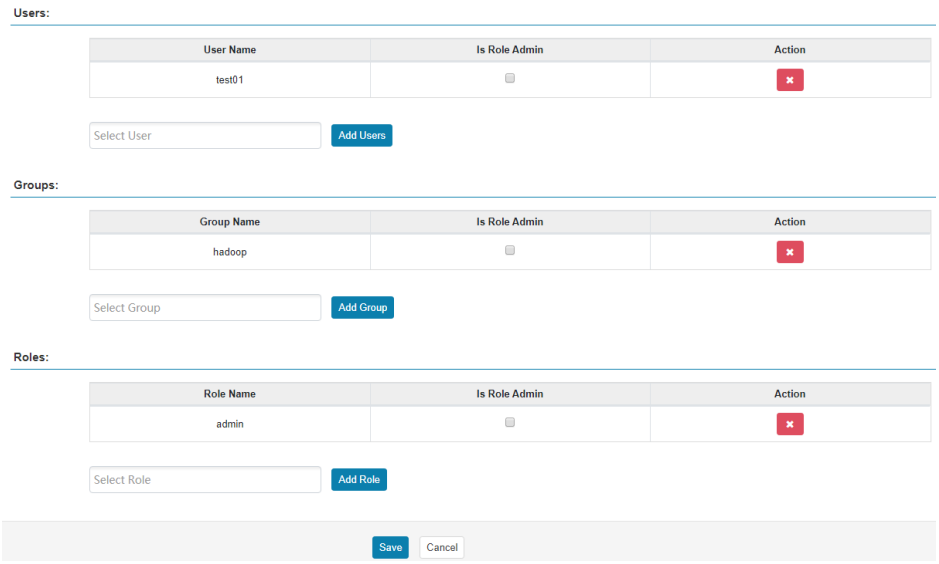
步骤1 登录Ranger管理页面。

步骤2 选择“Settings > Users/Groups/Roles > Roles > Add New Role”。

步骤3 根据界面提示填写Role的名称与描述信息。

步骤4 添加Role内需要包含的用户、用户组、子Role信息。

- 在“Users”区域，选择系统中已创建的用户，然后单击“Add Users”。
- 在“Groups”区域，选择系统中已创建的用户组，然后单击“Add Group”。
- 在“Roles”区域，选择系统中已创建的Role，然后单击“Add Role”。



步骤5 单击“Save”，Role添加成功。

----结束

12.21.8 添加 HDFS 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为HDFS用户配置HDFS目录或文件的读、写和执行权限。

前提条件

- 已安装Ranger服务且服务运行正常。
- 已创建需要配置权限的用户、用户组或Role。

操作步骤

步骤1 登录Ranger管理页面。

步骤2 在首页中单击“HDFS”区域的组件插件名称，例如“hacluster”。

步骤3 单击“Add New Policy”，添加HDFS权限控制策略。

步骤4 根据业务需求配置相关参数。

表 12-324 HDFS 权限参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，可以根据这些标签搜索报告和筛选策略。

参数名称	描述
Resource Path	资源路径，配置当前策略适用的HDFS路径文件夹或文件，可填写多个值，支持使用通配符“*”（例如“/test/*”）。 如需子目录继承上级目录权限，可打开递归开关按钮。 如果父目录开启递归，同时子目录也配置了策略，以子目录策略为准。 ● non-recursive：关闭递归 ● recursive：打开递归
Description	策略描述信息。
Audit Logging	是否审计此策略。
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外，例外条件优先级高于正常条件。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，单击“Add Permissions”，添加对应权限。 ● Read：读权限 ● Write：写权限 ● Execute：执行权限 ● Select/Deselect All：全选/取消全选 如需让当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”使这些用户或用户组成为受委托的管理员。被委托的管理员可以更新、删除本策略，还可以基于原始策略创建子策略。 如需添加多条权限控制规则，可单击  按钮添加。如需删除权限控制规则，可单击  按钮删除。 Exclude from Allow Conditions：配置排除在允许条件之外的例外规则。
Deny All Other Accesses	是否拒绝其它所有访问。 ● True：拒绝其它所有访问。 ● False：设置为false，可配置Deny Conditions。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类似，拒绝条件的优先级高于“Allow Conditions”中配置的允许条件。 Exclude from Deny Conditions：配置排除在拒绝条件之外的例外规则。

例如为用户“testuser”添加“/user/test”目录的写权限，配置如下：

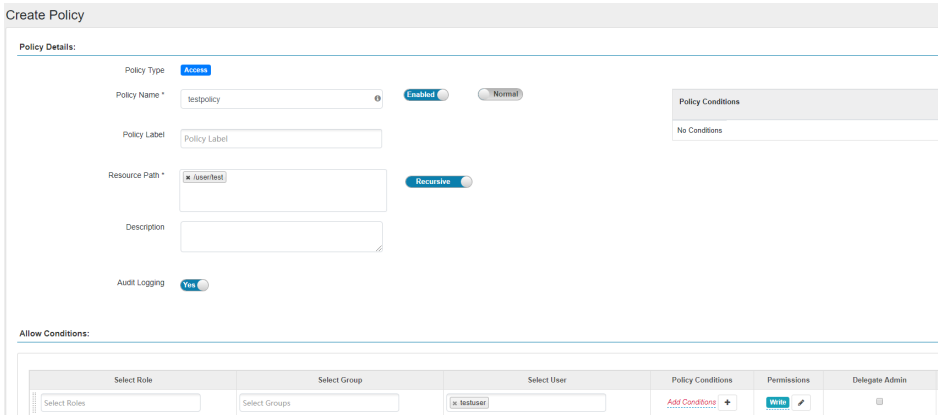


表 12-325 设置权限

任务场景	角色授权操作
设置HDFS管理员权限	1. 在首页中单击“HDFS”区域的组件插件名称，例如“hacluster”。 2. 选择“Policy Name”为“all - path”的策略，单击  按钮编辑策略。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。
设置用户执行HDFS检查和HDFS修复的权限	1. 在“Resource Path”配置文件夹或文件。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Read”和“Execute”。
设置用户读取其他用户的目录或文件的权限	1. 在“Resource Path”配置文件夹或文件。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Read”和“Execute”。
设置用户在其他用户的文件写入数据的权限	1. 在“Resource Path”配置文件夹或文件。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Write”和“Execute”。
设置用户在其他用户的目录新建或删除子文件、子目录的权限	1. 在“Resource Path”配置文件夹或文件。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Write”和“Execute”。

任务场景	角色授权操作
设置用户在其他用户的目录或文件执行的权限	1. 在“Resource Path”配置文件夹或文件。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Execute”。
设置子目录继承上级目录权限	1. 在“Resource Path”配置文件夹或文件。 2. 打开递归开关按钮，“Recursive”即为打开递归。

步骤5 （可选）添加策略有效期。在页面右上角单击“Add Validity period”，设置“Start Time”和“End Time”，选择“Time Zone”。单击“Save”保存。如需添加多条策略有效期，可单击  按钮添加。如需删除策略有效期，可单击  按钮删除。

步骤6 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如需禁用某条策略，可单击  按钮编辑策略，设置策略开关为“Disabled”。

如果不再使用策略，可单击  按钮删除策略。

----结束

12.21.9 添加 HBase 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为HBase用户配置HBase表和列族，列的权限。

前提条件

- 已安装Ranger服务且服务运行正常。
- 已创建需要配置权限的用户、用户组或Role。

操作步骤

步骤1 登录Ranger管理界面。

步骤2 在首页中单击“HBASE”区域的组件插件名称如“HBase”。

步骤3 单击“Add New Policy”，添加HBase权限控制策略。

步骤4 根据业务需求配置相关参数。

表 12-326 HBase 权限参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，可以根据这些标签搜索报告和筛选策略。
HBase Table	将适用该策略的表。 可支持通配符“*”，例如“table1:*“表示table1下的所有表。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。 **说明** Ranger界面上HBase服务插件的“hbase.rpc.protection”参数值必须和HBase服务端的“hbase.rpc.protection”参数值保持一致。具体请参考**Ranger界面添加或者修改HBase策略时，无法使用通配符搜索已存在的HBase表**。
HBase Column-family	将适用该策略的列族。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
HBase Column	将适用该策略的列。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，单击“Add Permissions”，添加对应权限。 ● Read：读权限 ● Write：写权限 ● Create：创建权限 ● Admin：管理权限 ● Select/Deselect All：全选/取消全选 如需让当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”使这些用户或用户组成为受委托的管理员。被委托的管理员可以更新、删除本策略，还可以基于原始策略创建子策略。 如需添加多条权限控制规则，可单击  按钮添加。如需删除权限控制规则，可单击  按钮删除。 Exclude from Allow Conditions：配置策略例外条件。
Deny All Other Accesses	是否拒绝其它所有访问。 ● True：拒绝其它所有访问 ● False：设置为False，可配置Deny Conditions。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类似。 拒绝条件的优先级高于“Allow Conditions”中配置的允许条件。 Exclude from Deny Conditions：配置排除在拒绝条件之外的例外规则。

表 12-327 设置权限

任务场景	角色授权操作
设置HBase管理员权限	1. 在首页中单击“HBase”区域的组件插件名称，例如“HBase”。 2. 选择“Policy Name”为“all - table, column-family, column”的策略，单击  按钮编辑策略。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。

任务场景	角色授权操作
设置用户创建表的权限	1. 在“HBase Table”配置表名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Create”。 4. 该用户具有以下操作权限： create table drop table truncate table alter table enable table flush table flush region compact disable enable desc
设置用户写入数据的权限	1. 在“HBase Table”配置表名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Write”。 4. 该用户具有put，delete，append，incr，bulkload等操作权限。
设置用户读取数据的权限	1. 在“HBase Table”配置表名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Read”。 4. 该用户具有get，scan操作权限。
设置用户管理命名空间或表的权限	1. 在“HBase Table”配置表名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Admin”。 4. 该用户具有rsgroup，peer，assign，balance等操作权限。
设置列的读取或写入权限	1. 在“HBase Table”配置表名。 2. 在“HBase Column-family”配置列族名。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Read”或者“Write”。

说明

如果用户在hbase shell中执行desc操作，需要同时给该用户赋予hbase:qouta表的读权限。

步骤5 （可选）添加策略有效期。在页面右上角单击“Add Validity period”，设置“Start Time”和“End Time”，选择“Time Zone”。单击“Save”保存。如需添加多条策略有效期，可单击  按钮添加。如需删除策略有效期，可单击  按钮删除。

步骤6 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如需禁用某条策略，可单击  按钮编辑策略，设置策略开关为“Disabled”。

如果不再使用策略，可单击  按钮删除策略。

----结束

12.21.10 添加 Hive 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为Hive用户进行相关的权限设置。Hive默认管理员帐号为hive，初始密码为Hive@123。

前提条件

- 已安装Ranger服务且服务运行正常。
- 已创建用户需要配置权限的用户、用户组或Role。
- 用户加入hive组。

操作步骤

步骤1 登录Ranger管理界面。

步骤2 在首页中单击“HADOOP SQL”区域的组件插件名称如“Hive”。

步骤3 在“Access”页签单击“Add New Policy”，添加Hive权限控制策略。

步骤4 根据业务需求配置相关参数。

表 12-328 Hive 权限参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。

参数名称	描述
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
database	将适用该策略的列Hive数据库名称。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
table	将适用该策略的Hive表名称。 如果需要添加基于UDF的策略，可切换为UDF，然后输入UDF的名称。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
Hive Column	将适用该策略的列名，填写*时表示所有列。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，然后在单击“Add Permissions”，添加对应权限。 ● select：查询权限 ● update：更新权限 ● Create：创建权限 ● Drop：drop操作权限 ● Alter：alter操作权限 ● Index：索引操作权限 ● All：所有执行权限 ● Read：可读权限 ● Write：可写权限 ● Temporary UDF Admin：临时UDF管理权限 ● Select/Deselect All：全选/取消全选 如需添加多条权限控制规则，可单击  按钮添加。 如需当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”，这些用户将成为受委托的管理员。被委托的管理员可以更新、删除本策略，它还可以基于原始策略创建子策略。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类型。

表 12-329 设置权限

任务场景	角色授权操作
role admin操作	1. 在首页中单击“Settings”，选择“Roles”。 2. 单击Role Name为admin的角色，在“Users”区域，单击“Select User”，选择对应用户名。 3. 单击Add Users按钮，在对应用户名所在行勾选“Is Role Admin”，单击“Save”保存配置。 **说明** Ranger页面的“Settings”选项只有rangeradmin用户有权限访问。用户绑定Hive管理员角色后，在每个维护操作会话中，还需要执行以下操作： 1. 以客户端安装用户，登录安装Hive客户端的节点。 2. 执行以下命令配置环境变量。 例如，Hive客户端安装目录为“/opt/hiveclient”，执行**source /opt/hiveclient/bigdata_env** 3. 执行以下命令认证用户。 **kinit** *Hive业务用户* 4. 执行以下命令登录客户端工具。 **beeline** 5. 执行以下命令更新用户的管理员权限。 **set role admin;**
创建库表操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库(如果是创建表则在“table”右侧填写或选择对应的表)，在“column”右侧填写或选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Create”。
删除库表操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库（如果是删除表则在“table”右侧填写或选择对应的表），在“column”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Drop”。
查询操作(select、desc、show)	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库(如果是表则在“table”右侧填写或选择对应的表)，在“column”右侧填写并选择对应的列（*代表所有列）。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“select”。

任务场景	角色授权操作
Alter操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库(如果是表则在“table”右侧填写或选择对应的表)，在“column”右侧填写或选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Alter”。
LOAD操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库，在“table”右侧填写或选择对应的表，在“column”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“update”。
INSERT、DELETE操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库，在“table”右侧填写或选择对应的表，在“column”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“update”。 5. 用户还需要具有Yarn任务队列的“submit”权限，权限配置参考**添加Yarn的Ranger访问权限策略**。
GRANT、REVOKE操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库，在“table”右侧填写或选择对应的表，在“column”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 勾选“Delegate Admin”。
ADD JAR操作	1. 在“Policy Name”填写策略名称。 2. 单击“database”并在下拉菜单中选择“global”。在“global”右侧填写或选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Temporary UDF Admin”。

任务场景	角色授权操作
UDF 操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库，“udf”右侧填写对应的udf 函数名。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，根据需求，给用户勾选相应权限（udf支持 Create，select， Drop）。
VIEW操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写或选择对应的数据库，在“table”右侧填写或选择对应的VIEW名称，在“column”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，参照表格上述相关操作，根据需求，给用户勾选相应权限。
dfs命令操作	执行set role admin 操作才可使用。
其他用户库表操作	1. 参照表格上述相关操作添加对应权限。 2. 给用户添加其他用户库表的HDFS路径的读、写、执行权限，详情请参考**添加HDFS的Ranger访问权限策略**。

说明

- 如果用户在执行命令时指定了HDFS路径，需要给该用户添加HDFS路径的读、写、执行权限，详情请参考**添加HDFS的Ranger访问权限策略**。也可以不配置HDFS的Ranger策略，通过之前Hive权限插件的方式，给角色添加权限，然后把角色赋予对应用户。如果HDFS Ranger策略可以匹配到Hive库表的文件或目录权限，则优先使用HDFS Ranger策略。
- Ranger策略中的URL策略是hive表存储在obs上的场景涉及，URL填写对象在obs上的完整路径。与URL联合使用的Read, Write 权限，其他场景不涉及URL策略。
- Ranger策略中global策略仅用于和Temprorary UDF Admin权限联合使用，控制UDF包的上传。
- Ranger策略中的hiveservice策略仅用于和Service Admin权限联合使用，用于控制命令：**kill query <queryId>** 结束正在执行的任务的权限。
- lock、index、refresh、replAdmin 权限暂不支持。
- 使用**show grant**命令查看表权限，表owner的grantor列统一显示为hive用户，其他用户Ranger页面赋权或后台采用grant命令赋权，则grantor显示为对应用户；若用户需要查看之前Hive权限插件的结果，可设置hive-ext.ranger.previous.privileges.enable为true后采用**show grant**查看。

步骤5 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如需禁用某条策略，可单击  按钮编辑策略，设置策略开关为“Disabled”。

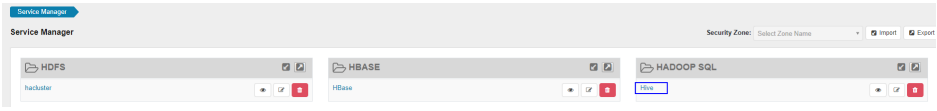
如果不再使用策略，可单击  按钮删除策略。

----结束

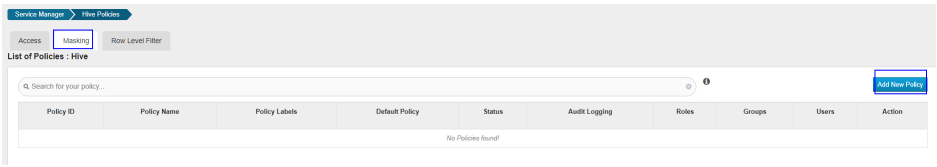
Hive 数据脱敏

Ranger支持对Hive数据进行脱敏处理（Data Masking），可对用户执行的select操作的返回结果进行处理，以屏蔽敏感信息。

步骤1 登录Ranger WebUI界面，在首页中单击“HADOOP SQL”区域的“Hive”



步骤2 在“Masking”页签单击“Add New Policy”，添加Hive权限控制策略。



步骤3 根据业务需求配置相关参数。

表 12-330 Hive 数据脱敏参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
Hive Database	配置当前策略适用的Hive中数据库名称。
Hive Table	配置当前策略适用的Hive中的表名称。
Hive Column	可添加列名。
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Mask Conditions	在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的对象，单击“Add Conditions”，添加策略适用的IP地址范围，然后在单击“Add Permissions”，勾选“select”权限。 单击“Select Masking Option”，选择数据脱敏时的处理策略： • Redact：用x屏蔽所有字母字符，用n屏蔽所有数字字符。 • Partial mask: show last 4：只显示最后的4个字符，其他用x代替。 • Partial mask: show first 4：只显示开始的4个字符，其他用x代替。 • Hash：用值的哈希值替换原值，采用的是hive的内置mask_hash函数，只对string、char、varchar类型的字段生效，其他类型的字段会返回NULL值。 • Nullify：用NULL值替换原值。 • Unmasked(retain original value)：原样显示。 • Date: show only year：仅显示日期字符串的年份部分，并将月份和日期默认为01/01。 • Custom：可使用任何有效返回与被屏蔽的列中的数据类型相同的数据类型来自定义策略。 如需添加多列的脱敏策略，可单击  按钮添加。

步骤4 单击“Add”，在策略列表可查看策略的基本信息。

步骤5 用户通过Hive客户端对配置了数据脱敏策略的表执行select操作，系统将对数据进行处理后进行展示。

说明

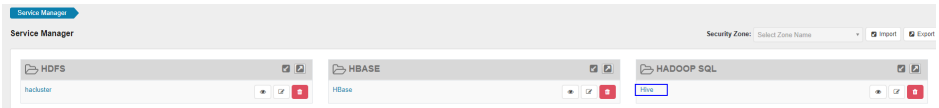
处理数据需要用户同时具有向Yarn队列提交任务的权限。

----结束

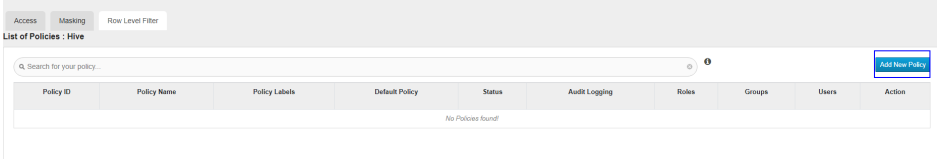
Hive 行级别数据过滤

Ranger支持用户对Hive数据表执行select操作时进行行级别的数据过滤。

步骤1 登录Ranger WebUI界面，在首页中单击“HADOOP SQL”区域的“Hive”。



步骤2 在“Row Level Filter”页签单击“Add New Policy”，添加行数据过滤策略。



步骤3 根据业务需求配置相关参数。

表 12-331 Hive 行数据过滤参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
Hive Database	配置当前策略适用的Hive中数据库名称。
Hive Table	配置当前策略适用的Hive中的表名称。
Description	策略描述信息。
Audit Logging	是否审计此策略。
Row Filter Conditions	在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的对象，单击“Add Conditions”，添加策略适用的IP地址范围，然后在单击“Add Permissions”，勾选“select”权限。 单击“Row Level Filter”，填写数据过滤规则。 例如过滤表A中“name”列“zhangsan”行的数据，过滤规则为：name <> 'zhangsan'。更多信息可参考Ranger官方文档。 如需添加更多规则，可单击  按钮添加。

步骤4 单击“Add”，在策略列表可查看策略的基本信息。

步骤5 用户通过Hive客户端对配置了数据脱敏策略的表执行select操作，系统将对数据进行处理后进行展示。

说明

处理数据需要用户同时具有向Yarn队列提交任务的权限。

----结束

12.21.11 添加 Yarn 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为Yarn用户配置Yarn管理员权限以及Yarn队列资源管理权限。

前提条件

- 已安装Ranger服务且服务运行正常。
- 已创建需要配置权限的用户、用户组或Role。

操作步骤

步骤1 登录Ranger管理界面。

步骤2 在首页中单击“YARN”区域的组件插件名称如“Yarn”。

步骤3 单击“Add New Policy”，添加Yarn权限控制策略。

步骤4 根据业务需求配置相关参数。

表 12-332 Yarn 权限参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，可以根据这些标签搜索报告和筛选策略。
Queue	队列名称，支持通配符“*”。 如需子队列继承上级队列权限，可打开递归开关按钮。 ● Non-recursive：关闭递归 ● Recursive：打开递归
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，单击“Add Permissions”，添加对应权限。 ● submit-app：提交队列任务权限 ● admin-queue：管理队列任务权限 ● Select/Deselect All：全选/取消全选 如需让当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”使这些用户成为受委托的管理员。被委托的管理员可以更新、删除本策略，它还可以基于原始策略创建子策略。 如需添加多条权限控制规则，可单击  按钮添加。如需删除权限控制规则，可单击  按钮删除。 Exclude from Allow Conditions：配置策略例外条件。
Deny All Other Accesses	是否拒绝其它所有访问。 ● True：拒绝其它所有访问 ● False：设置为False，可配置Deny Conditions。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类似。拒绝条件的优先级高于“Allow Conditions”中配置的允许条件。 Exclude from Deny Conditions：配置排除在拒绝条件之外的例外规则。

表 12-333 设置权限

任务场景	角色授权操作
设置Yarn管理员权限	1. 在首页中单击“YARN”区域的组件插件名称，例如“Yarn”。 2. 选择“Policy Name”为“all - queue”的策略，单击  按钮编辑策略。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。
设置用户在指定Yarn队列提交任务的权限	1. 在“Queue”配置队列名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“submit-app”。

任务场景	角色授权操作
设置用户在指定Yarn队列管理任务的权限	1. 在“Queue”配置队列名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“admin-queue”。

步骤5 （可选）添加策略有效期。在页面右上角单击“Add Validity period”，设置“Start Time”和“End Time”，选择“Time Zone”。单击“Save”保存。如需添加多条策略有效期，可单击  按钮添加。如需删除策略有效期，可单击  按钮删除。

步骤6 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如需禁用某条策略，可单击  按钮编辑策略，设置策略开关为“Disabled”。

如果不再使用策略，可单击  按钮删除策略。

----结束

说明

Ranger Yarn上面各个权限之间相互独立，没有语义上的包含与被包含关系。当前支持下面两种权限：

- submit-app：提交队列任务权限
- admin-queue：管理队列任务权限

虽然admin-queue也有提交任务的权限，但和submit-app权限之间并没有包含关系。

12.21.12 添加 Spark2x 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为Spark2x用户进行相关的权限设置。

说明

1. Spark2x开启或关闭Ranger鉴权后，需要重启Spark2x服务。
2. 需要重新下载客户端，或手动刷新客户端配置文件“*客户端安装目录*/Spark2x/spark/conf/spark-defaults.conf”：

   开启Ranger鉴权：spark.ranger.plugin.authorization.enable=true

   关闭Ranger鉴权：spark.ranger.plugin.authorization.enable=false
3. Spark2x中，spark-beeline（即连接到JDBCServer的应用）支持Ranger的IP过滤策略（即Ranger权限策略中的**Policy Conditions**），spark-submit与spark-sql不支持。

前提条件

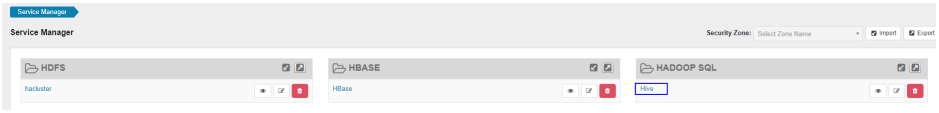
- 已安装Ranger服务且服务运行正常。
- 已启用Hive服务的Ranger鉴权功能，并且重启Hive服务后，重启了Spark2x服务。

- 已创建用户需要配置权限的用户、用户组或Role。
- 创建的用户已加入hive用户组。

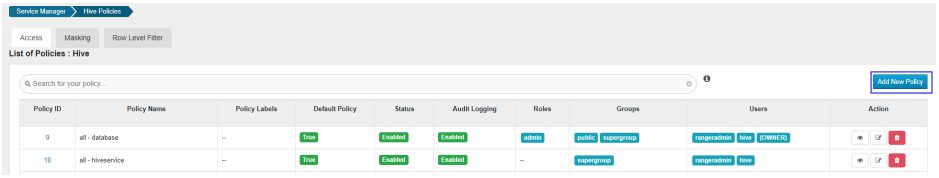
操作步骤

步骤1 登录Ranger管理界面。

步骤2 在首页中单击“HADOOP SQL”区域的组件插件名称如“Hive”。



步骤3 在“Access”页签单击“Add New Policy”，添加Spark2x权限控制策略。



步骤4 根据业务需求配置相关参数。

表 12-334 Spark2x 权限参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
database	适用该策略的Spark2x数据库名称。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
table	适用该策略的Spark2x表名称。 如果需要添加基于UDF的策略，可切换为UDF，然后输入UDF的名称。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
column	适用该策略的列名，填写*时表示所有列。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，然后在单击“Add Permissions”，添加对应权限。 ● select：查询权限 ● update：更新权限 ● Create：创建权限 ● Drop：drop操作权限 ● Alter：alter操作权限 ● Index：索引操作权限 ● All：所有执行权限 ● Read：可读权限 ● Write：可写权限 ● Temporary UDF Admin：临时UDF管理权限 ● Select/Deselect All：全选/取消全选 如需添加多条权限控制规则，可单击  按钮添加。 如需当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”，这些用户将成为受委托的管理员。被委托的管理员可以更新、删除本策略，它还可以基于原始策略创建子策略。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类型。

表 12-335 设置权限

任务场景	角色授权操作
role admin操作	1. 在首页中单击“Settings”，选择“Roles > Add New Role”。 2. 设置“Role Name”为“admin”，在“Users”区域，单击“Select User”，选择对应用户名。 3. 单击Add Users按钮，在对应用户名所在行勾选“Is Role Admin”，单击“Save”保存配置。 **说明** 用户绑定Hive管理员角色后，在每个维护操作会话中，还需要执行以下操作： 1. 以客户端安装用户，登录安装Hive客户端的节点。 2. 执行以下命令配置环境变量。 例如，Spark2x客户端安装目录为“/opt/client”，执行**source /opt/client/bigdata_env** 3. 执行以下命令认证用户。 **kinit Spark2x***业务用户* 4. 执行以下命令登录客户端工具。 **spark-beeline** 5. 执行以下命令更新用户的管理员权限。 **set role admin;**
创建库表操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库（如果是创建库，需填写将要创建的库名称，或填写“*”表示任意名称的数据库，然后选择所写名称），在“table”与“column”右侧填写并选择对应的表名称、列名称，均支持通配符（“*”）匹配。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Create”。
删除库表操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库（如果是删除库，需填写将要创建的库名称，或填写“*”表示任意名称的数据库，然后选择所写名称），在“table”与“column”右侧填写并选择对应的表名称、列名称，均支持通配符（“*”）匹配。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Drop”。 **说明** 对于CarbonData表，只有对应表的OWNER，才能执行“drop”操作。

任务场景	角色授权操作
ALTER操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库，在“table”右侧填写并选择对应的表，在“column”右侧填写并选择对应的列名称，支持通配符（“*”）匹配。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Alter”。
LOAD操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库，在“table”右侧填写并选择对应的表，在“column”右侧填写并选择对应的列名称，支持通配符（“*”）匹配。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“update”。
INSERT操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库，在“table”右侧填写并选择对应的表，在“column”右侧填写并选择对应的列名称，支持通配符（“*”）匹配。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“update”。 5. 用户还需要具有Yarn任务队列的“submit-app”权限，默认情况下，hadoop用户组具有向所有Yarn任务队列“submit-app”权限。具体配置请参考**添加Yarn的Ranger访问权限策略**。
GRANT操作	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库，在“table”右侧填写并选择对应的表，在“column”右侧填写并选择对应的列名称，支持通配符（“*”）匹配。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 勾选“Delegate Admin”。
ADD JAR操作	1. 在“Policy Name”填写策略名称。 2. 单击“database”并在下拉菜单中选择“global”。在“global”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Temporary UDF Admin”。

任务场景	角色授权操作
VIEW与INDEX权限	1. 在“Policy Name”填写策略名称。 2. “database”右侧填写并选择对应的数据库，在“table”右侧填写并选择对应的VIEW或INDEX名称，在“column”右侧填写并选择“*”。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，参照表格上述相关操作，根据需求，给用户勾选相应权限。
其他用户库表操作	1. 参照表格上述操作添加对应权限。 2. 给当前用户添加其他用户库表的HDFS路径的读、写、执行权限，具体配置请参考**添加HDFS的Ranger访问权限策略**。

说明

在Ranger上为用户添加Spark SQL的访问策略后，需要在HDFS的访问策略中添加相应的路径访问策略，否则无法访问数据文件，具体请参考**添加HDFS的Ranger访问权限策略**。

- Ranger策略中global策略仅用于联合Temprorary UDF Admin权限，用来控制UDF包的上传。
- 通过Ranger对Spark SQL进行权限控制时，不支持empower语法。

步骤5 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如果需要禁用某条策略，可单击  按钮编辑该策略，设置策略开关为“Disabled”。

如果不再使用某条策略，可单击  按钮删除该策略。

----结束

Spark2x 表数据脱敏

Ranger支持对Spark2x数据进行脱敏处理（Data Masking），可对用户执行的select操作的返回结果进行处理，以屏蔽敏感信息。

步骤1 登录Ranger WebUI界面，在首页单击“HADOOP SQL”区域的组件插件名称如“Hive”。

步骤2 在“Masking”页签单击“Add New Policy”，添加Spark2x权限控制策略。

步骤3 根据业务需求配置相关参数。

表 12-336 Spark2x 数据脱敏参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
Hive Database	配置当前策略适用的Spark2x中的数据库名称。
Hive Table	配置当前策略适用的Spark2x中的表名称。
Hive Column	配置当前策略适用的Spark2x中的列名称。
Description	策略描述信息。
Audit Logging	是否审计此策略。
Mask Conditions	在“Select Group”、“Select User”列选择已创建好的需要授予权限的用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，然后在单击“Add Permissions”，勾选“select”权限。 单击“Select Masking Option”，选择数据脱敏时的处理策略： ● Redact：用x屏蔽所有字母字符，用n屏蔽所有数字字符。 ● Partial mask: show last 4：只显示最后的4个字符。 ● Partial mask: show first 4：只显示开始的4个字符。 ● Hash：对数据进行Hash处理。 ● Nullify：用NULL值替换原值。 ● Unmasked(retain original value)：不脱敏，显示原数据。 ● Date: show only year：日期格式数据只显示年份信息。 ● Custom：可使用任何有效Hive UDF（返回与被屏蔽的列中的数据类型相同的数据类型）来自定义策略。 如需添加多列的脱敏策略，可单击  按钮添加。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类型。

----结束

Spark2x 行级别数据过滤

Ranger支持用户对Spark2x数据表执行select操作时进行行级别的数据过滤。

步骤1 登录Ranger WebUI界面，在首页单击“HADOOP SQL”区域的组件插件名称如“Hive”。

步骤2 在“Row Level Filter”页签单击“Add New Policy”，添加行数据过滤策略。

步骤3 根据业务需求配置相关参数。

表 12-337 Spark2x 行数据过滤参数

参数名称	描述
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
Hive Database	配置当前策略适用的Saprk2x中的数据库名称。
Hive Table	配置当前策略适用的Saprk2x中的表名称。
Description	策略描述信息。
Audit Logging	是否审计此策略。
Row Filter Conditions	在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的对象，单击“Add Conditions”，添加策略适用的IP地址范围，然后在单击“Add Permissions”，勾选“select”权限。 单击“Row Level Filter”，填写数据过滤规则。 例如过滤表A中“name”列“zhangsan”行的数据，过滤规则为：name <> 'zhangsan'。更多信息可参考Ranger官方文档。 如需添加更多规则，可单击  按钮添加。

步骤4 单击“Add”，在策略列表可查看策略的基本信息。

步骤5 用户通过Saprk2x客户端对配置了数据脱敏策略的表执行select操作，系统将对数据进行处理后进行展示。

----结束

12.21.13 添加 Kafka 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为Kafka用户配置Kafka主题的读、写、管理权限以及集群的管理权限，本章节以为用户“test”添加 “test” 主题的“生产”权限。

前提条件

- 已安装Ranger服务且服务运行正常。
- 已创建用户需要配置权限的用户、用户组或Role。

操作步骤

步骤1 登录Ranger管理界面。

步骤2 在首页中单击“KAFKA”区域的组件插件名称如“Kafka”。

步骤3 单击“Add New Policy”，添加Kafka权限控制策略。

步骤4 根据业务需求配置相关参数。

表 12-338 Kafka 权限参数

参数名称	描述
Policy Type	Access。
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
topic	配置当前策略适用的topic名，可以填写多个值。这里支持通配符，例如：test、test*、*。 “Include”策略适用于当前输入的对象，“Exclude”表示策略适用于除去当前输入内容之外的其他对象。
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外，例外条件优先级高于正常条件。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户。 单击“Add Conditions”，添加策略适用的IP地址范围，单击“Add Permissions”，添加对应权限。 ● Publish：生产权限。 ● Consume：消费权限。 ● Describe：查询权限。 ● Create： 创建主题权限。 ● Delete： 删除主题权限。 ● Describe Configs：查询配置权限。 ● Alter：修改topic的partition数量的权限。 ● Alter Configs：修改配置权限。 ● Select/Deselect All：全选/取消全选。 如需添加多条权限控制规则，可单击  按钮添加。 如需当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”，这些用户将成为受委托的管理员。被委托的管理员可以更新、删除本策略，它还可以基于原始策略创建子策略。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类型，拒绝条件的优先级高于“Allow Conditions”中配置的允许条件。

例如为用户“testuser”添加“test”主题的生产权限，配置如下：

图 12-42 Kafka 权限参数

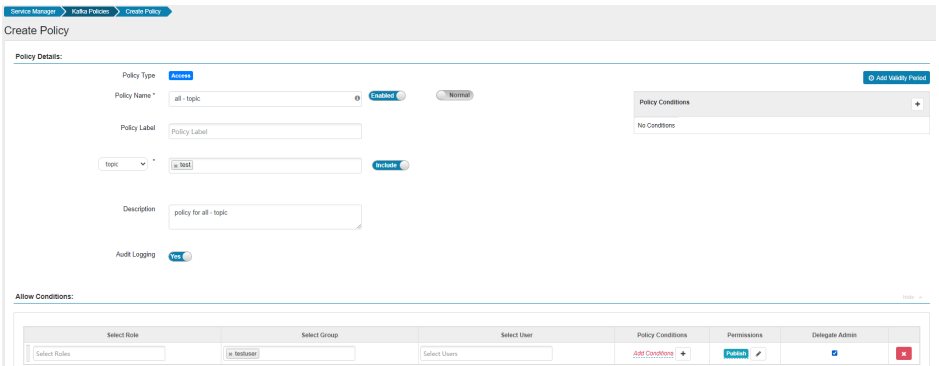


表 12-339 设置权限

任务场景	角色授权操作
设置Kafka管理员权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - topic”的策略，单击  按钮编辑策略。 3. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 4. 单击“Add Permissions”，勾选“Select/ Deselect All”。
设置用户对Topic的创建权限	1. 在“topic”配置Topic名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Create”。 **说明** 目前Kafka内核支持"--zookeeper"和"--bootstrap-server"两种方式创建Topic,社区将会在后续的版本中删掉对"--zookeeper"的支持，所以建议用户使用"--bootstrap-server"的方式创建Topic。 注意：目前Kafka只支持"--bootstrap-server"方式创建Topic行为的鉴权，不支持对"--zookeeper"方式的鉴权
设置用户对Topic的删除权限	1. 在“topic”配置Topic名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Delete”。 **说明** 目前Kafka内核支持"--zookeeper"和"--bootstrap-server"两种方式删除Topic，社区将会在后续的版本中删掉对"--zookeeper"的支持，所以建议用户使用"--bootstrap-server"的方式删除Topic。 注意：目前Kafka只支持对"--bootstrap-server"方式删除Topic行为的鉴权，不支持对"--zookeeper"方式的鉴权
设置用户对Topic的查询权限	1. 在“topic”配置Topic名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Describe”和“Describe Configs”。 **说明** 目前Kafka内核支持"--zookeeper"和"--bootstrap-server"两种方式查询Topic，社区将会在后续的版本中删掉对"--zookeeper"的支持，所以建议用户使用"--bootstrap-server"的方式查询Topic。 注意：目前Kafka只支持对"--bootstrap-server"方式查询Topic行为的鉴权，不支持对"--zookeeper"方式的鉴权

任务场景	角色授权操作
设置用户对Topic的生产权限	1. 在“topic”配置Topic名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Publish”。
设置用户对Topic的消费权限	1. 在“topic”配置Topic名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Consume”。 **说明** 因为消费Topic时，涉及到Offset的管理操作，必须同时开启ConsumerGroup的“Consume”权限，详见“设置用户对ConsumerGroup Offsets 的提交权限”
设置用户对Topic的扩容权限（增加分区）	1. 在“topic”配置Topic名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Alter”。
设置用户对Topic的配置修改权限	当前Kafka内核暂不支持基于“--bootstrap-server”的Topic参数修改行为，故当前Ranger不支持对此行为的鉴权操作。
设置用户对Cluster的所有管理权限	1. 在“cluster”右侧输入并选择集群名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Kafka Admin”。
设置用户对Cluster的创建权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - cluster”的策略，单击  按钮编辑策略。 3. 在“cluster”右侧输入并选择集群名。 4. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 5. 单击“Add Permissions”，勾选“Create”。 **说明** 对于Cluster的Create操作鉴权主要涉及以下两个场景： 1. 集群开启了“auto.create.topics.enable”参数后，客户端向服务的还未创建的Topic发送数据的场景，此时会判断用户是否有集群的Create权限 2. 对于用户创建大量Topic的场景，如果授予用户Cluster Create权限，那么该用户可以在集群内部创建任意Topic

任务场景	角色授权操作
设置用户对Cluster的配置修改权限	1. 在“cluster”右侧输入并选择集群名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Alter Configs”。 **说明** 此处的配置修改权限，指的是Broker、Broker Logger的配置权限。 当授予用户配置修改权限后，即使不授予配置查询权限也可查询配置详情（配置修改权限高于且包含配置查询权限）。
设置用户对Cluster的配置查询权限	1. 在“cluster”右侧输入并选择集群名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Describe”和“Describe Configs”。 **说明** 此处查询指的是查询集群内的Broker、Broker Logger信息。该查询不涉及Topic。
设置用户对Cluster的Idempotent Write权限	1. 在“cluster”右侧输入并选择集群名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Idempotent Write”。 **说明** 此权限会对用户客户端的Idempotent Produce行为进行鉴权。
设置用户对Cluster的分区迁移权限管理	1. 在“cluster”右侧输入并选择集群名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Alter”。 **说明** Cluster的Alter权限可以对以下三种场景进行权限控制： 1. Partition Reassign场景下，迁移副本的存储目录。 2. 集群里各分区内部leader选举。 3. Acl管理（添加或删除）。 其中**步骤4.1**和**步骤4.2**都是集群内部Controller与Broker间、Broker与Broker间的操作，创建集群时，默认授予内置kafka用户此权限，普通用户授予此权限没有意义。 **步骤4.3**涉及Acl的管理，Acl设计的就是用于鉴权，由于目前kafka鉴权已全部托管给Ranger，所以这个场景也基本不涉及（配置后亦不生效）。

任务场景	角色授权操作
设置用户对Cluster的Cluster Action权限	1. 在“cluster”右侧输入并选择集群名。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Cluster Action”。 **说明** 此权限主要对集群内部副本主从同步、节点间通信进行控制，在集群创建时已经授权给内置kakfa用户，普通用户授予此权限没有意义。
设置用户对TransactionalId的权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - transactionalid”的策略，单击  按钮编辑策略。 1. 在“transactionalid”配置事务ID。 2. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 3. 单击“Add Permissions”，勾选“Publish”和"Describe"。 **说明** “Publish”权限主要对用户开启了事务特性的客户端请求进行鉴权，例如事务开启、结束、提交offset、事务性数据生产等行为。 “Describe”权限主要对于开启事务特性的客户端与Coordinator的请求进行鉴权。 建议在开启事务特性的场景下，给用户同时授予“Publish”和“Describe”权限。
设置用户对DelegationToken的权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - delegationtoken”的策略，单击  按钮编辑策略。 3. 在“delegationtoken”配置delegationtoken。 4. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 5. 单击“Add Permissions”，勾选“ Describe”。 **说明** 当前Ranger对DelegationToken的鉴权控制仅限于对查询的权限控制，不支持对DelegationToken的create、renew、expire操作的权限控制。

任务场景	角色授权操作
设置用户对ConsumerGroup Offsets 的查询权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - consumergroup”的策略，单击  按钮编辑策略。 3. 在“consumergroup”配置需要管理的consumergroup。 4. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 5. 单击“Add Permissions”，勾选“Describe”。
设置用户对ConsumerGroup Offsets 的提交权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - consumergroup”的策略，单击  按钮编辑策略。 3. 在“consumergroup”配置需要管理的consumergroup。 4. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 5. 单击“Add Permissions”，勾选“Consume”。 **说明** 当给用户授予了ConsumerGroup的“Consume”权限后，用户会同时被授予“Describe”权限。
设置用户对ConsumerGroup Offsets 的删除权限	1. 在首页中单击“KAFKA”区域的组件插件名称，例如“Kafka”。 2. 选择“Policy Name”为“all - consumergroup”的策略，单击  按钮编辑策略。 3. 在“consumergroup”配置需要管理的consumergroup。 4. 在“Allow Conditions”区域，单击“Select User”下选择框选择用户。 5. 单击“Add Permissions”，勾选“Delete”。 **说明** 当给用户授予了ConsumerGroup的“Delete”权限后，用户会同时被授予“Describe”权限。

步骤5 （可选）添加策略有效期。在页面右上角单击“Add Validity period”，设置“Start Time”和“End Time”，选择“Time Zone”。单击“Save”保存。如需添加多条策略有效期，可单击  按钮添加。如需删除策略有效期，可单击  按钮删除。

步骤6 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如需禁用某条策略，可单击  按钮编辑策略，设置策略开关为“Disabled”。

如果不再使用策略，可单击  按钮删除策略。

----结束

12.21.14 添加 Storm 的 Ranger 访问权限策略

操作场景

管理员可通过Ranger为Storm用户进行相关的权限设置。

前提条件

- 已安装Ranger服务且服务运行正常。
- 已创建用户需要配置权限的用户、用户组或Role。
- 页面已启用Ranger鉴权开关，该按钮控制是否启用Ranger插件进行权限管控，启用则使用Ranger鉴权，否则使用组件自身鉴权机制。

操作步骤

步骤1 登录Ranger WebUI界面，在首页中单击“STORM”区域的“Storm”。

步骤2 单击“Add New Policy”，添加Storm权限控制策略。

步骤3 根据业务需求配置相关参数。

表 12-340 Storm 权限参数

参数名称	描述
Policy Conditions	IP过滤策略，可自定义，配置当前策略适用的主机节点，可填写一个或多个IP或IP段，并且IP填写支持“*”通配符，例如：192.168.1.10,192.168.1.20或者192.168.1.*。
Policy Name	策略名称，可自定义，不能与本服务内其他策略名称重复。 “include”策略适用于当前输入的对象，“exclude”表示策略适用于除去当前输入内容之外的其他对象。
Policy Label	为当前策略指定一个标签，您可以根据这些标签搜索报告和筛选策略。
Storm Topology	配置当前策略适用的拓扑名称。可以填写多个值。
Description	策略描述信息。
Audit Logging	是否审计此策略。

参数名称	描述
Allow Conditions	策略允许条件，配置本策略内允许的权限及例外。 在“Select Role”、“Select Group”、“Select User”列选择已创建好的需要授予权限的Role、用户组或用户，单击“Add Conditions”，添加策略适用的IP地址范围，单击“Add Permissions”，添加对应权限。 • Submit Topology：提交拓扑。 **说明** Submit Topology权限只有在Storm Topology为*的情况下可以赋权生效。 • File Upload：文件上传。 • File Download：文件下载。 • Kill Topology：删除拓扑。 • Rebalance：Rebalance操作权限。 • Activate：激活权限。 • Deactivate：去激活权限。 • Get Topology Conf：获取拓扑配置。 • Get Topology：获取拓扑。 • Get User Topology：获取用户拓扑。 • Get Topology Info：获取拓扑信息。 • Upload New Credential：上传新的凭证。 • Select/Deselect All：全选/取消全选。 如需添加多条权限控制规则，可单击  按钮添加。 如需当前条件中的用户或用户组管理本条策略，可勾选“Delegate Admin”，这些用户将成为受委托的管理员。被委托的管理员可以更新、删除本策略，它还可以基于原始策略创建子策略。
Deny Conditions	策略拒绝条件，配置本策略内拒绝的权限及例外，配置方法与“Allow Conditions”类似。

步骤4 （可选）添加策略有效期。在页面右上角单击“Add Validity period”，设置“Start Time”和“End Time”，选择“Time Zone”。单击“Save”保存。如需添加多条策略有效期，可单击  按钮添加。如需删除策略有效期，可单击  按钮删除。

步骤5 单击“Add”，在策略列表可查看策略的基本信息。等待策略生效后，验证相关权限是否正常。

如需禁用某条策略，可单击  按钮编辑策略，设置策略开关为“Disabled”。

如果不再使用策略，可单击  按钮删除策略。

----结束

12.21.15 Ranger 日志介绍

日志描述

日志存储路径：Ranger相关日志的默认存储路径为“/var/log/Bigdata/ranger/*角色名*”

- RangerAdmin：“/var/log/Bigdata/ranger/rangeradmin”（运行日志）。
- TagSync：“/var/log/Bigdata/ranger/tagsync”（运行日志）。
- UserSync“/var/log/Bigdata/ranger/usersync”（运行日志）。

日志归档规则：Ranger的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过20MB的时，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”，最多保留最近的20个压缩文件。

表 12-341 HDFS 日志列表

日志类型	日志文件名	描述
RangerAdmin运行日志	access_log.*<DATE>*.log	Tomcat访问日志。
	catalina.out	Tomcat服务运行日志。
	gc-worker.log	RangerAdmin的GC日志。
	postinstallDetail.log	实例安装前启动后工作日志。
	prestartDetail.log	实例启动前准备工作日志。
	ranger-admin-*<hostname>*.log	RangerAdmin运行日志。
	ranger_admin_sql-*<hostname>*.log	RangerAdmin检索DBService的日志。
	startDetail.log	实例启动日志。
TagSync运行日志	cleanupDetail.log	实例清理日志。
	gc-worker.log	实例GC日志。
	postinstallDetail.log	实例安装前启动后工作日志。
	prestartDetail.log	实例启动前准备工作日志。
	ranger-tagsync-*<hostname>*.log	TagSync运行日志。

日志类型	日志文件名	描述
	startDetail.log	实例启动日志。
	tagsync.out	TagSync的运行日志。
UserSync运行日志	auth.log	unixauth服务运行日志。
	cleanupDetail.log	实例清理日志。
	gc-worker.log	实例GC日志。
	postinstallDetail.log	实例安装前启动后工作日志。
	prestartDetail.log	实例启动前准备工作日志。
	ranger-usersync-*<hostname>*.log	USerSync运行日志。
	startDetail.log	实例启动日志。

日志级别

HDFS中提供了如表12-342所示的日志级别，日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-342 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO表示记录系统及各事件正常运行状态信息
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > 服务 > Ranger > 配置”。

步骤3 选择“全部配置”。

步骤4 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤5 选择所需修改的日志级别。

步骤6 单击“保存”，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后立即生效，不需要重启服务。

----结束

日志格式

Ranger的日志格式如下所示：

表 12-343 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-04-29 20:09:28,543 | INFO | http-bio-21401-exec-56 | Request comes from API call, skip cas filter. | CasAuthenticationFilterWrapper.java:25

12.21.16 Ranger 常见问题

12.21.16.1 安装集群过程中，Ranger 启动失败

问题

安装集群过程中，Ranger启动失败，Manager进程任务列表里打印“ERROR：cannot drop sequence X_POLICY_REF_ACCESS_TYPE_SEQ”等关于数据库信息，如何解决并正常安装Ranger?

回答

该现象可能出现在安装两个RangerAmdin实例的场景下，安装失败后，请先手动重启一个RangerAdmin，然后再逐步重启其他实例。

12.21.16.2 如何判断某个服务是否使用了 Ranger 鉴权

问题

如何判断某个支持使用Ranger鉴权的服务当前是否启用了Ranger鉴权?

回答

登录FusionInsight Manager，选择“集群 > 服务 > *服务名称*”，在服务详情页上继续单击“更多”，查看“启用Ranger鉴权”是否为可单击?

- 是，表示当前本服务未启用Ranger鉴权插件，可单击“启用Ranger鉴权”启用该功能。
- 否，表示当前本服务已启用Ranger鉴权插件，可通过Ranger管理界面配置访问该服务资源的权限策略。

12.21.16.3 新创建用户修改完密码后无法登录 Ranger

问题

使用新建用户登录Ranger页面，为什么在修改完密码后登录报401错误?

回答

由于UserSync同步用户数据有时间周期，默认是5分钟，因此在Manager上新创建的用户在用户同步成功前无法登录Ranger，因为Ranger的DB里暂时还没有该用户信息，需要等待同步周期所设置的时间后再尝试登录。

非安全模式下，由于Ranger并不从Manager同步用户数据，因此，仅有admin用户可以登录Ranger，暂时不支持其他用户登录。

12.21.16.4 Ranger 界面添加或者修改 HBase 策略时，无法使用通配符搜索已存在的 HBase 表

问题

添加HBase的Ranger访问权限策略时，在策略中使用通配符搜索已存在的HBase表时，搜索不到已存在的表，并且在/var/log/Bigdata/ranger/rangeradmin/ranger-admin-*log中报以下错误

```
Caused by: javax.security.sasl.SaslException: No common protection layer between client and server
at com.sun.security.sasl.gsskerb.GssKrb5Client.doFinalHandshake(GssKrb5Client.java:253)
at com.sun.security.sasl.gsskerb.GssKrb5Client.evaluateChallenge(GssKrb5Client.java:186)
at
org.apache.hadoop.hbase.security.AbstractHBaseSaslRpcClient.evaluateChallenge(AbstractHBaseSaslRpcClien
t.java:142)
at org.apache.hadoop.hbase.security.NettyHBaseSaslRpcClientHandler
$2.run(NettyHBaseSaslRpcClientHandler.java:142)
at org.apache.hadoop.hbase.security.NettyHBaseSaslRpcClientHandler
$2.run(NettyHBaseSaslRpcClientHandler.java:138)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1761)
at
org.apache.hadoop.hbase.security.NettyHBaseSaslRpcClientHandler.channelRead0(NettyHBaseSaslRpcClientH
andler.java:138)
at
org.apache.hadoop.hbase.security.NettyHBaseSaslRpcClientHandler.channelRead0(NettyHBaseSaslRpcClientH
andler.java:42)
at
org.apache.hbase.thirdparty.io.netty.channel.SimpleChannelInboundHandler.channelRead(SimpleChannelInbo
undHandler.java:105)
at
org.apache.hbase.thirdparty.io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractCh
annelHandlerContext.java:362)
```

回答

Ranger界面上HBase服务插件的“hbase.rpc.protection”参数值和HBase服务端的“hbase.rpc.protection”参数值必须保持一致。

步骤1 参考**登录Ranger管理界面**章节，登录Ranger管理界面。

步骤2 在首页中“HBASE”区域，单击组件插件名称，如HBase的  按钮

步骤3 搜索配置项“hbase.rpc.protection”，修改配置项的value值，与HBase服务端的“hbase.rpc.protection”的值保持一致。

步骤4 单击“保存”。

----结束

12.22 使用 Spark

12.22.1 使用前须知

本章节适用于MRS 3.x之前版本。

12.22.2 从零开始使用 Spark

本章节提供从零开始使用Spark提交sparkPi作业的操作指导，sparkPi是最经典的Spark作业，它用来计算Pi（π）值。

操作步骤

步骤1 准备sparkPi程序。

开源的Spark的样例程序包含多个例子，其中包含sparkPi。可以从**https://archive.apache.org/dist/spark/spark-2.1.0/spark-2.1.0-bin-hadoop2.7.tgz**中下载Spark的样例程序。

解压后在“spark-2.1.0-bin-hadoop2.7/examples/jars”路径下获取“spark-examples_2.11-2.1.0.jar”，即为Spark的样例程序。spark-examples_2.11-2.1.0.jar样例程序包含sparkPi程序。

步骤2 上传数据至OBS。

1. 登录OBS控制台。
2. 单击“并行文件系统 > 创建并行文件系统”，创建一个名称为sparkpi的文件系统。

   sparkpi仅为示例，文件系统名称必须全局唯一，否则会创建并行文件系统失败。其他参数分别保持默认值。
3. 单击sparkpi文件系统名称，并选择“文件”。
4. 单击“新建文件夹”，分别创建program文件夹。
5. 进入program文件夹，单击上传文件，从本地选择**步骤1**中下载的程序包，“存储类别”选择“标准存储”。

步骤3 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤4 提交sparkPi作业。

在MRS控制台选择“作业管理”，单击“添加”，进入“添加作业”页面。

- 作业类型选择“SparkSubmit”。
- 作业名称为“sparkPi”。
- 执行程序路径配置为OBS上存放程序的地址。例如：obs://sparkpi/program/spark-examples_2.11-2.1.0.jar。
- 运行程序参数选择“--class”，值填写“org.apache.spark.examples.SparkPi”。
- 执行程序参数中填写的参数为： 10。
- 服务配置参数无需填写。

只有集群处于“运行中”状态时才能提交作业。

作业提交成功后默认为“已接受”状态，不需要用户手动执行作业。

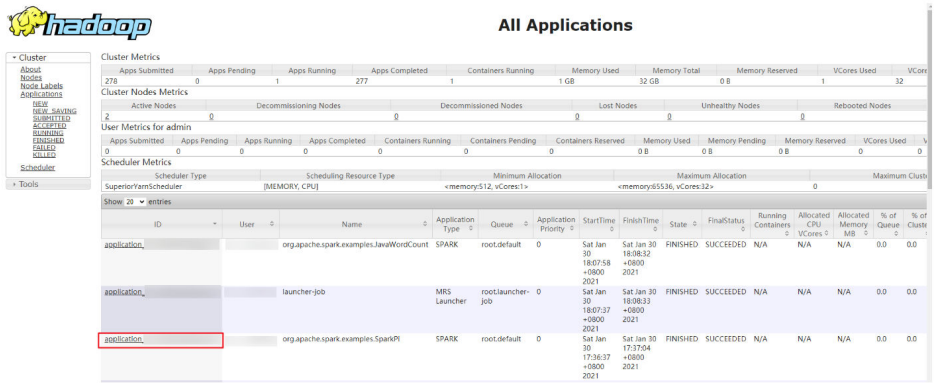
步骤5 查看作业执行结果。

1. 进入“作业管理”页面，查看作业是否执行完成。

   作业运行需要时间，作业运行结束后，刷新作业列表。

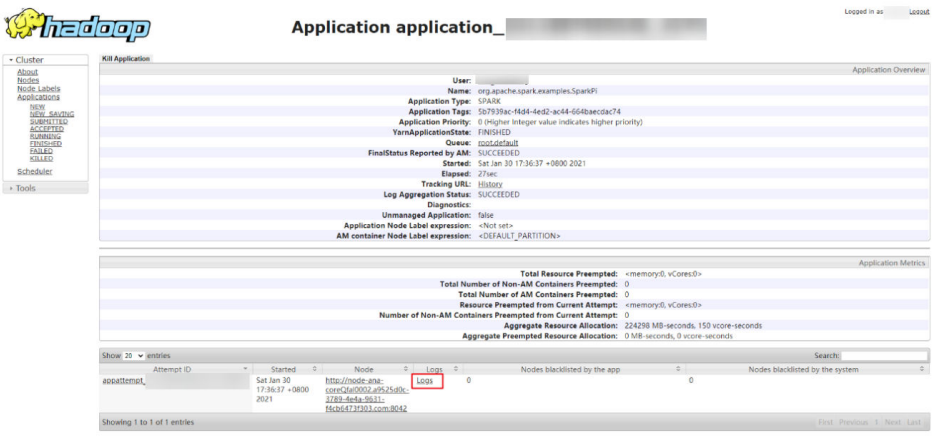
   作业执行成功或失败后都不能再次执行，只能新增作业，配置作业参数后重新提交作业。
2. 进入Yarn原生界面，查看作业输出信息。
   a. 进入“作业管理”页面，单击对应作业所在行“操作”列的“查看详情”，获取“作业实际编号”。
   b. 登录Manager页面，选择“服务管理 > Yarn > ResourceManager WebUI > ResourceManager (主)”进入Yarn界面。
   c. 单击“作业实际编号”对应ID。

   图 12-43 Yarn 界面

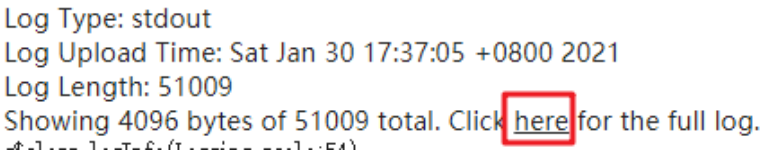
   d. 单击作业日志中的“Logs”。

图 12-44 sparkPi 作业日志



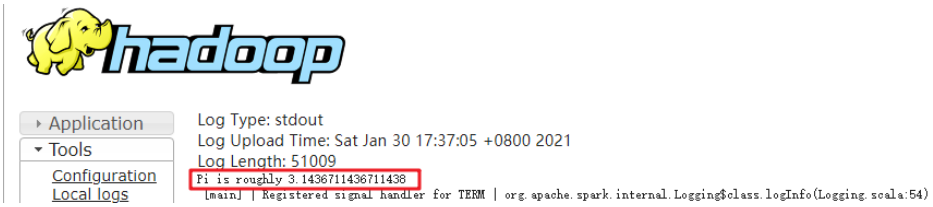
e. 单击“here”获取更详细日志。

图 12-45 sparkPi 作业更详细日志



f. 获取作业执行结果。

图 12-46 sparkPi 作业执行结果



----结束

12.22.3 从零开始使用 Spark SQL

Spark提供类似SQL的Spark SQL语言操作结构化数据，本章节提供从零开始使用Spark SQL，创建一个名称为src_data的表，然后在src_data表中每行写入一条数据，最后将数据存储在“mrs_20160907”集群中。再使用SQL语句查询src_data表中的数据，最后可将src_data表删除。

前提条件

将OBS数据源中的数据写入Spark SQL表中时，需要先获取AK/SK。获取方法如下：

1. 登录管理控制台。
2. 单击用户名，在下拉列表中单击“我的凭证”。
3. 单击“访问密钥”。
4. 单击“新增访问密钥”，进入“新增访问密钥”页面。

5. 输入登录密码和，单击“确定”，下载密钥，请妥善保。

操作步骤

步骤1 准备使用Spark SQL分析的数据源。

样例txt文件如下：

```
abcd3ghji
efgh658ko
1234jjyu9
7h8kodfg1
kk99icxz3
```

步骤2 上传数据至OBS。

1. 登录OBS控制台。
2. 单击“并行文件系统 > 创建并行文件系统”，创建一个名称为sparksql的文件系统。

   sparksql仅为示例，文件系统名称必须全局唯一，否则会创建并行文件系统失败。
3. 单击sparksql文件系统名称，并选择“文件”。
4. 单击“新建文件夹”，创建input文件夹。
5. 进入input文件夹，单击“上传文件 > 添加文件”，选择本地的txt文件，然后单击“上传”。

步骤3 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤4 将OBS中的txt文件导入至HDFS中。

1. 选择“文件管理”。
2. 在“HDFS文件列表”页签中单击“新建”，创建一个名称为userinput的文件夹。
3. 进入userinput文件夹，单击“导入数据”。
4. 选择OBS和HDFS路径，单击“确定”。

   OBS路径：obs://sparksql/input/sparksql-test.txt

   HDFS路径：/user/userinput

步骤5 提交Spark SQL语句。

1. 在MRS控制台选择“作业管理”。

   只有“mrs_20160907”集群处于“运行中”状态时才能提交Spark SQL语句。
2. 输入创建表的Spark SQL语句。

   输入Spark SQL语句时，总字符数应当小于或等于10000字符，否则会提交语句失败。

   语法格式：

   CREATE [EXTERNAL] **TABLE** [IF NOT EXISTS] *table_name* [(col_name data_type [COMMENT col_comment], ...)] [COMMENT table_comment] [PARTITIONED **BY** (col_name data_type [COMMENT col_comment], ...)] [CLUSTERED **BY** (col_name, col_name, ...) [SORTED **BY** (col_name [ASC|DESC], ...)] INTO num_buckets BUCKETS] [ROW FORMAT row_format] [STORED **AS** file_format] [LOCATION hdfs_path];

   创建表样例存在以下两种方式。

- 方式一：创建一个src_data表，将数据源中的数据一行一行写入src_data表中。
  - 数据源存储在HDFS的文件夹下：***create external table*** *src_data****(line string) row format delimited fields terminated by '\|\|n' stored as textfile location*** *'/user/userinput'*;
  - 数据源存储在OBS的“/sparksql/input”文件夹下：**create external table** *src_data***(line string) row format delimited fields terminated by '\\n' stored as textfile location** *'obs://AK:SK@sparksql/input'*;

    AK/SK获取方法，请参见**前提条件**。
- 方式二：创建一个表src_data1，将数据源中的数据批量load到src_data1表中。

  create table *src_data1* **(line string) row format delimited fields terminated by ','** ;

  load data inpath *'/user/userinput/sparksql-test.txt'* **into table** *src_data1*;

  说明

  采用方式二时，只能将HDFS上的数据load到新建的表中，OBS上的数据不支持直接load到新建的表中。

3. 输入查询表的Spark SQL语句。

   语法格式：

   SELECT col_name **FROM** *table_name*;

   查询表样例，查询src_data表中的所有数据：

   select * from src_data;
4. 输入删除表的Spark SQL语句。

   语法格式：

   DROP TABLE [IF EXISTS] *table_name*;

   删除表样例：

   drop table src_data;
5. 单击“检查”，检查输入语句的语法是否正确。
6. 单击“确定”。

   Spark SQL语句提交后，是否执行成功会在“执行结果”列中展示。

步骤6 删除集群。

----结束

12.22.4 使用 Spark 客户端

MRS集群创建完成后，可以通过客户端去创建和提交作业。客户端可以安装在集群内部节点或集群外部节点上：

- 集群内部节点：MRS集群创建完成后，集群内的master和core节点默认已经安装好客户端，详情请参见章节，登录安装客户端的节点。
- 集群外部节点：用户可以将客户端安装在集群外部节点上，详情请参见章节，登录安装客户端的节点。

使用 Spark 客户端

步骤1 请根据客户端所在位置，参考，或者章节，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行Spark Shell命令。例如：

spark-beeline

----结束

12.22.5 访问 Spark Web UI 界面

Spark Web UI界面主要用于查看Spark应用程序运行情况，推荐使用Google chrome浏览器以获得更好的体验。

Spark主要有两个Web页面。

- Spark UI页面，用于展示正在执行的应用的运行情况。

  页面主要包括了Jobs、Stages、Storage、Environment、Executors、SQL、JDBC/ODBC Server等部分。Streaming应用会多一个Streaming标签页。
- History Server页面，用于展示已经完成的和未完成的Spark应用的运行情况。

  页面包括了应用ID、应用名称、开始时间、结束时间、执行时间、所属用户等信息。

Spark UI

步骤1 进入组件管理页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 ”。

步骤2 选择“Yarn”并在“Yarn 概述”中“ResourceManager Web UI”中单击“ResourceManager Web UI”对应的“ResourceManager”进入Web界面。

步骤3 查找到对应的Spark应用程序，单击应用信息的最后一列“ApplicationMaster”，即可进入Spark UI页面。

图 12-47 ApplicationMaster

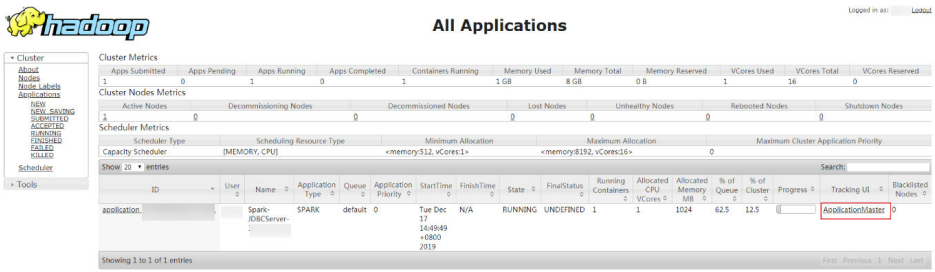


图 12-48 Spark UI 页面

----结束

History Server

步骤1 进入组件管理页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理”。

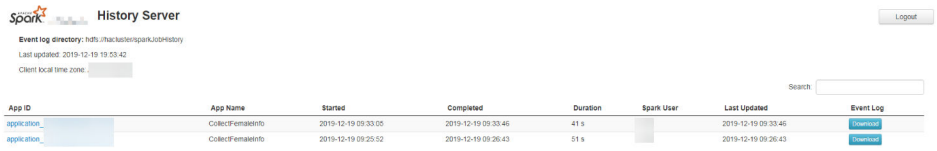
  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务”。

步骤2 选择“Spark”并在“Spark 概述”中“Spark Web UI”中单击“Spark Web UI”对应的“JobHistory”进入Web界面。

图 12-49 Spark History Server



----结束

12.22.6 Spark 对接 OpenTSDB

12.22.6.1 创建表关联 OpenTSDB

功能描述

MRS的Spark实现了访问OpenTSDB的Datasource，能够在Spark中创建关联表，查询和插入OpenTSDB数据。

使用CREATE TABLE命令创建表并关联OpenTSDB上已有的metric。

说明

若OpenTSDB上不存在metric，查询对应的表会报错。

语法格式

```
CREATE TABLE [IF NOT EXISTS] OPENTSDB_TABLE_NAME   USING OPENTSDB OPTIONS (
'metric' = 'METRIC_NAME',
'tags' = 'TAG1,TAG2'
);
```

关键字

参数	描述
metric	所创建的表对应的OpenTSDB中的指标名称。
tags	metric对应的标签，用于归类、过滤、快速检索等操作。可以是1个到8个，以“,”分隔，包括对应metric下所有tagk的值。

注意事项

创建表时，不需要指定timestamp和value字段，系统会根据指定的tags自动构建字段，包含以下字段，其中TAG1和TAG2由tags指定。

- TAG1 String
- TAG2 String
- timestamp Timestamp
- value double

示例

创建opentsdb_table表并关联到OpenTSDB组件的city.temp这个metric。

```
CREATE table opentsdb_table using opentsdb OPTIONS ('metric'='city.temp',  'tags'='city,location');
```

12.22.6.2 插入数据至 OpenTSDB 表

功能描述

使用INSERT INTO命令将表中的数据插入到已关联的OpenTSDB metric中。

语法格式

```
INSERT INTO TABLE_NAME SELECT * FROM SRC_TABLE;
INSERT INTO TABLE_NAME VALUES(XXX);
```

关键字

参数	描述
TABLE_NAME	所关联的OpenTSDB表名。
SRC_TABLE	获取数据的表名，普通表即可。

注意事项

- 插入的数据不能为null；插入的数据相同，会覆盖原数据；插入的数据只有value值不同，也会覆盖原数据。
- 不支持INSERT OVERWRITE语法。
- 不建议对同一张表并发插入数据，因为有一定概率发生并发冲突，导致插入失败。
- 时间戳格式只支持yyyy-MM-dd hh:mm:ss。

示例

在opentsdb_table表中插入数据。

```
insert into opentsdb_table values('city1','city2','2018-05-03 00:00:00',21);
```

12.22.6.3 查询 OpenTSDB 表

SELECT命令用于查询OpenTSDB表中的数据。

语法格式

```
SELECT * FROM table_name WHERE tagk=tagv LIMIT number;
```

关键字

参数	描述
LIMIT	对查询结果进行限制。
number	参数仅支持INT类型。

注意事项

- 所查询的表必须是已经存在的表，否则会出错。
- 查询的tagv必须是已经有的值，否则会出错。

示例

查询表opentsdb_table中的数据。

```
SELECT * FROM opentsdb_table LIMIT 100;
SELECT * FROM opentsdb_table WHERE city='city1';
```

12.22.6.4 默认配置修改

默认会连接Spark的Executor所在节点本地的TSD进程，在MRS中一般使用默认配置即可，无需修改。

表 12-344 OpenTSDB 数据源相关配置

配置名	描述	样例值
spark.sql.datasource.opentsdb.host	连接的TSD进程地址	空（默认值） xx.xx.xx.xx，多个地址间用英文逗号间隔。
spark.sql.datasource.opentsdb.port	TSD进程端口号	4242（默认值）
spark.sql.datasource.opentsdb.randomSeed	当spark.sql.datasource.opentsdb.host配置多个地址时，是否使用随机种子。配置为否时，所有在相同节点的executor会连接相同的host，这样可以配合spark.blacklist.enabled=true来实现Task容错。	false（默认）

示例

在spark-sql，spark-beeline执行set语句后，再执行其他SQL：

```
set spark.sql.datasource.opentsdb.host = 192.168.2.143,192.168.2.158;
SELECT * FROM opentsdb_table ;
```

12.23 使用 Spark2x

12.23.1 使用前须知

本章节适用于MRS 3.x及后续版本。

12.23.2 基本操作

12.23.2.1 快速入门

本章节提供从零开始使用Spark2x提交spark应用程序，包括Spark Core及Spark SQL。其中，Spark Core为Spark的内核模块，主要负责任务的执行，用于编写spark应用程序；Spark SQL为执行SQL的模块。

场景说明

假定用户有某个周末网民网购停留时间的日志文本，基于某些业务要求，要求开发Spark应用程序实现如下要求：

- 统计日志文件中本周末网购停留总时间超过2个小时的女性网民信息。
- 周末两天的日志文件第一列为姓名，第二列为性别，第三列为本次停留时间，单位为分钟，分隔符为“,”。

log1.txt：周六网民停留日志

```
LiuYang,female,20
YuanJing,male,10
GuoYijun,male,5
CaiXuyu,female,50
Liyuan,male,20
FangBo,female,50
LiuYang,female,20
YuanJing,male,10
GuoYijun,male,50
CaiXuyu,female,50
FangBo,female,60
```

log2.txt：周日网民停留日志

```
LiuYang,female,20
YuanJing,male,10
CaiXuyu,female,50
FangBo,female,50
GuoYijun,male,5
CaiXuyu,female,50
Liyuan,male,20
CaiXuyu,female,50
FangBo,female,50
LiuYang,female,20
YuanJing,male,10
FangBo,female,50
GuoYijun,male,50
CaiXuyu,female,50
FangBo,female,60
```

前提条件

- 在Manager界面创建用户并开通其HDFS、YARN、Kafka和Hive权限。
- 根据所用的开发语言安装并配置IntelliJ IDEA及JDK等工具。
- 已完成Spark2x客户端的安装及客户端网络连接的配置。
- 对于Spark SQL程序，需要先在客户端启动Spark SQL或Beeline以输入SQL语句。

操作步骤

步骤1 获取样例工程并将其导入IDEA，导入样例工程依赖jar包。通过IDEA配置并生成jar包。

步骤2 准备样例工程所需数据。

将场景说明中的原日志文件放置在HDFS系统中。

1. 本地新建两个文本文件，分别将log1.txt及log2.txt中的内容复制保存到input_data1.txt和input_data2.txt。
2. 在HDFS上建立一个文件夹“/tmp/input”，并上传input_data1.txt、input_data2.txt到此目录。

步骤3 将生成的jar包上传至Spark2x运行环境下（Spark2x客户端），如“/opt/female”。

步骤4 进入客户端目录，执行以下命令加载环境变量并登录。若安装了Spark2x多实例或者同时安装了Spark和Spark2x，在使用客户端连接具体实例时，请执行以下命令加载具体实例的环境变量。

source bigdata_env

source Spark2x/component_env

kinit <用于认证的业务用户>

步骤5 在bin目录下调用以下脚本提交Spark应用程序。

spark-submit --class *com.xxx.bigdata.spark.examples.FemaleInfoCollection* ***--master yarn-client*** */opt/female/FemaleInfoCollection.jar <inputPath>*

说明

- FemaleInfoCollection.jar为**步骤1**生成的jar包。
- <inputPath>是**步骤2.2**创建的目录。

步骤6 （可选）在bin目录下调用**spark-sql**或**spark-beeline**脚本后便可直接输入SQL语句执行查询等操作。

如创建一个表，插入一条数据再对表进行查询。

```
spark-sql> CREATE TABLE TEST(NAME STRING, AGE INT);
Time taken: 0.348 seconds
spark-sql>INSERT INTO TEST VALUES('Jack', 20);
Time taken: 1.13 seconds
spark-sql> SELECT * FROM TEST;
Jack	20
Time taken: 0.18 seconds, Fetched 1 row(s)
```

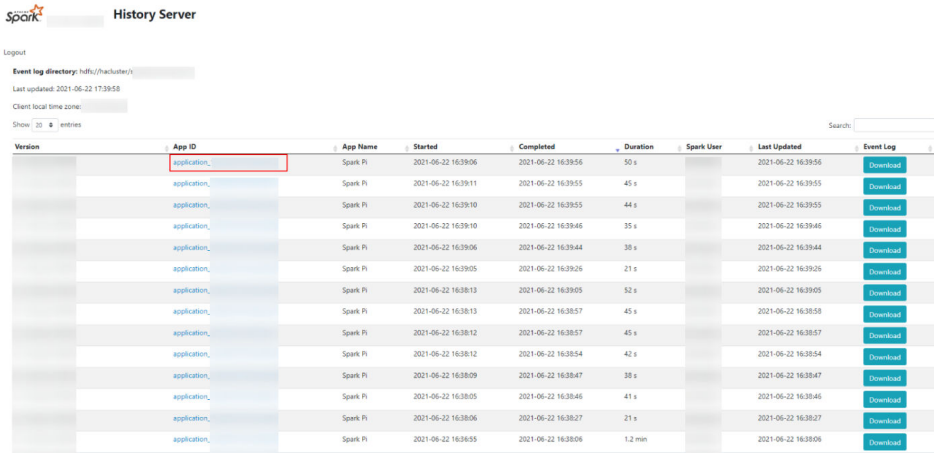
步骤7 查看Spark应用运行结果。

- 通过指定文件查看运行结果数据。

  结果数据的存储路径和格式由Spark应用程序指定。
- 通过Web页面查看运行情况。
  a. 登录Manager主页面。在服务中选择Spark2x。
  b. 进入Spark2x概览页面，单击SparkWebUI任意一个实例，如JobHistory2x(host2)。
  c. 进入History Server页面。

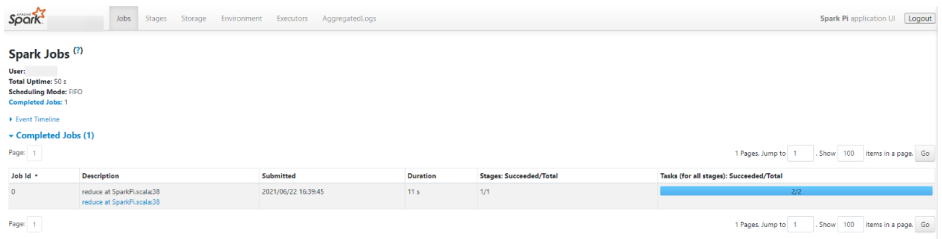
     History Server页面用于展示已完成和未完成的应用的运行情况。

图 12-50 History Server 页面



d. 选择一个应用ID，单击此页面将跳转到该应用的Spark UI页面。
Spark UI页面，用于展示正在执行的应用的运行情况。

图 12-51 Spark UI 页面



- 通过查看Spark日志获取应用运行情况。
  通过查看**Spark2x日志介绍**了解应用运行情况，并根据日志信息调整应用程序。

----结束

12.23.2.2 快速配置参数

概述

本节介绍Spark2x使用过程中快速配置常用参数和不建议修改的配置参数。

快速配置常用参数

其他参数在安装集群时已进行了适配，以下参数需要根据使用场景进行调整。以下参数除特别指出外，一般在Spark2x客户端的“spark-defaults.conf”文件中配置。

表 12-345 快速配置常用参数

配置项	说明	默认值
spark.sql.parquet.compression.codec	对于非分区parquet表，设置其存储文件的压缩格式。 在JDBCServer服务端的“spark-defaults.conf”配置文件中进行设置。	snappy

配置项	说明	默认值
spark.dynamicAllocation.enabled	是否使用动态资源调度，用于根据规模调整注册于该应用的executor的数量。目前仅在YARN模式下有效。 JDBCServer默认值为true，client默认值为false。	false
spark.executor.memory	每个Executor进程使用的内存数量，与JVM内存设置字符串的格式相同（例如：512m，2g）。	4G
spark.sql.autoBroadcastJoinThreshold	当进行join操作时，配置广播的最大值。 ● 当SQL语句中涉及的表中相应字段的大小小于该值时，进行广播。 ● 配置为-1时，将不进行广播。	10485760
spark.yarn.queue	JDBCServer服务所在的Yarn队列。 在JDBCServer服务端的“spark-defaults.conf”配置文件中进行设置。	default
spark.driver.memory	大集群下推荐配置32~64g驱动程序进程使用的内存数量，即SparkContext初始化的进程（例如：512m, 2g）。	4G
spark.yarn.security.credentials.hbase.enabled	是否打开获取HBase token的功能。如果需要Spark-on-HBase功能，并且配置了安全集群，参数值设置为“true”。否则设置为“false”。	false
spark.serializer	用于串行化将通过网络发送或需要缓存的对象的类以序列化形式展现。 Java序列化的默认值适用于任何Serializable Java对象，但运行速度相当慢，所以建议使用org.apache.spark.serializer.KryoSerializer并配置Kryo序列化。可以是org.apache.spark.serializer.Serializer的任何子类。	org.apache.spark.serializer.JavaSerializer
spark.executor.cores	每个执行者使用的内核个数。 在独立模式和Mesos粗粒度模式下设置此参数。当有足够多的内核时，允许应用程序在同样的worker上执行多个执行程序；否则，在每个worker上，每个应用程序只能运行一个执行程序。	1
spark.shuffle.service.enabled	NodeManager中一个长期运行的辅助服务，用于提升Shuffle计算性能。	fasle
spark.sql.adaptive.enabled	是否开启自适应执行框架。	false

配置项	说明	默认值
spark.executor.memoryOverhead	每个执行器要分配的堆内存量（单位为兆字节）。 这是占用虚拟机开销的内存，类似于内部字符串，其他内置开销等等。会随着执行器大小（通常为6-10%）而增长。	1GB
spark.streaming.kafka.direct.lifo	配置是否开启Kafka后进先出功能。	false

不建议修改的参数

以下参数在安装集群时已进行了适配，不建议用户进行修改。

表 12-346 不建议修改的参数说明

配置项	说明	默认值或配置示例
spark.password.factory	用于选择密钥解析方式。	org.apache.spark.om.util.FIPasswordFactory
spark.ssl.ui.protocol	配置ui的ssl协议。	TLSv1.2
spark.yarn.archive	Spark jars的存档，用于分发到YARN缓存。如果设置，此配置值将替换<code>spark.yarn.jars </code>，并存档在所有应用程序的容器中使用。存档应包含其根目录中的jar文件。与以前的选项一样，存档也可以在HDFS上托管，用来加快文件分发速度。	hdfs://hacluster/user/spark2x/jars/8.1.0.1/spark-archive-2x.zip **说明** 此处版本号8.1.0.1为示例，具体以实际环境的版本号为准。
spark.yarn.am.extraJavaOptions	在Client模式下传递至YARN Application Master的一系列额外JVM选项。在Cluster模式下使用“spark.driver.extraJavaOptions”。	-Dlog4j.configuration=./__spark_conf__/__hadoop_conf__/log4j-executor.properties -Djava.security.auth.login.config=./__spark_conf__/__hadoop_conf__/jaas-zk.conf -Dzookeeper.server.principal=zookeeper/hadoop.*<系统域名>* -Djava.security.krb5.conf=./__spark_conf__/__hadoop_conf__/kdc.conf -Djdk.tls.ephemeralDHKeySize=2048

配置项	说明	默认值或配置示例
spark.shuffle.servicev2.port	Shuffle服务监听数据获取请求的端口。	27338
spark.ssl.historyServer.enabled	配置history server是否使用SSL。	true
spark.files.overwrite	当目标文件存在时，且其内容与源的文件不匹配。是否覆盖通过SparkContext.addFile()添加的文件。	false
spark.yarn.cluster.driver.extraClassPath	YARN-Cluster模式下，Driver使用的extraClassPath，配置为服务端的路径和参数。	${BIGDATA_HOME}/common/runtime/security
spark.driver.extraClassPath	附加至driver的classpath的额外classpath条目。	${BIGDATA_HOME}/common/runtime/security
spark.yarn.dist.innerfiles	配置YARN模式下Spark内部需要上传到HDFS的文件。	/Spark_path/spark/conf/s3p.file,/Spark_path/spark/conf/locals3.jceks *Spark_path*为Spark客户端的安装路径。
spark.sql.bigdata.register.dialect	用于注册sql解析器。	org.apache.spark.sql.hbase.HBaseSQLParser
spark.shuffle.manager	处理数据的方式。有两种实现方式可用：sort和hash。sort shuffle对内存的使用率更高，是Spark 1.2及后续版本的默认选项。	SORT
spark.deploy.zookeeper.url	Zookeeper的地址，多个地址以逗号隔开。	For example: host1:2181,host2:2181,host3:2181
spark.broadcast.factory	使用的广播方式。	org.apache.spark.broadcast.TorrentBroadcastFactory
spark.sql.session.state.builder	指定会话状态构造器。	org.apache.spark.sql.hive.FIHiveACLSessionStateBuilder

配置项	说明	默认值或配置示例
spark.executor.extraLibraryPath	设置启动executor JVM时所使用的特殊的library path。	${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-Hadoop-3.1.1/hadoop/lib/native
spark.ui.customErrorPage	配置网页有错误时是否允许显示自定义的错误信息页面。	true
spark.httpdProxy.enable	配置是否使用httpd代理。	true
spark.ssl.ui.enabledAlgorithms	配置ui ssl算法。	TLS_ECDHE_ECDSA_WITH_AES_256_GCM_SHA384,TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384,TLS_ECDHE_ECDSA_WITH_AES_128_GCM_SHA256,TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256,TLS_DHE_RSA_WITH_AES_256_GCM_SHA384,TLS_DHE_DSS_WITH_AES_256_GCM_SHA384,TLS_DHE_RSA_WITH_AES_128_GCM_SHA256,TLS_DHE_DSS_WITH_AES_128_GCM_SHA256
spark.ui.logout.enabled	针对Spark组件的WebUI，设置logout按钮。	true
spark.security.hideInfo.enabled	配置UI界面是否隐藏敏感信息。	true
spark.yarn.cluster.driver.extraLibraryPath	YARN-Cluster模式下driver的extraLibraryPath，配置成服务端的路径和参数。	${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-Hadoop-3.1.1/hadoop/lib/native
spark.driver.extraLibraryPath	设置一个特殊的library path在启动驱动程序JVM时使用。	${DATA_NODE_INSTALL_HOME}/hadoop/lib/native
spark.ui.killEnabled	允许停止Web UI中的stage和相应的job。	true
spark.yarn.access.hadoopFileSystems	Spark可以访问多个NameService。有多个NameService时，需要把所使用的NameService都配置进该配置项，之间以逗号分隔。	hdfs://hacluster,hdfs://hacluster

配置项	说明	默认值或配置示例
spark.yarn.cluster.driver.extraJavaOptions	传递至Executor的额外JVM选项。例如，GC设置或其他日志记录。请注意不能通过此选项设置Spark属性或heap大小。Spark属性应该使用SparkConf对象或调用spark-submit脚本时指定的spark-defaults.conf文件来设置。Heap大小可以通过spark.executor.memory来设置。	-Xloggc:<LOG_DIR>/gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=10M -Dlog4j.configuration=./__spark_conf__/__hadoop_conf__/log4j-executor.properties -Djava.security.auth.login.config=./__spark_conf__/__hadoop_conf__/jaas-zk.conf -Dzookeeper.server.principal=zookeeper/hadoop.*<系统域名>* -Djava.security.krb5.conf=./__spark_conf__/__hadoop_conf__/kdc.conf -Djetty.version=x.y.z -Dorg.xerial.snappy.tempdir=${BIGDATA_HOME}/tmp/spark2x_app -Dcarbon.properties.filepath=./__spark_conf__/__hadoop_conf__/carbon.properties -Djdk.tls.ephemeralDHKeySize=2048
spark.driver.extraJavaOptions	传递至driver（驱动程序）的一系列额外JVM选项。	-Xloggc:${SPARK_LOG_DIR}/indexserver-omm-%p-gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:MaxDirectMemorySize=512M -XX:MaxMetaspaceSize=512M -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=10M -XX:OnOutOfMemoryError='kill -9 %p' -Djetty.version=x.y.z -Dorg.xerial.snappy.tempdir=${BIGDATA_HOME}/tmp/spark2x/JDBCServer/snappy_tmp -Djava.io.tmpdir=${BIGDATA_HOME}/tmp/spark2x/JDBCServer/io_tmp -Dcarbon.properties.filepath=${SPARK_CONF_DIR}/carbon.properties -Djdk.tls.ephemeralDHKeySize=2048 -Dspark.ssl.keyStore=${SPARK_CONF_DIR}/child.keystore #{java_stack_prefer}
spark.eventLog.overwrite	是否覆盖任何现有的文件。	false

配置项	说明	默认值或配置示例
spark.eventLog.dir	如果**spark.eventLog.enabled**为**true**，记录Spark事件的目录。在此目录下，Spark为每个应用程序创建文件，并将应用程序的事件记录到文件中。用户也可设置为统一的与HDFS目录相似的地址，这样History server就可以读取历史文件。	hdfs://hacluster/spark2xJobHistory2x
spark.random.port.min	设置随机端口的最小值。	22600
spark.authenticate	是否Spark认证其内部连接。如果不是运行在YARN上，请参见spark.authenticate.secret的相关内容。	true
spark.random.port.max	设置随机端口的最大值。	22899
spark.eventLog.enabled	是否记录Spark事件，用于应用程序在完成后重构webUI。	true

配置项	说明	默认值或配置示例
spark.executor.extraJavaOptions	传递至Executor的额外JVM选项。例如，GC设置或其他日志记录。请注意不能通过此选项设置Spark属性或heap大小。	-Xloggc:<LOG_DIR>/gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=10M -Dlog4j.configuration=./log4j-executor.properties -Djava.security.auth.login.config=./jaas-zk.conf -Dzookeeper.server.principal=zookeeper/hadoop.*<系统域名>* -Djava.security.krb5.conf=./kdc.conf -Dcarbon.properties.filepath=./carbon.properties -Xloggc:<LOG_DIR>/gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=10M -Dlog4j.configuration=./__spark_conf__/__hadoop_conf__/log4j-executor.properties -Djava.security.auth.login.config=./__spark_conf__/__hadoop_conf__/jaas-zk.conf -Dzookeeper.server.principal=zookeeper/hadoop.*<系统域名>* -Djava.security.krb5.conf=./__spark_conf__/__hadoop_conf__/kdc.conf -Dcarbon.properties.filepath=./__spark_conf__/__hadoop_conf__/carbon.properties -Djdk.tls.ephemeralDHKeySize=2048
spark.sql.authorization.enabled	配置Hive client是否开启认证。	true

12.23.2.3 常用参数

概述

本节介绍Spark使用过程中的常用配置项。以特性为基础划分子章节，以便用户快速搜索到相应的配置项。如果用户使用MRS集群，本节介绍的参数大部分已经适配好，用户无需再进行配置。少数需要用户根据实际场景配置的参数，请参见**快速配置参数**。

配置 Stage 失败重试次数

Spark任务在遇到FetchFailedException时会触发Stage重试。为了防止Stage无限重试，对Stage重试次数进行限制。重试次数可以根据实际需要进行调整。

在Spark客户端的“spark-defaults.conf”文件中配置如下参数。

表 12-347 参数说明

参数	说明	默认值
spark.stage.maxConsecutiveAttempts	Stage失败重试最大次数。	4

配置是否使用笛卡尔积功能

要启动使用笛卡尔积功能，需要在Spark的“spark-defaults.conf”配置文件中进行如下设置。

表 12-348 笛卡尔积参数说明

参数	说明	默认值
spark.sql.crossJoin.enabled	是否允许隐性执行笛卡尔积。 ● “true”表示允许 ● “false”表示不允许，此时只允许query中显式包含CROSS JOIN语法。	true

说明

- JDBC应用在服务端的“spark-defaults.conf”配置文件中设置该参数。
- Spark客户端提交的任务在客户端配的“spark-defaults.conf”配置文件中设置该参数。

Spark 长时间任务安全认证配置

安全模式下，使用Spark CLI（如spark shell、spark sql、spark submit）时，如果使用kinit命令进行安全认证，当执行长时间运行任务时，会因为认证过期导致任务失败。

在客户端的“spark-defaults.conf”配置文件中设置如下参数，配置完成后，重新执行Spark CLI即可。

说明

当参数值为“true”时，需要保证“spark-defaults.conf”和“hive-site.xml”中的Keytab和principal的值相同。

表 12-349 参数说明

参数名称	含义	默认值
spark.kerberos.principal	具有Spark操作权限的principal。请联系管理员获取对应principal。	-
spark.kerberos.keytab	具有Spark操作权限的Keytab文件名称和文件路径。请联系管理员获取对应Keytab文件。	-
spark.security.bigdata.loginOnce	Principal用户是否只登录一次。true为单次登录；false为多次登录。 单次登录与多次登录的区别在于：Spark社区使用多次Kerberos用户登录多次的方案，但容易出现TGT过期或者Token过期异常导致应用无法长时间运行。DataSight修改了Kerberos登录方式，只允许用户登录一次，可以有效的解决过期问题。限制在于，Hive相关的principal与keytab的配置项必须与Spark配置相同。 **说明** 当参数值为true时，需要保证“spark-defaults.conf”和“hive-site.xml”中的Keytab和principal的值相同。	true

Python Spark

Python Spark是Spark除了Scala、Java两种API之外的第三种编程语言。不同于Java和Scala都是在JVM平台上运行，Python Spark不仅会有JVM进程，还会有自身的Python进程。以下配置项只适用于Python Spark场景，而其他配置项也同样可以在Python Spark中生效。

表 12-350 参数说明

参数	描述	默认值
spark.python.profile	在Python worker中开启profiling。通过sc.show_profiles()展示分析结果。或者在driver退出前展示分析结果。可以通过sc.dump_profiles(path) 将结果转储到磁盘中。如果一些分析结果已经手动展示，那么在Driver退出前，它们将不会再自动展示。 默认使用pyspark.profiler.BasicProfiler，可以在初始化SparkContext时传入指定的profiler来覆盖默认的profiler。	false

参数	描述	默认值
spark.python.worker.memory	聚合过程中每个python worker进程所能使用的内存大小，其值格式同指定JVM内存一致，如512m，2g。如果进程在聚集期间所用的内存超过了该值，数据将会被写入磁盘。	512m
spark.python.worker.reuse	是否重用python worker。如是，它将使用固定数量的Python workers，那么下一批提交的task将重用这些Python workers，而不是为每个task重新fork一个Python进程。 该功能在大型广播下非常有用，因为此时对下一批提交的task不需要将数据从JVM再一次传输至Python worker。	true

Dynamic Allocation

动态资源调度是On Yarn模式特有的特性，并且必须开启Yarn External Shuffle才能使用这个功能。在使用Spark作为一个常驻的服务时候，动态资源调度将大大的提高资源的利用率。例如JDBCServer服务，大多数时间该进程并不接受JDBC请求，因此将这段空闲时间的资源释放出来，将极大的节约集群的资源。

表 12-351 参数说明

参数	描述	默认值
spark.dynamicAllocation.enabled	是否使用动态资源调度，用于根据规模调整注册于该应用的executor的数量。注意目前仅在YARN模式下有效。 启用动态资源调度必须将spark.shuffle.service.enabled设置为true。以下配置也与此相关：spark.dynamicAllocation.minExecutors、spark.dynamicAllocation.maxExecutors和spark.dynamicAllocation.initialExecutors。	• JDBCServer2x：true • SparkResource2x：false
spark.dynamicAllocation.minExecutors	最小Executor个数。	0
spark.dynamicAllocation.initialExecutors	初始Executor个数。	spark.dynamicAllocation.minExecutors
spark.dynamicAllocation.maxExecutors	最大executor个数。	2048
spark.dynamicAllocation.schedulerBacklogTimeout	调度第一次超时时间。单位为秒。	1s

参数	描述	默认值
spark.dynamicAllocation.sustainedSchedulerBacklogTimeout	调度第二次及之后超时时间。	1s
spark.dynamicAllocation.executorIdleTimeout	普通Executor空闲超时时间。单位为秒。	60
spark.dynamicAllocation.cachedExecutorIdleTimeout	含有cached blocks的Executor空闲超时时间。	• JDBCServer2x：2147483647s • IndexServer2x：2147483647s • SparkResource2x：120

Spark Streaming

Spark Streaming是在Spark批处理平台提供的流式数据的处理能力，以“mini-batch”的方式处理从外部输入的数据。

在Spark客户端的“spark-defaults.conf”文件中配置如下参数。

表 12-352 参数说明

参数	描述	默认值
spark.streaming.receiver.writeAheadLog.enable	启用预写日志（WAL）功能。所有通过Receiver接收的输入数据将被保存至预写日志，预写日志可以保证Driver程序出错后数据可以恢复。	false
spark.streaming.unpersist	由Spark Streaming产生和保存的RDDs自动从Spark的内存中强制移除。Spark Streaming接收的原始输入数据也将自动清除。设置为false时原始输入数据和存留的RDDs不会自动清除，因此在streaming应用外部依然可以访问，但是这会占用更多的Spark内存。	true

Spark Streaming Kafka

Receiver是Spark Streaming一个重要的组成部分，它负责接收外部数据，并将数据封装为Block，提供给Streaming消费。最常见的数据源是Kafka，Spark Streaming对

Kafka的集成也是最完善的，不仅有可靠性的保障，而且也支持从Kafka直接作为RDD输入。

表 12-353 参数说明

参数	描述	默认值
spark.streaming.kafka.maxRatePerPartition	使用Kafka direct stream API时，从每个Kafka分区读取数据的最大速率（每秒记录数量）。	-
spark.streaming.blockInterval	在被存入Spark之前Spark Streaming Receiver接收数据累积成数据块的间隔（毫秒）。推荐最小值为50毫秒。	200ms
spark.streaming.receiver.maxRate	每个Receiver接收数据的最大速率（每秒记录数量）。配置设置为0或者负值将不会对速率设限。	-
spark.streaming.receiver.writeAheadLog.enable	是否使用ReliableKafkaReceiver。该Receiver支持流式数据不丢失。	false

Netty/NIO 及 Hash/Sort 配置

Shuffle是大数据处理中最重要的一个性能点，网络是整个Shuffle过程的性能点。目前Spark支持两种Shuffle方式，一种是Hash，另外一种Sort。网络也有两种方式，Netty和NIO。

表 12-354 参数说明

参数	描述	默认值
spark.shuffle.manager	处理数据的方式。有两种实现方式可用：sort和hash。sort shuffle对内存的使用率更高，是Spark 1.2及后续版本的默认选项。	SORT
spark.shuffle.consolidateFiles	（仅hash方式）若要合并在shuffle过程中创建的中间文件，需要将该值设置为“true”。文件创建的少可以提高文件系统处理性能，降低风险。使用ext4或者xfs文件系统时，建议设置为“true”。由于文件系统限制，在ext3上该设置可能会降低8核以上机器的处理性能。	false
spark.shuffle.sort.bypassMergeThreshold	该参数只适用于spark.shuffle.manager设置为sort时。在不做map端聚合并且reduce任务的partition数小于或等于该值时，避免对数据进行归并排序，防止系统处理不必要的排序引起性能下降。	200

参数	描述	默认值
spark.shuffle.io.maxRetries	（仅Netty方式）如果设为非零值，由于IO相关的异常导致的fetch失败会自动重试。该重试逻辑有助于大型shuffle在发生长GC暂停或者网络闪断时保持稳定。	12
spark.shuffle.io.numConnectionsPerPeer	（仅Netty方式）为了减少大型集群的连接创建，主机间的连接会被重新使用。对于拥有较多硬盘和少数主机的集群，此操作可能会导致并发性不足以占用所有磁盘，所以用户可以考虑增加此值。	1
spark.shuffle.io.preferDirectBufs	（仅Netty方式）使用off-heap缓冲区减少shuffle和高速缓存块转移期间的垃圾回收。对于off-heap内存被严格限制的环境，用户可以将其关闭以强制所有来自Netty的申请使用堆内内存。	true
spark.shuffle.io.retryWait	（仅Netty方式）等待fetch重试期间的时间（秒）。重试引起的最大延迟为maxRetries * retryWait，默认是15秒。	5

普通 Shuffle 配置

表 12-355 参数说明

参数	描述	默认值
spark.shuffle.spill	若设为“true”，通过将数据溢出至磁盘来限制reduce任务期间内存的使用量。	true
spark.shuffle.spill.compress	是否压缩shuffle期间溢出的数据。使用spark.io.compression.codec指定的算法进行数据压缩。	true
spark.shuffle.file.buffer	每个shuffle文件输出流的内存缓冲区大小（单位：KB）。这些缓冲区可以减少创建中间shuffle文件流过程中产生的磁盘寻道和系统调用次数。也可以通过配置项spark.shuffle.file.buffer.kb设置。	32KB
spark.shuffle.compress	是否压缩map任务输出文件。建议压缩。使用spark.io.compression.codec进行压缩。	true
spark.reducer.maxSizeInFlight	从每个reduce任务同时fetch的map任务输出最大值（单位：MB）。由于每个输出要求创建一个缓冲区进行接收，这代表了每个reduce任务固定的内存开销，所以除非拥有大量内存，否则保持低值。也可以通过配置项spark.reducer.maxMbInFlight设置。	48MB

Driver 配置

Spark Driver可以理解为Spark提交应用的客户端，所有的代码解析工作都在这个进程中完成，因此该进程的参数尤其重要。下面将以如下顺序介绍Spark中进程的参数设置：

- JavaOptions：Java命令中“-D”后面的参数，可以由System.getProperty获取。
- ClassPath：包括Java类和Native的Lib加载路径。
- Java Memory and Cores：Java进程的内存和CPU使用量。
- Spark Configuration：Spark内部参数，与Java进程无关。

表 12-356 参数说明

参数	描述	默认值
spark.driver.extraJavaOptions	传递至driver（驱动程序）的一系列额外JVM选项。例如，GC设置或其他日志记录。 注意：在Client模式中，该配置禁止直接在应用程序中通过SparkConf设置，因为驱动程序JVM已经启动。请通过--driver-java-options命令行选项或默认property文件进行设置。	参考**快速配置参数**
spark.driver.extraClassPath	附加至driver的classpath的额外classpath条目。 注意：在Client模式中，该配置禁止直接在应用程序中通过SparkConf设置，因为驱动程序JVM已经启动。请通过--driver-java-options命令行选项或默认property文件进行设置。	参考**快速配置参数**
spark.driver.userClassPathFirst	（试验性）当在驱动程序中加载类时，是否授权用户添加的jar优先于Spark自身的jar。这种特性可用于减缓Spark依赖和用户依赖之间的冲突。目前该特性仍处于试验阶段，仅用于Cluster模式中。	false
spark.driver.extraLibraryPath	设置一个特殊的library path在启动驱动程序JVM时使用。 注意：在Client模式中，该配置禁止直接在应用程序中通过SparkConf设置，因为驱动程序JVM已经启动。请通过--driver-java-options命令行选项或默认property文件进行设置。	• JDBCServer2x：${SPARK_INSTALL_HOME}/spark/native • SparkResource2x：${DATA_NODE_INSTALL_HOME}/hadoop/lib/native

参数	描述	默认值
spark.driver.cores	驱动程序进程使用的核数。仅适用于Cluster模式。	1
spark.driver.memory	驱动程序进程使用的内存数量，即SparkContext初始化的进程（例如：512M, 2G）。 注意：在Client模式中，该配置禁止直接在应用程序中通过SparkConf设置，因为驱动程序JVM已经启动。请通过--driver-java-options命令行选项或默认property文件进行设置。	4G
spark.driver.maxResultSize	对每个Spark action操作（例如“collect”）的所有分区序列化结果的总量限制，至少1M，设置成0表示不限制。如果总量超过该限制，工作任务会中止。限制值设置过高可能会引起驱动程序的内存不足错误（取决于spark.driver.memory和JVM的对象内存开销）。设置合理的限制可以避免驱动程序出现内存不足的错误。	1G
spark.driver.host	Driver监听的主机名或IP地址，用于Driver与Executor进行通信。	(local hostname)
spark.driver.port	Driver监听的端口，用于Driver与Executor进行通信。	(random)

ExecutorLaucher 配置

ExecutorLauncher只有在Yarn-Client模式下才会存在的角色，Yarn-Client模式下，ExecutorLauncher和Driver不在同一个进程中，需要对ExecutorLauncher的参数进行特殊的配置。

表 12-357 参数说明

参数	描述	默认值
spark.yarn.am.extraJavaOptions	在Client模式下传递至YARN Application Master的一系列额外JVM选项。在Cluster模式下使用spark.driver.extraJavaOptions。	参考**快速配置参数**
spark.yarn.am.memory	针对Client模式下YARN Application Master使用的内存数量，与JVM内存设置字符串格式一致（例如：512m，2g）。在集群模式下，使用spark.driver.memory。	1G
spark.yarn.am.memoryOverhead	和“spark.yarn.driver.memoryOverhead”一样，但只针对Client模式下的Application Master。	-

参数	描述	默认值
spark.yarn.am.cores	针对Client模式下YARN Application Master使用的核数。在Cluster模式下，使用spark.driver.cores。	1

Executor 配置

Executor也是单独一个Java进程，但不像Driver和AM只有一个，Executor可以有多个进程，而目前Spark只支持相同的配置，即所有Executor的进程参数都必然是一样的。

表 12-358 参数说明

参数	描述	默认值
spark.executor.extraJavaOptions	传递至Executor的额外JVM选项。例如，GC设置或其他日志记录。请注意不能通过此选项设置Spark属性或heap大小。Spark属性应该使用SparkConf对象或调用spark-submit脚本时指定的spark-defaults.conf文件来设置。Heap大小可以通过spark.executor.memory来设置。	参考**快速配置参数**
spark.executor.extraClassPath	附加至Executor classpath的额外的classpath。这主要是为了向后兼容Spark的历史版本。用户一般不用设置此选项。	-
spark.executor.extraLibraryPath	设置启动executor JVM时所使用的特殊的library path。	参考**快速配置参数**
spark.executor.userClassPathFirst	（试验性）与spark.driver.userClassPathFirst相同的功能，但应用于Executor实例。	false
spark.executor.memory	每个Executor进程使用的内存数量，与JVM内存设置字符串的格式相同（例如：512M，2G）。	4G
spark.executorEnv.[EnvironmentVariableName]	添加由EnvironmentVariableName指定的环境变量至executor进程。用户可以指定多个来设置多个环境变量。	-
spark.executor.logs.rolling.maxRetainedFiles	设置系统即将保留的最新滚动日志文件的数量。旧的日志文件将被删除。默认关闭。	-
spark.executor.logs.rolling.size.maxBytes	设置滚动Executor日志的文件的最大值。默认关闭。数值以字节为单位设置。若要自动清除旧日志，请查看spark.executor.logs.rolling.maxRetainedFiles。	-

参数	描述	默认值
spark.executor.logs.rolling.strategy	设置executor日志的滚动策略。默认滚动关闭。可以设置为“time”（基于时间的滚动）或“size”（基于大小的滚动）。当设置为“time”，使用spark.executor.logs.rolling.time.interval属性的值作为日志滚动的间隔。当设置为“size”，使用spark.executor.logs.rolling.size.maxBytes设置滚动的最大文件大小滚动。	-
spark.executor.logs.rolling.time.interval	设置executor日志滚动的时间间隔。默认关闭。合法值为“daily”、“hourly”、“minutely”或任意秒。若要自动清除旧日志，请查看spark.executor.logs.rolling.maxRetainedFiles。	daily

WebUI

WebUI展示了Spark应用运行的过程和状态。

表 12-359 参数说明

参数	描述	默认值
spark.ui.killEnabled	允许停止Web UI中的stage和相应的job。 **说明** 出于安全考虑，将此配置项的默认值设置成false，以避免用户发生误操作。如果需要开启此功能，则可以在spark-defaults.conf配置文件中将此配置项的值设为true。请谨慎操作。	true
spark.ui.port	应用程序dashboard的端口，显示内存和工作量数据。	• JDBCServer2x：4040 • SparkResource2x：0 • IndexServer2x：22901
spark.ui.retainedJobs	在垃圾回收之前Spark UI和状态API记住的job数。	1000
spark.ui.retainedStages	在垃圾回收之前Spark UI和状态API记住的stage数。	1000

HistoryServer

HistoryServer读取文件系统中的EventLog文件，展示已经运行完成的Spark应用在运行时的状态信息。

表 12-360 参数说明

参数	描述	默认值
spark.history.fs.logDirectory	History server的日志目录	-
spark.history.ui.port	JobHistory侦听连接的端口。	18080
spark.history.fs.updateInterval	History server所显示信息的更新周期，单位为秒。每次更新检查持久存储中针对事件日志进行的更改。	10s
spark.history.fs.update.interval.seconds	每个事件日志更新检查的间隔。与spark.history.fs.updateInterval功能相同，推荐使用spark.history.fs.updateInterval。	10s
spark.history.updateInterval	该配置项与spark.history.fs.update.interval.seconds和spark.history.fs.updateInterval功能相同，推荐使用spark.history.fs.updateInterval。	10s

HistoryServer UI 超时和最大访问数

表 12-361 参数说明

参数	描述	默认值
spark.session.maxAge	设置会话的超时时间，单位秒。此参数只适用于安全模式。普通模式下，无法设置此参数。	600
spark.connection.maxRequest	设置客户端访问Jobhistory的最大并发数量。	5000

EventLog

Spark应用在运行过程中，实时将运行状态以JSON格式写入文件系统，用于HistoryServer服务读取并重现应用运行时状态。

表 12-362 参数说明

参数	描述	默认值
spark.eventLog.enabled	是否记录Spark事件，用于应用程序在完成后重构webUI。	true

参数	描述	默认值
spark.eventLog.dir	如果**spark.eventLog.enabled**为**true**，记录Spark事件的目录。在此目录下，Spark为每个应用程序创建文件，并将应用程序的事件记录到文件中。用户也可设置为统一的与HDFS目录相似的地址，这样History server就可以读取历史文件。	hdfs://hacluster/spark2xJobHistory2x
spark.eventLog.compress	spark.eventLog.enabled为true时，是否压缩记录的事件。	false

EventLog 的周期清理

JobHistory上的Event log是随每次任务的提交而累积的，任务提交的次数多了之后会造成太多文件的存放。Spark提供了周期清理Evnet log的功能，用户可以通过配置开关和相应的清理周期参数来进行控制。

表 12-363 参数说明

参数	描述	默认值
spark.history.fs.cleaner.enabled	是否打开清理功能。	true
spark.history.fs.cleaner.interval	清理功能的检查周期。	1d
spark.history.fs.cleaner.maxAge	日志的最长保留时间。	4d

Kryo

Kryo是一个非常高效的Java序列化框架，Spark中也默认集成了该框架。几乎所有的Spark性能调优都离不开将Spark默认的序列化器转化为Kryo序列化器的过程。目前Kryo序列化只支持Spark数据层面的序列化，还不支持闭包的序列化。设置Kryo序列元，需要将配置项“spark.serializer”设置为“org.apache.spark.serializer.KryoSerializer”，同时也搭配设置以下的配置项，优化Kryo序列化的性能。

表 12-364 参数说明

参数	描述	默认值
spark.kryo.classesToRegister	使用Kryo序列化时，需要注册到Kryo的类名，多个类之间用逗号分隔。	-

参数	描述	默认值
spark.kryo.referenceTracking	当使用Kryo序列化数据时，是否跟踪对同一个对象的引用情况。适用于对象图有循环引用或同一对象有多个副本的情况。否则可以设置为关闭以提升性能。	true
spark.kryo.registrationRequired	是否需要使用Kryo来注册对象。当设为“true”时，如果序列化一个未使用Kryo注册的对象则会抛出异常。当设为“false”（默认值）时，Kryo会将未注册的类名称一同写到序列化对象中。该操作会带来大量性能开销，所以在用户还没有从注册队列中删除相应的类时应该开启该选项。	false
spark.kryo.registrator	如果使用Kryo序列化，使用Kryo将该类注册至定制类。如果需要以定制方式注册类，例如指定一个自定义字段序列化器，可使用该属性。否则spark.kryo.classesToRegister会更简单。它应该设置为一个扩展KryoRegistrator的类。	-
spark.kryoserializer.buffer.max	Kryo序列化缓冲区允许的最大值，单位为兆字节。这个值必须大于尝试序列化的对象。当在Kryo中遇到“buffer limit exceeded”异常时可以适当增大该值。也可以通过配置项spark.kryoserializer.buffer.max配置。	64MB
spark.kryoserializer.buffer	Kryo序列化缓冲区的初始值，单位为兆字节。每个worker的每个核心都会有一个缓冲区。如果有需要，缓冲区会增大到spark.kryoserializer.buffer.max设置的值。也可以通过配置项spark.kryoserializer.buffer配置。	64KB

Broadcast

Broadcast用于Spark进程间数据块的传输。Spark中无论Jar包、文件还是闭包以及返回的结果都会使用Broadcast。目前的Broadcast支持两种方式，Torrent与HTTP。前者将会把数据切成小片，分布到集群中，有需要时从远程获取；后者将文件存入到本地磁盘，有需要时通过HTTP方式将整个文件传输到远端。前者稳定性优于后者，因此Torrent为默认的Broadcast方式。

表 12-365 参数说明

参数	描述	默认值
spark.broadcast.factory	使用的广播方式。	org.apache.spark.broadcast.TorrentBroadcastFactory
spark.broadcast.blockSize	TorrentBroadcastFactory的块大小。该值过大会降低广播时的并行度（速度变慢），过小可能会影响BlockManager的性能。	4096

参数	描述	默认值
spark.broadcast.compress	在发送广播变量之前是否压缩。建议压缩。	true

Storage

内存计算是Spark的最大亮点，Spark的Storage主要管理内存资源。Storage中主要存储RDD在Cache过程中产生的数据块。JVM中堆内存是整体的，因此在Spark的Storage管理中，“Storage Memory Size”变成了一个非常重要的概念。

表 12-366 参数说明

参数	描述	默认值
spark.storage.memoryMapThreshold	超过该块大小的Block，Spark会对该磁盘文件进行内存映射。这可以防止Spark在内存映射时映射过小的块。一般情况下，对接近或低于操作系统的页大小的块进行内存映射会有高开销。	2m

PORT

表 12-367 参数说明

参数	描述	默认值
spark.ui.port	应用仪表盘的端口，显示内存和工作负载数据。	• JDBCServer2x：4040 • SparkResource2x：0
spark.blockManager.port	所有BlockManager监听的端口。这些同时存在于Driver和Executor上。	**随机端口范围**
spark.driver.port	Driver监听的端口，用于Driver与Executor进行通信。	**随机端口范围**

随机端口范围

所有随机端口必须在一定端口范围内。

表 12-368 参数说明

参数	描述	默认值
spark.random.port.min	设置随机端口的最小值。	22600
spark.random.port.max	设置随机端口的最大值。	22899

TIMEOUT

Spark默认配置能很好的处理中等数据规模的计算任务，但一旦数据量过大，会经常出现超时导致任务失败的场景。在大数据量场景下，需调大Spark中的超时参数。

表 12-369 参数说明

参数	描述	默认值
spark.files.fetchTimeout	获取通过驱动程序的SparkContext.addFile()添加的文件时的通信超时（秒）。	60s
spark.network.timeout	所有网络交互的默认超时（秒）。如未配置，则使用该配置代替 spark.core.connection.ack.wait.timeout, spark.akka.timeout, spark.storage.blockManagerSlaveTimeoutMs或 spark.shuffle.io.connectionTimeout。	360s
spark.core.connection.ack.wait.timeout	连接时应答的超时时间（单位：秒）。为了避免由于GC带来的长时间等待，可以设置更大的值。	60

加密

Spark支持Akka和HTTP（广播和文件服务器）协议的SSL，但WebUI和块转移服务仍不支持SSL。

SSL必须在每个节点上配置，并使用特殊协议为通信涉及到的每个组件进行配置。

表 12-370 参数说明

参数	描述	默认值
spark.ssl.enabled	是否在所有被支持协议上开启SSL连接。 与spark.ssl.xxx类似的所有SSL设置指示了所有被支持协议的全局配置。为了覆盖特殊协议的全局配置，在协议指定的命名空间中必须重写属性。 使用“spark.ssl.YYY.XXX”设置覆盖由YYY指示的特殊协议的全局配置。目前YYY可以是基于Akka连接的akka或广播与文件服务器的fs。	false

参数	描述	默认值
spark.ssl.enabledAlgorithms	以逗号分隔的密码列表。指定的密码必须被JVM支持。	-
spark.ssl.keyPassword	key-store的私人密钥密码。	-
spark.ssl.keyStore	key-store文件的路径。该路径可以绝对或相对于开启组件的目录。	-
spark.ssl.keyStorePassword	key-store的密码。	-
spark.ssl.protocol	协议名。该协议必须被JVM支持。本页所有协议的参考表。	-
spark.ssl.trustStore	trust-store文件的路径。该路径可以绝对或相对于开启组件的目录。	-
spark.ssl.trustStorePassword	trust-store的密码。	-

安全性

Spark目前支持通过共享密钥认证。可以通过spark.authenticate配置参数配置认证。该参数控制Spark通信协议是否使用共享密钥执行认证。该认证是确保双边都有相同的共享密钥并被允许通信的基本握手。如果共享密钥不同，通信将不被允许。共享密钥通过如下方式创建：

- 对于YARN部署的Spark，将spark.authenticate配置为真会自动处理生成和分发共享密钥。每个应用程序会独占一个共享密钥。
- 对于其他类型部署的Spark，应该在每个节点上配置Spark参数spark.authenticate.secret。所有Master/Workers和应用程序都将使用该密钥。

表 12-371 参数说明

参数	描述	默认值
spark.acls.enable	是否开启Spark acls。如果开启，它将检查用户是否有访问和修改job的权限。请注意这要求用户可以被识别。如果用户被识别为无效，检查将不被执行。UI可以使用过滤器认证和设置用户。	true
spark.admin.acls	逗号分隔的有权限访问和修改所有Spark job的用户/管理员列表。如果在共享集群上运行并且工作时有管理员或开发人员帮助调试，可以使用该列表。	admin
spark.authenticate	是否Spark认证其内部连接。如果不是运行在YARN上，请参见spark.authenticate.secret。	true

参数	描述	默认值
spark.authenticate.secret	设置Spark各组件之间验证的密钥。如果不是运行在YARN上且认证未开启，需要设置该项。	-
spark.modify.acls	逗号分隔的有权限修改Spark job的用户列表。默认情况下只有开启Spark job的用户才有修改列表的权限（例如删除列表）。	-
spark.ui.view.acls	逗号分隔的有权限访问Spark web ui的用户列表。默认情况下只有开启Spark job的用户才有访问权限。	-

开启 Spark 进程间的认证机制

目前Spark进程间支持共享密钥方式的认证机制，通过配置spark.authenticate可以控制Spark在通信过程中是否做认证。这种认证方式只是通过简单的握手来确定通信双方享有共同的密钥。

在Spark客户端的“spark-defaults.conf”文件中配置如下参数。

表 12-372 参数说明

参数	描述	默认值
spark.authenticate	在Spark on YARN模式下，将该参数配置成true即可。密钥的生成和分发过程是自动完成的，并且每个应用独占一个密钥。	true

Compression

数据压缩是一个以CPU换内存的优化策略，因此当Spark内存严重不足的时候（由于内存计算的特质，这种情况非常常见），使用压缩可以大幅提高性能。目前Spark支持三种压缩算法：snappy，lz4，lzf。Snappy为默认压缩算法，并且调用native方法进行压缩与解压缩，在Yarn模式下需要注意堆外内存对Container进程的影响。

表 12-373 参数说明

参数	描述	默认值
spark.io.compression.codec	用于压缩内部数据的codec，例如RDD分区、广播变量和shuffle输出。默认情况下，Spark支持三种压缩算法：lz4，lzf和snappy。可以使用完全合格的类名称指定算法，例如org.apache.spark.io.LZ4CompressionCodec、org.apache.spark.io.LZFCompressionCodec及org.apache.spark.io.SnappyCompressionCodec。	lz4
spark.io.compression.lz4.block.size	当使用LZ4压缩算法时LZ4压缩中使用的块大小（字节）。当使用LZ4时降低块大小同样也会降低shuffle内存使用。	32768

参数	描述	默认值
spark.io.compression.snappy.block.size	当使用Snappy压缩算法时Snappy压缩中使用的块大小（字节）。当使用Snappy时降低块大小同样也会降低shuffle内存使用。	32768
spark.shuffle.compress	是否压缩map任务输出文件。建议压缩。使用spark.io.compression.codec进行压缩。	true
spark.shuffle.spill.compress	是否压缩在shuffle期间溢出的数据。使用spark.io.compression.codec进行压缩。	true
spark.eventLog.compress	设置当spark.eventLog.enabled设置为true时是否压缩记录的事件。	false
spark.broadcast.compress	在发送之前是否压缩广播变量。建议压缩。	true
spark.rdd.compress	是否压缩序列化的RDD分区（例如StorageLevel.MEMORY_ONLY_SER的分区）。牺牲部分额外CPU的时间可以节省大量空间。	false

在资源不足的情况下，降低客户端运行异常概率

在资源不足的情况下，Application Master会因等待资源出现超时，导致任务被删除。调整如下参数，降低客户端应用运行异常概率。

在客户端的“spark-defaults.conf”配置文件中调整如下参数。

表 12-374 参数说明

参数	说明	默认值
spark.yarn.applicationMaster.waitTries	设置Application Master等待Spark master的次数，同时也是等待SparkContext初始化的次数。增大该参数值，可以防止AM任务被删除，降低客户端应用运行异常的概率。	10
spark.yarn.am.memory	调整AM的内存。增大该参数值，可以防止AM因内存不足而被RM删除任务，降低客户端应用运行异常的概率。	1G

12.23.2.4 SparkOnHBase 概述及基本应用

操作场景

Spark on HBase为用户提供了在Spark SQL中查询HBase表，通过Beeline工具为HBase表进行存数据等操作。通过HBase接口可实现创建表、读取表、往表中插入数据等操作。

操作步骤

步骤1 登录Manager界面，选择“集群 > *待操作集群的名称* > 集群属性”查看集群是否为安全模式。

- 是，执行**步骤2**。
- 否，执行**步骤5**。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置 > 全部配置 > JDBCServer2x > 默认”，修改以下参数：

表 12-375 参数列表 1

参数	默认值	修改结果
spark.yarn.security.credentials.hbase.enabled	false	true

说明

为了保证Spark2x可以长期访问HBase，建议不要修改HBase与HDFS服务的以下参数：

- dfs.namenode.delegation.token.renew-interval
- dfs.namenode.delegation.token.max-lifetime
- hbase.auth.key.update.interval
- hbase.auth.token.max.lifetime（不可修改，固定值为604800000毫秒，即7天）

如果必须要修改以上参数，请务必保证HDFS参数“dfs.namenode.delegation.token.renew-interval”的值不大于HBase参数“hbase.auth.key.update.interval”、“hbase.auth.token.max.lifetime”的值和HDFS参数“dfs.namenode.delegation.token.max-lifetime”的值。

步骤3 选择“SparkResource2x > 默认”，修改以下参数：

表 12-376 参数列表 2

参数	默认值	修改结果
spark.yarn.security.credentials.hbase.enabled	false	true

步骤4 重启Spark2x服务，配置生效。

说明

如果需要在Spark2x客户端用Spark on HBase功能，需要重新下载并安装Spark2x客户端。

步骤5 在Spark2x客户端使用spark-sql或者spark-beeline连接，可以查询由Hive on HBase所创建的表，支持通过SQL命令创建HBase表或创建外表关联HBase表。建表前，确认HBase中已存在对应 HBase表，下面以HBase表table1为例说明。

1. 通过Beeline工具创建HBase表，命令如下：

   create table *hbaseTable*

   (

id string,
name string,
age int
)
using org.apache.spark.sql.hbase.HBaseSource
options(
hbaseTableName *"table1"***,**
keyCols *"id"***,**
colsMapping "
name=cf1.cq1,
age=cf1.cq2
");

说明

- hbaseTable：是创建的spark表的表名。
- id string,name string, age int：是spark表的字段名和字段类型。
- table1：HBase表名。
- id：HBase表的rowkey列名。
- name=cf1.cq1, age=cf1.cq2：spark表的列和HBase表的列的映射关系。spark的name列映射HBase表的cf1列簇的cq1列，spark的age列映射HBase表的cf1列簇的cq2列。

2. 通过csv文件导入数据到HBase表，命令如下：

   hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator="," -Dimporttsv.columns=HBASE_ROW_KEY,*cf1:cq1,cf1:cq2,cf1:cq3,cf1:cq4,cf1:cq5 table1 /hperson*

   其中：table1为HBase表名，/hperson为csv文件存放的路径。

3. 在spark-sql或spark-beeline中查询数据，*hbaseTable*为对应的spark表名。命令如下：

   select * from *hbaseTable;*

----结束

12.23.2.5 SparkOnHBasev2 概述及基本应用

操作场景

Spark on HBaseV2为用户提供了在Spark SQL中查询HBase表，通过Beeline工具为HBase表进行存数据等操作。通过HBase接口可实现创建表、读取表、往表中插入数据等操作。

操作步骤

步骤1 登录Manager界面，选择“集群 > *待操作集群的名称* > 集群属性”查看集群是否为安全模式。

- 是，执行**步骤2**。
- 否，执行**步骤5**。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置 > 全部配置 > JDBCServer2x > 默认”，修改以下参数：

表 12-377 参数列表 1

参数	默认值	修改结果
spark.yarn.security.credentials.hbase.enabled	false	true

说明

为了保证Spark2x可以长期访问HBase，建议不要修改HBase与HDFS服务的以下参数：

- dfs.namenode.delegation.token.renew-interval
- dfs.namenode.delegation.token.max-lifetime
- hbase.auth.key.update.interval
- hbase.auth.token.max.lifetime（不可修改，固定值为604800000毫秒，即7天）

如果必须要修改以上参数，请务必保证HDFS参数“dfs.namenode.delegation.token.renew-interval”的值不大于HBase参数“hbase.auth.key.update.interval”、“hbase.auth.token.max.lifetime”的值和HDFS参数“dfs.namenode.delegation.token.max-lifetime”的值。

步骤3 选择“SparkResource2x > 默认”，修改以下参数：

表 12-378 参数列表 2

参数	默认值	修改结果
spark.yarn.security.credentials.hbase.enabled	false	true

步骤4 重启Spark2x服务，配置生效。

说明

如果需要在Spark2x客户端用Spark on HBase功能，需要重新下载并安装Spark2x客户端。

步骤5 在Spark2x客户端使用spark-sql或者spark-beeline连接，可以查询由Hive on HBase所创建的表，支持通过SQL命令创建HBase表或创建外表关联HBase表。具体见下面说明。下面以HBase表table1为例说明。

1. 通过spark-beeline工具创建表的语法命令如下：

   create table *hbaseTable*1
   (*id string, name string, age int***)**
   using org.apache.spark.sql.hbase.HBaseSourceV2
   options(
   hbaseTableName *"table2"***,**
   keyCols *"id"***,**
   colsMapping "*name=cf1.cq1,age=cf1.cq2***");**

说明

- hbaseTable1：是创建的spark表的表名。
- id string,name string, age int：是spark表的字段名和字段类型。
- table2：HBase表名。
- id：HBase表的rowkey列名。
- name=cf1.cq1, age=cf1.cq2：spark表的列和HBase表的列的映射关系。spark的name列映射HBase表的cf1列簇的cq1列，spark的age列映射HBase表的cf1列簇的cq2列。

2. 通过csv文件导入数据到HBase表，命令如下：

   hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator="," -Dimporttsv.columns=HBASE_ROW_KEY,*cf1:cq1,cf1:cq2,cf1:cq3,cf1:cq4,cf1:cq5 table2 /hperson*

   其中：table2为HBase表名，/hperson为csv文件存放的路径。

3. 在spark-sql或spark-beeline中查询数据，*hbaseTable*1为对应的spark表名，命令如下：

   select * from *hbaseTable1;*

----结束

12.23.2.6 SparkSQL 权限管理（安全模式）

12.23.2.6.1 SparkSQL 权限介绍

SparkSQL 权限

类似于Hive，SparkSQL也是建立在Hadoop上的数据仓库框架，提供类似SQL的结构化数据。

MRS提供用户、用户组和角色，集群中的各类权限需要先授予角色，然后将用户或者用户组与角色绑定。用户只有绑定角色或者加入绑定角色的用户组，才能获得权限。

说明

- 如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Spark2x的Ranger访问权限策略**。
- Spark2x开启或关闭Ranger鉴权后，需要重启Spark2x服务，并重新下载客户端，或刷新客户端配置文件spark/conf/spark-defaults.conf：

  开启Ranger鉴权：spark.ranger.plugin.authorization.enable=true

  关闭Ranger鉴权：spark.ranger.plugin.authorization.enable=false

权限管理介绍

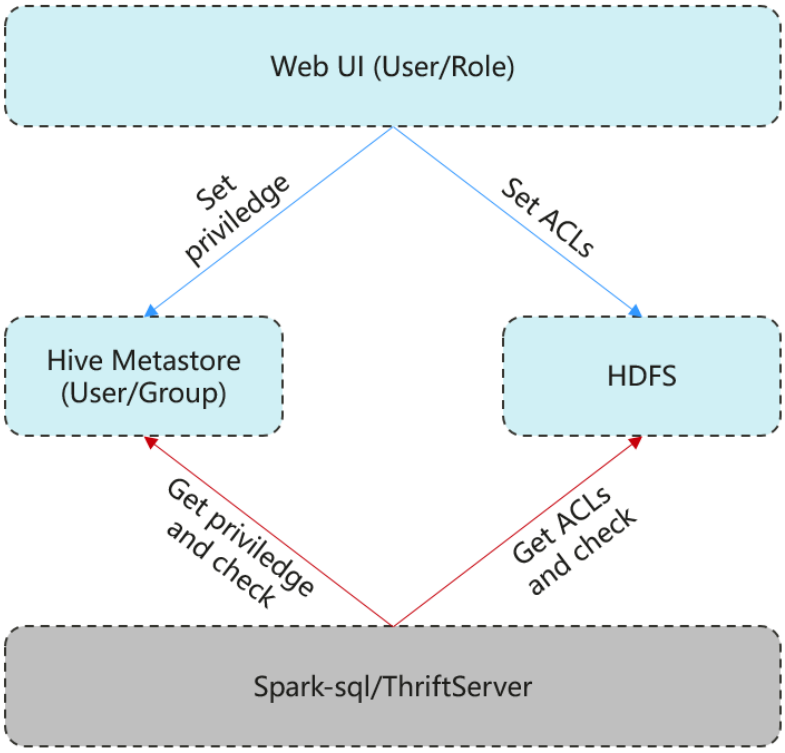
SparkSQL的权限管理是指SparkSQL中管理用户操作数据库的权限系统，以保证不同用户之间操作数据库的独立性和安全性。如果一个用户想操作另一个用户的表、数据库等，需要获取相应的权限才能进行操作，否则会被拒绝。

SparkSQL权限管理部分集成了Hive权限管理的功能。使用SparkSQL权限管理功能需要使用Hive的MetaStore服务和页面上的赋权功能。

图12-52展示了SparkSQL权限管理的基本架构。主要包含了两部分：页面赋权和服务获权并判断。

- 页面赋权：SparkSQL仅支持页面赋权的方式。在FusionInsight Manager的“系统 > 权限”中，可以进行用户、用户组和角色的添加/删除操作，可以对某个角色进行赋权/撤权。
- 服务获权并判断：当接收到客户端的DDL、DML的SQL命令时，SparkSQL服务会向MetaStore服务获取客户端用户对数据库信息的已有权限，并检查是否包含了所需的所有权限，如果是则继续执行，否则拒绝该用户的操作。当通过了MetaStore的权限检查后，还需进行HDFS的ACLs权限检查。

图 12-52 SparkSQL 权限管理架构图



SparkSQL还提供了列权限和视图权限，以满足用户不同场景的需求。

- 列权限介绍

  SparkSQL权限控制由元数据权限控制和HDFS ACL权限控制两部分组成。Hive MetaStore会将表权限自动同步到HDFS ACL中时，不会同步列级别的权限。也就是说，当用户对表具有部分列权限或全部列权限时，不能通过HDFS Client访问HDFS文件。

  - 在spark-sql模式下，用户仅具有列级别权限（即列权限用户）将不能访问HDFS文件，因此无法访问相应表的列。
  - Beeline/JDBCServer模式下，用户间赋权，例如将A用户创建的表赋权给B用户时。
    - “hive.server2.enable.doAs”=true（在Spark服务端的“hive-site.xml”文件中配置）

      此时用户B不可查询。需在HDFS上手动为文件赋读权限。
    - “hive.server2.enable.doAs”=false
      - 用户A和B均通过Beeline连接，用户B可查询。

- A用户通过SQL方式建表，B用户可在Beeline进行查询。

而其他情况，如A用户使用Beeline建表，B用户通过SQL查询，或者A用户通过SQL方式建表，B用户使用SQL方式查询的情况均不支持。需在HDFS上手动为文件赋读权限。

说明

由于“spark”用户在HDFS ACL的权限控制上为管理员用户权限，Beeline客户端用户的权限控制仅取决于Spark侧的元数据权限。

- 视图权限介绍

视图权限是指仅对表的视图具有查询、修改等操作的权限，不再依赖于视图所在的表的相应权限。即用户拥有视图的查询权限时，不管是否有表权限都可以进行查询。视图的权限是针对整个表而言的，不支持对其中的部分列创建视图权限。

视图权限在SparkSQL权限上的限制与列权限相似，详细如下：

- 在spark-sql模式下，只有视图权限而没有表权限，且没有HDFS的读取权限时，用户不能访问HDFS上存储的表的数据，即该情况下不支持对该表的视图进行查询。
- Beeline/JDBCServer模式下，用户间赋权，例如将A用户创建的视图赋权给B用户时。
  - “hive.server2.enable.doAs”=true（在Spark服务端的“hive-site.xml”文件中配置）

    此时用户B不可查询。需在HDFS上手动为文件赋读权限。
  - “hive.server2.enable.doAs”=false
    - 用户A和B均通过Beeline连接，用户B可查询。
    - A用户通过SQL方式创建视图，B用户可在Beeline进行查询。

    而其他情况，如A用户使用Beeline创建视图，B用户通过SQL查询，或者A用户通过SQL方式创建视图，B用户使用SQL方式查询的情况均不支持。需在HDFS上手动为文件赋读权限。

对表的视图进行相应操作，分别需要具有以下权限。

- 创建视图时，需要数据库的CREATE权限、表的SELECT、SELECT_of_GRANT权限。
- 查询、描述视图时，只需要视图的SELECT权限，不需要视图所依赖的表或依赖的视图的SELECT权限。若同时查询视图和其他表，则仍然需要其他表的SELECT权限，例如：select * from v1 join t1时，需要有视图v1和表t1的SELECT权限，即使v1是基于t1的视图，也需要表t1的SELECT权限。

  说明

  在Beeline/JDBCServer模式下，查询视图只需表的SELECT权限；而在spark-sql模式下，查询视图需要视图的SELECT权限和表的SELECT权限。
- 删除、修改视图时，必须要有视图的owner权限。

SparkSQL 权限模型

用户使用SparkSQL服务进行SQL操作，必须对SparkSQL数据库和表（含外表和视图）拥有相应的权限。完整的SparkSQL权限模型由元数据权限与HDFS文件权限组成。使用数据库或表时所需要的各种权限都是SparkSQL权限模型中的一种。

- 元数据权限

  元数据权限即在元数据层上进行权限控制，与传统关系型数据库类似，SparkSQL数据库包含“创建”和“查询”权限，表和列包含“查询”、“插入”、“UPDATE”和“删除”权限。SparkSQL中还包含拥有者权限“OWNERSHIP”和管理员权限“管理”。

- 数据文件权限，即HDFS文件权限

  SparkSQL的数据库、表对应的文件保存在HDFS中。默认创建的数据库或表保存在HDFS目录“/user/hive/warehouse”。系统自动以数据库名称和数据库中表的名称创建子目录。访问数据库或者表，需要在HDFS中拥有对应文件的权限，包含“读”、“写”和“执行”权限。

用户对SparkSQL数据库或表执行不同操作时，需要关联不同的元数据权限与HDFS文件权限。例如，对SparkSQL数据表执行查询操作，需要关联元数据权限“查询”，以及HDFS文件权限“读”和“执行”。

使用Manager界面图形化的角色管理功能来管理SparkSQL数据库和表的权限，只需要设置元数据权限，系统会自动关联HDFS文件权限，减少界面操作，提高效率。

SparkSQL 使用场景及对应权限

用户通过SparkSQL服务创建数据库需要加入Hive组，不需要角色授权。用户在Hive和HDFS中对自己创建的数据库或表拥有完整权限，可直接创建表、查询数据、删除数据、插入数据、更新数据以及授权他人访问表与对应HDFS目录与文件。

如果用户访问别人创建的表或数据库，需要授予权限。所以根据SparkSQL使用场景的不同，用户需要的权限可能也不相同。

表 12-379 SparkSQL 使用场景

主要场景	用户需要的权限
使用SparkSQL表、列或数据库	使用其他用户创建的表、列或数据库，不同的场景需要不同的权限，例如： ● 创建表，需要“创建”。 ● 查询数据，需要“查询”。 ● 插入数据，需要“插入”。
关联使用其他组件	部分场景除了SparkSQL权限，还可能需要组件的权限，例如： 使用Spark on HBase，在SparkSQL中查询HBase表数据，需要设置HBase权限。

在一些特殊SparkSQL使用场景下，需要单独设置其他权限。

表 12-380 SparkSQL 授权注意事项

场景	用户需要的权限
创建SparkSQL数据库、表、外表，或者为已经创建的表或外表添加分区，且Hive用户指定数据文件保存在“/user/hive/warehouse”以外的HDFS目录。	● 需要此目录已经存在，客户端用户是目录的属主，且用户对目录拥有“读”、“写”和“执行”权限。同时用户对此目录上层的每一级目录都拥有“读”和“执行”权限。 ● 在Spark2x中，在创建HBase的外表时，需要拥有Hive端database的“创建”权限。而在Spark 1.5中，在创建HBase的外表时，需要拥有Hive端database的“创建”权限，也需要拥有HBase端Namespace的“创建”权限。
用户使用load将指定目录下所有文件或者指定文件，导入数据到表中。	● 数据源为Linux本地磁盘，指定目录时需要此目录已经存在，系统用户“omm”对此目录以及此目录上层的每一级目录拥有“r”和“x”的权限。指定文件时需要此文件已经存在，“omm”对此文件拥有“r”的权限，同时对此文件上层的每一级目录拥有“r”和“x”的权限。 ● 数据源为HDFS，指定目录时需要此目录已经存在，SparkSQL用户是目录属主，且用户对此目录及其子目录拥有“读”、“写”和“执行”权限，并且其上层的每一级目录拥有“读”和“执行”权限。指定文件时需要此文件已经存在，SparkSQL用户是文件属主，且用户对文件拥有“读”、“写”和“执行”权限，同时对此文件上层的每一级目录拥有“读”和“执行”权限。
创建函数、删除函数或者修改任意数据库。	需要授予“管理”权限。
操作Hive中所有的数据库和表。	需加入到supergroup用户组，并且授予“管理”权限。
对部分datasource表赋予insert权限后，执行insert|analyze操作前需要单独对hdfs上的表目录赋予写权限。	当前对spark datasource表赋予Insert权限时，若表格式为：text|csv|json|parquet|orc,则不会修改表目录的权限。因此，对以上几种类型的datasource表赋予Insert权限后，还需要单独对hdfs上的表目录赋予写权限，用户才能成功对表执行insert|analyze操作。

12.23.2.6.2 创建 SparkSQL 角色

操作场景

该任务指导系统管理员在Manager创建并设置SparkSQL的角色。SparkSQL角色可设置管理员权限以及数据表的数据操作权限。

用户使用Hive并创建数据库需要加入hive组，不需要角色授权。用户在Hive和HDFS中对自己创建的数据库或表拥有完整权限，可直接创建表、查询数据、删除数据、插入数据、更新数据以及授权他人访问表与对应HDFS目录与文件。默认创建的数据库或表保存在HDFS目录“/user/hive/warehouse”。

说明

- 如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理，具体操作可参考**添加Spark2x的Ranger访问权限策略**。
- Spark2x开启或关闭Ranger鉴权后，需要重启Spark2x服务，并重新下载客户端，或刷新客户端配置文件spark/conf/spark-defaults.conf：

  开启Ranger鉴权：spark.ranger.plugin.authorization.enable=true

  关闭Ranger鉴权：spark.ranger.plugin.authorization.enable=false

操作步骤

1. 登录Manager页面，选择“系统 > 权限 > 角色”。
2. 单击“添加角色”，然后“角色名称”和“描述”输入角色名字与描述。
3. 设置角色“配置资源权限”请参见**表12-381**。
   - “Hive管理员权限”：Hive管理员权限。
   - “Hive读写权限”：Hive数据表管理权限，可设置与管理已创建的表的数据操作权限。

     说明

     - Hive角色管理支持授予管理员权限、访问表和视图的权限，不支持数据库的授权。
     - Hive管理员权限不支持管理HDFS的权限。
     - 如果数据库中的表或者表中的文件数量比较多，在授权时可能需要等待一段时间。例如表的文件数量为1万时，可能需要等待2分钟。

表 12-381 设置角色

任务场景	角色授权操作
设置Hive管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive”，勾选“Hive管理权限”。 用户绑定Hive管理员角色后，在每个维护操作会话中，还需要执行以下操作： 1. 以客户端安装用户，登录安装Spark2x客户端的节点。 2. 执行以下命令配置环境变量。 例如，Spark2x客户端安装目录为“/opt/client”，执行**source /opt/client/bigdata_env** **source /opt/client/Spark2x/component_env** 3. 执行以下命令认证用户。 ***kinit*** *Hive业务用户* 4. 执行以下命令登录客户端工具。 **/opt/client/Spark2x/spark/bin/beeline -u "jdbc:hive2://<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;user.principal=spark2x/hadoop.*<系统域名>**@***<系统域名>**;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>**@***<系统域名>**;"*** **说明** ● 其中“<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>”是Zookeeper的URL。例如“192.168.81.37:2181,192.168.195.232:2181,192.168.169.84:2181”。 ● 其中“sparkthriftserver”是Zookeeper上的目录，表示客户端从该目录下随机选择Triftserver实例或proxyThriftServer进行连接。 ● 用户可登录Manager，选择“系统 > 权限 > 域和互信”，查看“本端域”参数，即为当前系统域名。“spark2x/hadoop.<系统域名>”为用户名，用户的用户名所包含的系统域名所有字母为小写。例如“本端域”参数为“9427068F-6EFA-4833-B43E-60CB641E5B6C.COM”，用户名为“spark2x/hadoo.9427068f-6efa-4833-b43e-60cb641e5b6c.com”。 5. 执行以下命令更新用户的管理员权限。 **set role admin;**

任务场景	角色授权操作
设置在默认数据库中，查询其他用户表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”。 2. 在数据库列表中单击指定的数据库名称，显示数据库中的表。 3. 在指定表的“权限”列，勾选“查询”。
设置在默认数据库中，导入数据到其他用户表的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”。 2. 在数据库列表中单击指定的数据库名称，显示数据库中的表。 3. 在指定表的“权限”列，勾选“删除”和“插入”。

4. 单击“确定”完成。

12.23.2.6.3 配置表、列和数据库的权限

操作场景

使用SparkSQL操作表或者数据库时，如果用户访问别人创建的表或数据库，需要授予对应的权限。为了实现更严格权限控制，SparkSQL也支持列级别的权限控制。如果要访问别人创建的表上某些列，需要授予列权限。以下介绍使用Manager角色管理功能在表授权、列授权和数据库授权三个场景下的操作。

操作步骤

SparkSQL表授权、列授权、数据库授权与Hive的操作相同，详情请参见**权限管理**。

说明

- 在权限管理中，为了方便用户使用，授予数据库下表的任意权限将自动关联该数据库目录的HDFS权限。为了避免产生性能问题，取消表的任意权限，系统不会自动取消数据库目录的HDFS权限，但对应的用户只能登录数据库和查看表名。
- 若为角色添加或删除数据库的查询权限，数据库中的表也将自动添加或删除查询权限。此机制为Hive实现，SparkSQL与Hive保持一致。
- Spark不支持struct数据类型中列名称含有特殊字符（除字母、数字、下划线外的其他字符）。如果struct类型中列名称含有特殊字符，在FusionInsight Manager的“编辑角色”页面进行授权时，该列将无法正确显示。

相关概念

SparkSQL的语句在SparkSQL中进行处理，权限要求如**表12-382**所示。

表 12-382 使用 SparkSQL 表、列或数据库场景权限一览

操作场景	用户需要的权限
CREATE TABLE	“创建”, RWX+ownership（for create external table - the location） **说明** 按照指定文件路径创建datasource表时，需要path后面文件的RWX+ownership权限。
DROP TABLE	“Ownership”（of table）
DROP TABLE PROPERTIES	“Ownership”
DESCRIBE TABLE	“查询”
SHOW PARTITIONS	“查询”
ALTER TABLE LOCATION	“Ownership”, RWX+ownership (for new location)
ALTER PARTITION LOCATION	“Ownership”, RWX+ownership (for new partition location)
ALTER TABLE ADD PARTITION	“插入”, RWX+ownership (for partition location)
ALTER TABLE DROP PARTITION	“删除”
ALTER TABLE(all of them except the ones above)	“Update”, “Ownership”
TRUNCATE TABLE	“Ownership”
CREATE VIEW	“查询”, “Grant Of Select”, “创建”
ALTER VIEW PROPERTIES	“Ownership”
ALTER VIEW RENAME	“Ownership”
ALTER VIEW ADD PARTS	“Ownership”
ALTER VIEW AS	“Ownership”
ALTER VIEW DROPPARTS	“Ownership”
ANALYZE TABLE	“查询”, “插入”
SHOW COLUMNS	“查询”
SHOW TABLE PROPERTIES	“查询”
CREATE TABLE AS SELECT	“查询”, “创建”
SELECT	“查询” **说明** 与表一样，对视图进行SELECT操作的时候需要有该视图的“查询”权限。
INSERT	“插入”, “删除 (for overwrite)”

操作场景	用户需要的权限
LOAD	“插入”，“删除”，RWX+ownership(input location)
SHOW CREATE TABLE	“查询”，“Grant Of Select”
CREATE FUNCTION	“管理”
DROP FUNCTION	“管理”
DESC FUNCTION	-
SHOW FUNCTIONS	-
MSCK (metastore check)	“Ownership”
ALTER DATABASE	“管理”
CREATE DATABASE	-
SHOW DATABASES	-
EXPLAIN	“查询”
DROP DATABASE	“Ownership”
DESC DATABASE	-
CACHE TABLE	“查询”
UNCACHE TABLE	“查询”
CLEAR CACHE TABLE	“管理”
REFRESH TABLE	“查询”
ADD FILE	“管理”
ADD JAR	“管理”
HEALTHCHECK	-

12.23.2.6.4 配置 SparkSQL 业务使用其他组件的权限

操作场景

SparkSQL业务还可能需要关联使用其他组件，例如spark on HBase需要HBase权限。以下介绍SparkSQL关联HBase服务的操作。

前提条件

- 完成Spark客户端的安装，例如安装目录为“/opt/client”。
- 获取一个拥有管理员权限的用户，例如“admin”。

操作步骤

- **Spark on HBase授权**

  用户如果需要使用类似SQL语句的方式来操作HBase表，授予权限后可以使用SparkSQL访问HBase表。以授予用户在SparkSQL中查询HBase表的权限为例，操作步骤如下：

  说明

  设置“spark.yarn.security.credentials.hbase.enabled”为“true”。

  a. 在Manager角色界面创建一个角色，例如“hive_hbase_create”，并授予创建HBase表的权限。

  在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global”，勾选命名空间“default”的“创建”，单击“确定”保存。

  说明

  本例中建表是保存在Hive的“default”数据库中，默认具有“default”数据库的“建表”权限。如果Hive的数据库不是“default”，则还需要执行以下步骤：

  在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”，勾选所需指定的数据库的“建表”，单击“确定”保存。

  b. 在Manager角色界面创建一个角色，例如“hive_hbase_submit”，并授予提交任务到Yarn的队列的权限。

  在“配置资源权限”的表格中选择“*待操作集群的名称* > Yarn > 调度队列 > root”，勾选队列“default”的“提交”，单击“确定”保存。

  c. 在Manager用户界面创建一个“人机”用户，例如“hbase_creates_user”，加入“hive”组，绑定角色“hive_hbase_create”和“hive_hbase_submit”，用于创建SparkSQL表和HBase表。

  d. 以客户端安装用户登录安装客户端的节点。

  e. 执行以下命令，配置环境变量。

  source /opt/client/bigdata_env

  source /opt/client/Spark2x/component_env

  f. 执行以下命令，认证用户。

  kinit hbase_creates_user

  g. 执行以下命令，进入Spark JDBCServer客户端shell环境：

  /opt/client/Spark2x/spark/bin/beeline -u "jdbc:hive2://<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;user.principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;"

  h. 执行以下命令，同时在SparkSQL和HBase中创建表。例如创建表hbaseTable。

  create table hbaseTable (id string, name string, age int) using org.apache.spark.sql.hbase.HBaseSource options (hbaseTableName "table1", keyCols "id", colsMapping = ", name=cf1.cq1, age=cf1.cq2");

创建好的SparkSQL表和HBase表分别保存在Hive的数据库“default”和HBase的命名空间“default”。

i. 在Manager角色界面创建一个角色，例如“hive_hbase_select”，并授予查询SparkSQL on HBase表hbaseTable和HBase表hbaseTable的权限。

- 在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global > default”，勾选表hbaseTable的“读”，单击“确定”保存，授予HBase角色查询表的权限。
- 编辑角色，在“配置资源权限”的表格中选择“*待操作集群的名称* > HBase > HBase Scope > global > hbase”，勾选表“hbase:meta”的“执行”，单击“确定”保存。
- 编辑角色，在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限 > default”，勾选表hbaseTable的“查询”，单击“确定”保存。

j. 在Manager用户界面创建一个“人机”用户，例如“hbase_select_user”，加入“hive”组，绑定角色“hive_hbase_select”，用于查询SparkSQL表和HBase表。

k. 执行以下命令，配置环境变量。

source /opt/client/bigdata_env

source /opt/client/Spark2x/component_env

l. 执行以下命令，认证用户。

kinit hbase_select_user

m. 执行以下命令，进入Spark JDBCServer客户端shell环境：

/opt/client/Spark2x/spark/bin/beeline -u "jdbc:hive2://<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;user.principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;"

n. 执行以下命令，使用SparkSQL语句查询HBase表的数据。

select * from hbaseTable;

12.23.2.6.5 客户端和服务端配置

SparkSQL权限管理功能相关的配置如下所示，客户端与服务端的配置相同。要使用表权限功能，需要在服务端和客户端添加如下配置。

- “spark-defaults.conf”配置文件

表 12-383 参数说明（1）

参数	描述	默认值
spark.sql.authorization.enabled	是否开启datasource语句的权限认证功能。建议将此参数修改为true，开启权限认证功能。	true

- “hive-site.xml”配置文件

  表 12-384 参数说明（2）

参数	描述	默认值
hive.metastore.uris	Hive组件中MetaStore服务的地址，如“thrift://10.10.169.84:21088,thrift://10.10.81.37:21088”	-
hive.metastore.sasl.enabled	MetaStore服务是否使用SASL安全加固。表权限功能需要设置为“true”。	true
hive.metastore.kerberos.principal	Hive组件中MetaStore服务的Principal，如“hive/hadoop.*<系统域名>*@*<系统域名>*”。	hive-metastore/_HOST@EXAMPLE.COM
hive.metastore.thrift.sasl.qop	开启SparkSQL权限管理功能后，需将此参数设置为“auth-conf”。	auth-conf
hive.metastore.token.signature	MetaStore服务对应的token标识，设为“HiveServer2ImpersonationToken”。	HiveServer2ImpersonationToken
hive.security.authenticator.manager	Hive客户端授权的管理器，需设为“org.apache.hadoop.hive.ql.security.SessionStateUserGroupAuthenticator”。	org.apache.hadoop.hive.ql.security.SessionStateUserMSGroupAuthenticator
hive.security.authorization.enabled	是否开启客户端的授权，需设为“true”。	true
hive.security.authorization.createtable.owner.grants	将哪些权限赋给创建表的owner，建议设置为“ALL”。	ALL

- MetaStore服务的core-site.xml配置文件

表 12-385 参数说明 (3)

参数	描述	默认值
hadoop.proxyuser.spark.hosts	允许Spark用户伪装成来自哪些host的用户，需设为“*”，代表所有节点。	-
hadoop.proxyuser.spark.groups	允许Spark用户伪装成哪些用户组的用户，需设为“*”，代表所有用户组。	-

12.23.2.7 场景化参数

12.23.2.7.1 配置多主实例模式

配置场景

集群中支持同时共存多个ThriftServer服务，通过客户端可以随机连接其中的任意一个服务进行业务操作。即使集群中一个或多个ThriftServer服务停止工作，也不影响用户通过同一个客户端接口连接其他正常的ThriftServer服务。

配置描述

登录Manager，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索并修改以下参数。

表 12-386 多主实例参数说明

参数	说明	默认值
spark.thriftserver.zookeeper.connection.timeout	Zookeeper客户端连接超时时间，单位毫秒。	60000
spark.thriftserver.zookeeper.session.timeout	Zookeeper客户端会话超时时间，单位毫秒。	90000
spark.thriftserver.zookeeper.retry.times	Zookeeper客户端失联后，重试次数。	3
spark.yarn.queue	JDBCServer服务所在的Yarn队列。	default

12.23.2.7.2 配置多租户模式

配置场景

多租户模式是将JDBCServer和租户绑定，每一个租户对应一个或多个JDBCServer，一个JDBCServer只给一个租户提供服务。不同的租户可以配置不同的Yarn队列，从而达到资源隔离。

配置描述

登录Manager，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索并修改以下参数。

表 12-387 参数说明

参数	说明	默认值
spark.proxyserver.hash.enabled	是否使用Hash算法连接ProxyServer。 ● true为使用Hash算法，使用多租户模式时，该参数需配置为true。 ● false为使用随机连接，多主实例模式，配置为false。	true **说明** 该参数修改后需要重新下载客户端。
spark.thriftserver.proxy.enabled	是否使用多租户模式。 ● false表示使用多实例模式 ● true表示使用多租户模式	true
spark.thriftserver.proxy.maxThriftServerPerTenancy	多租户模式下，一个租户可启动JDBCServer实例的最大个数。	1
spark.thriftserver.proxy.maxSessionPerThriftServer	多租户模式下，单个JDBCServer实例的session数量超过该值时，如果租户的JDBCServer最大实例数量没超过限制，则启动新的JDBCServer，否则输出警告日志。	50
spark.thriftserver.proxy.sessionWaitTime	多租户模式下，当JDBCServer的session连接数为0时，停止JDBCServer前的等待时间。	180000
spark.thriftserver.proxy.sessionThreshold	多租户模式下，当JDBCServer的session使用率（公式：当前session数 / (spark.thriftserver.proxy.maxSessionPerThriftServer * 当前JDBCServer个数)）达到阈值时，自动新增JDBCServer。	100
spark.thriftserver.proxy.healthcheck.period	多租户模式下，JDBCServer代理检查JDBCServer健康状态周期。	60000
spark.thriftserver.proxy.healthcheck.recheckTimes	多租户模式下，JDBCServer代理检查JDBCServer健康状态失败后重试次数。	3
spark.thriftserver.proxy.healthcheck.waitTime	多租户模式下，JDBCServer代理发送健康检查，等待JDBCServer响应的超时时间。	10000
spark.thriftserver.proxy.session.check.interval	多租户模式下，JDBCServer代理检查session的周期。	6h

参数	说明	默认值
spark.thriftserver.proxy.idle.session.timeout	多租户模式下，JDBCServer代理session的空闲超时时间。如果在这段时间内没有做任何操作，session会被关闭。	7d
spark.thriftserver.proxy.idle.session.check.operation	多租户模式下，JDBCServer代理session的过期是否要判断该session上还存在operation。	true
spark.thriftserver.proxy.idle.operation.timeout	多租户模式下，operation的超时时间。如果operation超时，operation会被关闭。	5d

12.23.2.7.3 配置多主实例与多租户模式切换

配置场景

在使用集群中，如果需要在多主实例模式与多租户模式之间切换，则还需要进行如下参数的设置。

- 多租户切换成多主实例模式

  修改Spark2x服务的以下参数：

  - spark.thriftserver.proxy.enabled=false
  - spark.scheduler.allocation.file=#{conf_dir}/fairscheduler.xml
  - spark.proxyserver.hash.enabled=false
- 多主实例切换成多租户模式

  修改Spark2x服务的以下参数：

  - spark.thriftserver.proxy.enabled=true
  - spark.scheduler.allocation.file=./__spark_conf__/__hadoop_conf__/fairscheduler.xml
  - spark.proxyserver.hash.enabled=true

配置描述

登录Manager，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索并修改以下参数。

表 12-388 参数说明

参数	说明	默认值
spark.thriftserver.proxy.enabled	是否使用多租户模式。 - false表示使用多实例模式 - true表示使用多租户模式	true

参数	说明	默认值
spark.scheduler.allocation.file	公平调度文件路径。 ● 多主实例配置为：#{conf_dir}/fairscheduler.xml ● 多租户配置为：./__spark_conf__/__hadoop_conf__/fairscheduler.xml	./__spark_conf__/__hadoop_conf__/fairscheduler.xml
spark.proxyserver.hash.enabled	是否使用Hash算法连接ProxyServer。 ● true为使用Hash算法，使用多租户模式时，该参数需配置为true。 ● false为使用随机连接，多主实例模式，配置为false。	true **说明** 该参数修改后需要重新下载客户端。

12.23.2.7.4 配置事件队列的大小

配置场景

Spark中见到的UI、EventLog、动态资源调度等功能都是通过事件传递实现的。事件有SparkListenerJobStart、SparkListenerJobEnd等，记录了每个重要的过程。

每个事件在发生后都会保存到一个队列中，Driver在创建SparkContext对象时，会启动一个线程循环的从该队列中依次拿出一个事件，然后发送给各个Listener，每个Listener感知到事件后就会做各自的处理。

因此当队列存放的速度大于获取的速度时，就会导致队列溢出，从而丢失了溢出的事件，影响了UI、EventLog、动态资源调度等功能。所以为了更灵活的使用，在这边添加一个配置项，用户可以根据Driver的内存大小设置合适的值。

配置描述

参数入口：

在执行应用之前，在Spark服务配置中修改。在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”。在搜索框中输入参数名称。

表 12-389 参数说明

参数	描述	默认值
spark.scheduler.listenerbus.eventqueue.capacity	事件队列的大小，可以根据Driver的内存做适当的配置。	1000000

说明

当Driver日志中出现如下的日志时，表示队列溢出了。

1. 普通应用：
   ```
   Dropping SparkListenerEvent because no remaining room in event queue.
   This likely means one of the SparkListeners is too slow and cannot keep
   up with the rate at which tasks are being started by the scheduler.
   ```
2. Spark Streaming应用：
   ```
   Dropping StreamingListenerEvent because no remaining room in event queue.
   This likely means one of the StreamingListeners is too slow and cannot keep
   up with the rate at which events are being started by the scheduler.
   ```

12.23.2.7.5 配置 executor 堆外内存大小

配置场景

当分配的内存太小或者被更高优先级的进程抢占资源时，会出现物理内存超限的情况。调整如下参数，可以防止物理内存超限。

配置描述

参数入口：

在应用提交时通过“--conf”设置这些参数，或者在客户端的“spark-defaults.conf”配置文件中调整如下参数。

表 12-390 参数说明

参数	说明	默认值
spark.executor.memoryOverhead	用于指定每个executor的堆外内存大小(MB)，增大该参数值，可以防止物理内存超限。该值是通过max(384,executor-memory*0.1)计算所得，最小值为384。	1024

12.23.2.7.6 增强有限内存下的稳定性

配置场景

当前Spark SQL执行一个查询时需要使用大量的内存，尤其是在做聚合（Aggregate）和关联（Join）操作时，此时如果内存有限的情况下就很容易出现OutOfMemoryError。有限内存下的稳定性就是确保在有限内存下依然能够正确执行相关的查询，而不出现OutOfMemoryError。

说明

有限内存并不意味着内存无限小，它只是在内存不足于放下大于内存可用总量几倍的数据时，通过利用磁盘来做辅助从而确保查询依然稳定执行，但依然有一些数据是必须留在内存的，如在做涉及到Join的查询时，对于当前用于Join的相同key的数据还是需要放在内存中，如果该数据量较大而内存较小依然会出现OutOfMemoryError。

有限内存下的稳定性涉及到3个子功能：

1. ExternalSort

外部排序功能，当执行排序时如果内存不足会将一部分数据溢出到磁盘中。

2. TungstenAggregate

   新Hash聚合功能，默认对数据调用外部排序进行排序，然后再进行聚合，因此内存不足时在排序阶段会将数据溢出到磁盘，在聚合阶段因数据有序，在内存中只保留当前key的聚合结果，使用的内存较小。

3. SortMergeJoin、SortMergeOuterJoin

   基于有序数据的等值连接。该功能默认对数据调用外部排序进行排序，然后再进行等值连接，因此内存不足时在排序阶段会将数据溢出到磁盘，在连接阶段因数据有序，在内存中只保留当前相同key的数据，使用的内存较小。

配置描述

参数入口：

在应用提交时通过“--conf”设置这些参数，或者在客户端的“spark-defaults.conf”配置文件中调整如下参数。

表 12-391 参数说明

参数	场景	描述	默认值
spark.sql.tungsten.enabled	/	类型为Boolean。 ● 当设置的值等于true时，表示开启tungsten功能，即逻辑计划等同于开启codegeneration，同时物理计划使用对应的tungsten执行计划。 ● 当设置的值等于false时，表示关闭tungsten功能。	true
spark.sql.codegen.wholeStage		类型为Boolean。 ● 当设置的值等于true时，表示开启codegeneration功能，即运行时对于某些特定的查询将动态生成各逻辑计划代码。 ● 当设置的值等于false时，表示关闭codegeneration功能，运行时使用当前已有静态代码。	true

说明

1. 开启ExternalSort除配置spark.sql.planner.externalSort=true外，还需配置spark.sql.unsafe.enabled=false或者spark.sql.codegen.wholeStage =false。
2. 如果您需要开启TungstenAggregate，有如下几种方式：

   将spark.sql.codegen.wholeStage 和spark.sql.unsafe.enabled的值都设置为true（通过配置文件或命令行方式设置）。

   如果spark.sql.codegen.wholeStage 和spark.sql.unsafe.enabled都不为true或者其中一个不为true，只要spark.sql.tungsten.enabled的值设置为true时，TungstenAggregate会开启。

12.23.2.7.7 配置 WebUI 上查看聚合后的 container 日志

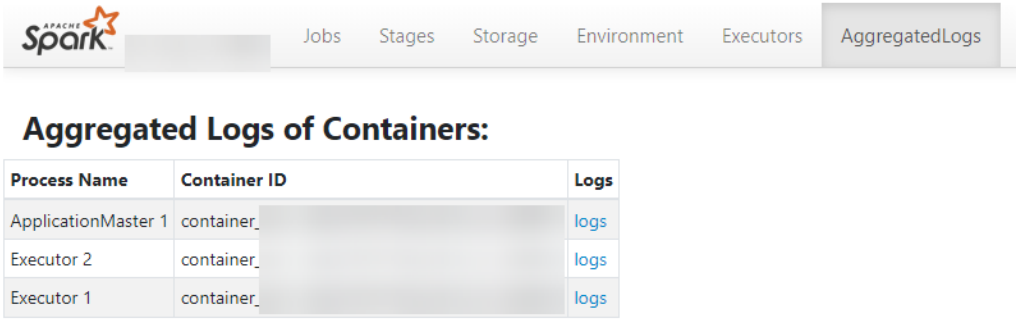
配置场景

当Yarn配置“yarn.log-aggregation-enable”为“true”时，就开启了container日志聚合功能。日志聚合功能是指：当应用在Yarn上执行完成后，NodeManager将本节点中所有container的日志聚合到HDFS中，并删除本地日志。详情请参见**配置Container日志聚合功能**。

然而，开启container日志聚合功能之后，其日志聚合至HDFS目录中，只能通过获取HDFS文件来查看日志。开源Spark和Yarn服务不支持通过WebUI查看聚合后的日志。

因此，Spark在此基础上进行了功能增强。如**图12-53**所示，在HistoryServer页面添加“AggregatedLogs”页签，可以通过“logs”链接查看聚合的日志。

图 12-53 聚合日志显示页面



配置描述

为了使WebUI页面显示日志，需要将聚合日志进行解析和展现。Spark是通过Hadoop的JobHistoryServer来解析聚合日志的，所以您可以通过“spark.jobhistory.address”参数，指定JobHistoryServer页面地址，即可完成解析和展现。

参数入口：

在应用提交时通过“--conf”设置这些参数，或者在客户端的“spark-defaults.conf”配置文件中调整如下参数。

说明

- 此功能依赖Hadoop中的JobHistoryServer服务，所以使用聚合日志之前需要保证JobHistoryServer服务已经运行正常。
- 如果参数值为空，“AggregatedLogs”页签仍然存在，但是无法通过logs链接查看日志。
- 只有当App已经running，HDFS上已经有该App的事件日志文件时才能查看到聚合的container日志。
- 正在运行的任务的日志，用户可以通过“Executors”页面的日志链接进行查看，任务结束后日志会汇聚到HDFS上，“Executors”页面的日志链接就会失效，此时用户可以通过“AggregatedLogs”页面的logs链接查看聚合日志。

表 12-392 参数说明

参数	描述	默认值
spark.jobhistory.address	JobHistoryServer页面的地址，格式：*http(s)://ip:port/jobhistory*。例如，将参数值设置为“https://10.92.115.1:26014/jobhistory”。 默认值为空，表示不能从WebUI查看container聚合日志。 修改参数后，需重启服务使得配置生效。	-

12.23.2.7.8 配置 YARN-Client 和 YARN-Cluster 不同模式下的环境变量

配置场景

当前，在YARN-Client和YARN-Cluster模式下，两种模式的客户端存在冲突的配置，即当客户端为一种模式的配置时，会导致在另一种模式下提交任务失败。

为避免出现如上情况，添加**表12-393**中的配置项，避免两种模式下来回切换参数，提升软件易用性。

- YARN-Cluster模式下，优先使用新增配置项的值，即服务端路径和参数。
- YARN-Client模式下，直接使用原有的三个配置项的值。
  原有的三个配置项为：“spark.driver.extraClassPath”、“spark.driver.extraJavaOptions”、“spark.driver.extraLibraryPath”。

说明

不添加**表12-393**中配置项时，使用方式与原有方式一致，程序可正常执行，只是在不同模式下需切换配置。

配置参数

参数入口：

在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，在搜索框中输入参数名称。

表 12-393 参数介绍

参数	描述	默认值
spark.yarn.cluster.driver.extraClassPath	YARN-Cluster模式下，Driver使用的extraClassPath，配置为服务端的路径和参数。 同时，“spark.driver.extraClassPath”配置成Spark客户端路径，可以保证在YARN-Client模式下和YARN-Cluster模式下不需要切换配置。	${BIGDATA_HOME}/common/runtime/security

参数	描述	默认值
spark.yarn.cluster.driver.extraJavaOptions	YARN-Cluster模式下Driver的extraJavaOptions，配置成服务端的路径和参数。 同时，“spark.driver.extraJavaOptions”配置成Spark客户端路径，可以保证YARN-Client模式和YARN-Cluster模式不需要切换配置。	-Xloggc:<LOG_DIR>/indexserver-%p-gc.log -XX:+PrintGCDetails -XX:-OmitStackTraceInFastThrow -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=20 -XX:GCLogFileSize=10M -Dlog4j.configuration=./__spark_conf__/__hadoop_conf__/log4j-executor.properties -Dlog4j.configuration.watch=true -Djava.security.auth.login.config=./__spark_conf__/__hadoop_conf__/jaas-zk.conf -Dzookeeper.server.principal=${ZOOKEEPER_SERVER_PRINCIPAL} -Djava.security.krb5.conf=./__spark_conf__/__hadoop_conf__/kdc.conf -Djetty.version=x.y.z -Dorg.xerial.snappy.tempdir=${BIGDATA_HOME}/tmp -Dcarbon.properties.filepath=./__spark_conf__/__hadoop_conf__/carbon.properties -Djdk.tls.ephemeralDHKeySize=2048 -Dspark.ssl.keyStore=./child.keystore #{java_stack_prefer}

12.23.2.7.9 配置 SparkSQL 的分块个数

配置场景

SparkSQL在进行shuffle操作时默认的分块数为200。在数据量特别大的场景下，使用默认的分块数就会造成单个数据块过大。如果一个任务产生的单个shuffle数据块大于2G，该数据块在被fetch的时候还会报类似错误：

```
Adjusted frame length exceeds 2147483647: 2717729270 - discarded
```

例如，SparkSQL运行TPCDS 500G的测试时，使用默认配置出现错误。所以当数据量较大时需要适当的调整该参数。

配置参数

参数入口：

在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”。在搜索框中输入参数名称。

表 12-394 参数介绍

参数	描述	默认值
spark.sql.shuffle.partitions	SparkSQL在进行shuffle操作时默认的分块数。	200

12.23.2.7.10 配置 parquet 表的压缩格式

配置场景

当前版本对于parquet表的压缩格式分以下两种情况进行配置：

1. 对于分区表，需要通过parquet本身的配置项“parquet.compression”设置parquet表的数据压缩格式。如在建表语句中设置tblproperties："parquet.compression"="snappy"。
2. 对于非分区表，需要通过“spark.sql.parquet.compression.codec”配置项来设置parquet类型的数据压缩格式。直接设置“parquet.compression”配置项是无效的，因为它会读取“spark.sql.parquet.compression.codec”配置项的值。当“spark.sql.parquet.compression.codec”未做设置时默认值为“snappy”，“parquet.compression”会读取该默认值。

因此，“spark.sql.parquet.compression.codec”配置项只适用于设置非分区表的parquet压缩格式。

配置参数

参数入口：

在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，在搜索框中输入参数名称。

表 12-395 参数介绍

参数	描述	默认值
spark.sql.parquet.compression.codec	对于非分区parquet表，设置其存储文件的压缩格式。	snappy

12.23.2.7.11 配置 WebUI 上显示的 Lost Executor 信息的个数

配置场景

Spark WebUI中“Executor”页面支持展示Lost Executor的信息，对于JDBCServer长任务来说，Executor的动态回收是常态，Lost Executor个数太多，会撑爆“Executor”页面，因此需要控制页面显示的Lost Executor个数。

配置描述

在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-396 参数说明

参数	说明	默认值
spark.ui.retainedDeadExecutors	Spark UI页面显示的Lost Executor的最大个数。	100

12.23.2.7.12 动态设置日志级别

配置场景

在某些场景下，当任务已经启动后，用户想要修改日志级别以定位问题或者查看想要的信息。

用户可以在进程启动前，在进程的JVM参数中增加参数“-Dlog4j.configuration.watch=true”来打开动态设置日志级别的功能。进程启动后，就可以通过修改进程对应的log4j配置文件，来调整日志打印级别。

目前支持动态设置日志级别功能的有：Driver日志、Executor日志、AM日志、JobHistory日志、JDBCServer日志。

允许设置的日志级别是：FATAL，ERROR，WARN，INFO，DEBUG，TRACE和ALL。

配置描述

在进程对应的JVM参数配置项中增加以下参数。

表 12-397 参数描述

参数	描述	默认值
-Dlog4j.configuration.watch	进程JVM参数，设置成“true”用于打开动态设置日志级别功能。	未配置，即为false。

Driver、Executor、AM进程的JVM参数如表12-398所示。在Spark客户端的配置文件“spark-defaults.conf”中进行配置。Driver、Executor、AM进程的日志级别在对应的JVM参数中的“-Dlog4j.configuration”参数指定的log4j配置文件中设置。

表 12-398 进程的 JVM 参数 1

参数	说明	默认日志级别
spark.driver.extraJavaOptions	Driver的JVM参数。	INFO
spark.executor.extraJavaOptions	Executor的JVM参数。	INFO
spark.yarn.am.extraJavaOptions	AM的JVM参数。	INFO

JobHistory Server和JDBCServer的JVM参数如**表12-399**所示。在服务端配置文件“ENV_VARS”中进行配置。JobHistory Server和JDBCServer的日志级别在服务端配置文件“log4j.properties”中设置。

表 12-399 进程的 JVM 参数 2

参数	说明	默认日志级别
GC_OPTS	JobHistory Server的JVM参数。	INFO
SPARK_SUBMIT_OPTS	JDBCServer的JVM参数。	INFO

示例：

为了动态修改Executor日志级别为DEBUG，在进程启动之前，修改“spark-defaults.conf”文件中的Executor的JVM参数“spark.executor.extraJavaOptions”，增加如下配置：

```
-Dlog4j.configuration.watch=true
```

提交用户应用后，修改“spark.executor.extraJavaOptions”中“-Dlog4j.configuration”参数指定的log4j日志配置文件（例如：“-Dlog4j.configuration=file:${BIGDATA_HOME}/FusionInsight_Spark2x_8.1.0.1/install/FusionInsight-Spark2x-3.1.1/spark/conf/log4j-executor.properties”）中的日志级别为DEBUG，如下所示：

```
log4j.rootCategory=DEBUG, sparklog
```

DEBUG级别生效会有一定的时延。

12.23.2.7.13 配置 Spark 是否获取 HBase Token

配置场景

使用Spark提交任务时，Driver默认会去HBase获取Token，访问HBase则需要配置文件“jaas.conf”进行安全认证。此时若用户未配置“jaas.conf”文件，会导致应用运行失败。

因此，根据应用是否涉及HBase进行以下处理：

- 当应用不涉及HBase时，即无需获取HBase Token。此时，将“spark.yarn.security.credentials.hbase.enabled”设置为“false”即可。
- 当应用涉及HBase时，将“spark.yarn.security.credentials.hbase.enabled”设置为“true”，且需要在Driver端配置“jaas.conf”文件，配置如下：

  ```
  {client}/spark/bin/spark-sql  --master yarn-client --principal {principal} --keytab {keytab} --driver-java-options "-Djava.security.auth.login.config={LocalPath}/jaas.conf"
  ```

  在“jaas.conf”中指定Keytab和Prinical，示例如下：

  ```
  Client {
  com.sun.security.auth.module.Krb5LoginModule required
  useKeyTab=true
  keyTab = "{LocalPath}/user.keytab"
  principal="super@<系统域名>"
  useTicketCache=false
  debug=false;
  };
  ```

配置描述

在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-400 参数说明

参数	说明	默认值
spark.yarn.security.credentials.hbase.enabled	HBase是否获取Token： ● true：获取 ● false：不获取	false

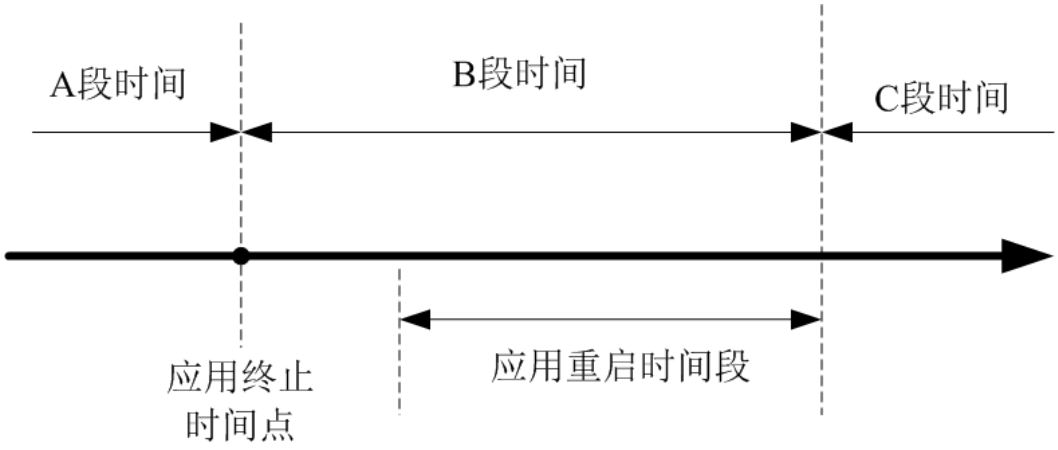
12.23.2.7.14 配置 Kafka 后进先出

配置场景

当Spark Streaming应用与Kafka对接，Spark Streaming应用异常终止并从checkpoint恢复重启后，对于进入Kafka数据的任务，系统默认优先处理应用终止前（A段时间）未完成的任务和应用终止到重启完成这段时间内（B段时间）进入Kafka数据生成的任务，最后再处理应用重启完成后（C段时间）进入Kafka数据生成的任务。并且对于B段时间进入Kafka的数据，Spark将按照终止时间（batch时间）生成相应个数的任务，其中第一个任务读取全部数据，其余任务可能不读取数据，造成任务处理压力不均匀。

若A段时间的任务和B段时间任务处理得较慢，则会影响C段时间任务的处理。针对上述场景，Spark提供Kafka后进先出功能。

图 12-54 Spark Streaming 应用重启时间轴



开启此功能后，Spark将优先调度C段时间内的任务，若存在多个C段任务，则按照任务产生的先后顺序调度执行，再执行A段时间和B段时间的任务。另外，对于B段时间进入Kafka的数据，Spark除了按照终止时间生成相应任务，还将这个期间进入Kafka的所有数据均匀分配到各个任务，避免任务处理压力不均匀。

约束条件：

- 目前该功能只适用于Spark Streaming中的Direct方式，且执行结果与上一个batch时间处理结果没有依赖关系（即无state操作，如updatestatebykey）。对多条数据输入流，需要相对独立无依赖的状态，否则可能导致数据切分后结果发生变化。
- Kafka后进先出功能的开启要求应用只能对接Kafka输入源。
- 若提交应用的同时开启Kafka后进先出和流控功能，对于B段时间进入Kafka的数据，将不启动流控功能，以确保读取这些数据的任务调度优先级最低。应用重新启动后C段时间的任务启用流控功能。

配置描述

在Spark Driver端的“spark-defaults.conf”配置文件中进行设置。

表 12-401 参数说明

参数	说明	默认值
spark.streaming.kafka.direct.lifo	配置是否开启Kafka后进先出功能。	false
spark.streaming.kafka010.inputstream.class	获取解耦在FusionInsight侧的类	org.apache.spark.streaming.kafka010.HWDirectKafkaInputDStream

12.23.2.7.15 配置对接 Kafka 可靠性

配置场景

Spark Streaming对接Kafka时，当Spark Streaming应用重启后，应用根据上一次读取的topic offset作为起始位置和当前topic最新的offset作为结束位置从Kafka上读取数据的。

Kafka服务的topic的leader异常后，若Kafka的leader和follower的offset相差太大，用户重启Kafka服务，Kafka的follower和leader相互切换，则Kafka服务重启后，topic的offset变小。

- 若Spark Streaming应用一直在运行，由于Kafka上topic的offset变小，会导致读取Kafka数据的起始位置比结束位置大，这样将无法从Kafka读取数据，应用报错。
- 若在重启Kafka服务前，先停止Spark Streaming应用，等Kafka重启后，再重启Spark Streaming应用使应用从checkpoint恢复。此时，Spark Streaming应用会记录终止前读取到的offset位置，以此为基准读取后面的数据，而Kafka offset变小（例如从10万变成1万），Spark Streaming会等待Kafka leader的offset增长至10万之后才会去消费，导致新发送的offset在1万至10万之间的数据丢失。

针对上述背景，提供配置Streaming对接Kafka更高级别的可靠性。对接Kafka可靠性功能开启后，上述场景处理方式如下。

- 若Spark Streaming应用在运行应用时Kafka上topic的offset变小，则会将Kafka上topic最新的offset作为读取Kafka数据的起始位置，继续读取后续的数据。

  对于已经生成但未调度处理的任务，若读取的Kafka offset区间大于Kafka上topic的最新offset，则该任务会运行失败。

  说明

  若任务失败过多，则会将executor加入黑名单，从而导致后续的任务无法部署运行。此时用户可以通过配置“spark.blacklist.enabled”参数关闭黑名单功能，黑名单功能默认为开启。
- 若Kafka上topic的offset变小后，Spark Streaming应用进行重启恢复终止前未处理完的任务若读取的Kafka offset区间大于Kafka上topic的最新offset，则该任务直接丢弃，不进行处理。

说明

若Streaming应用中使用了state函数，则不允许开启对接Kafka可靠性功能。

配置描述

在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-402 参数说明

参数	说明	默认值
spark.streaming.Kafka.reliability	Spark Streaming对接Kafka是否开启可靠性功能： ● true：开启可靠性功能 ● false：不开启可靠性功能	false

12.23.2.7.16 配置流式读取 driver 执行结果

配置场景

在执行查询语句时，返回结果有可能会很大（10万数量以上），此时很容易导致JDBCServer OOM（Out of Memory）。因此，提供数据汇聚功能特性，在基本不牺牲性能的情况下尽力避免OOM。

配置描述

提供两种不同的数据汇聚功能配置选项，两者在Spark JDBCServer服务端的tunning选项中进行设置，设置完后需要重启JDBCServer。

表 12-403 参数说明

参数	说明	默认值
spark.sql.bigdata.thriftServer.useHdfsCollect	是否将结果数据保存到HDFS中而不是内存中。 优点：由于查询结果保存在hdfs端，因此基本不会造成JDBCServer的OOM。 缺点：速度慢。 ● true：保存至HDFS中 ● false：不使用该功能 **须知** spark.sql.bigdata.thriftServer.useHdfsCollect参数设置为true时，将结果数据保存到HDFS中，但JobHistory原生页面上Job的描述信息无法正常关联到对应的SQL语句，同时spark-beeline命令行中回显的Execution ID为null，为解决JDBCServer OOM问题，同时显示信息正确，建议选择 spark.sql.userlocalFileCollect参数进行配置。	false
spark.sql.uselocalFileCollect	是否将结果数据保存在本地磁盘中而不是内存里面。 优点：结果数据小数据量情况下和原生内存的方式相比性能损失可以忽略，大数据情况下（亿级数据）性能远比使用hdfs，以及原生内存方式好。 缺点：需要调优。大数据情况下建议JDBCServer driver端内存10G，executor端每个核心分配3G内存。 ● true：使用该功能 ● false: 不使用该功能	false

参数	说明	默认值
spark.sql.collect.Hive	该参数在spark.sql.uselocalFileCollect开启的情况下生效。直接序列化的方式，还是间接序列化的方式保存结果数据到磁盘。 优点：针对分区数特别多的表查询结果汇聚性能优于直接使用结果数据保证在磁盘的方式。 缺点：和spark.sql.uselocalFileCollect开启时候的缺点一样。 ● true：使用该功能 ● false：不使用该功能	false
spark.sql.collect.serialize	该参数在spark.sql.uselocalFileCollect，spark.sql.collect.Hive同时开启的情况下生效。 作用是进一步提升性能 ● java：采用java序列化方式收集数据。 ● kryo：采用kryo序列化方式收集数据，性能要比采用java好。	java

说明

参数spark.sql.bigdata.thriftServer.useHdfsCollect和spark.sql.uselocalFileCollect不能同时设置为true。

12.23.2.7.17 配置过滤掉分区表中路径不存在的分区

配置场景

当读取HIVE分区表时，如果指定的分区路径在HDFS上不存在，则执行***select***查询时会报FileNotFoundException异常。此时可以通过配置“spark.sql.hive.verifyPartitionPath”参数来过滤掉分区路径不存在的分区，来避免读取时报错。

配置描述

可以通过以下两种方式配置是否过滤掉分区表分区路径不存在的分区。

- 在Spark Driver端的“spark-defaults.conf”配置文件中进行设置。

  表 12-404 参数说明

参数	说明	默认值
spark.sql.hive.verifyPartitionPath	配置读取HIVE分区表时，是否过滤掉分区表分区路径不存在的分区。 “true”：过滤掉分区路径不存在的分区； “false”：不进行过滤。	false

- 在spark-submit命令提交应用时，通过“--conf”参数配置是否过滤掉分区表分区路径不存在的分区。

  示例：

  ```
  spark-submit --class org.apache.spark.examples.SparkPi  --conf spark.sql.hive.verifyPartitionPath=true $SPARK_HOME/lib/spark-examples_*.jar
  ```

12.23.2.7.18 配置 Spark2x Web UI ACL

配置场景

当Spark2x Web UI中有一些不允许其他用户看到的数据时，用户可能想对UI进行安全防护。用户一旦登录，Spark2x可以比较与这个用户相对应的视图ACLs来确认是否授权用户访问 UI。

Spark2x存在两种类型的Web UI，一种为运行中任务的Web UI，可以通过Yarn原生页面的应用链接或者REST接口访问。一种为已结束任务的Web UI，可以通过Spark2x JobHistory服务或者REST接口访问。

说明

本章节仅支持安全模式（开启了Kerberos认证）集群。

- 运行中任务Web UI ACL配置。

  运行中的任务，可通过服务端对如下参数进行配置。

  - “spark.admin.acls”：指定Web UI的管理员列表。
  - “spark.admin.acls.groups”：指定管理员组列表。
  - “spark.ui.view.acls”：指定yarn界面的访问者列表。
  - “spark.modify.acls.groups”：指定yarn界面的访问者组列表。
  - “spark.modify.acls”：指定Web UI的修改者列表。
  - “spark.ui.view.acls.groups”：指定Web UI的修改者组列表。

- 运行结束后Web UI ACL配置。

  运行结束的任务通过客户端的参数“spark.history.ui.acls.enable”控制是否开启ACL访问权限。

  如果开启了ACL控制，由客户端的“spark.admin.acls”和s“park.admin.acls.groups”配置指定Web UI的管理员列表和管理员组列表，由客户端的“spark.ui.view.acls”和“spark.modify.acls.groups”配置指定查看Web UI任务明细的访问者列表和组列表，由客户端的“spark.modify.acls”和“spark.ui.view.acls.groups”配置指定修改Web UI任务明细的访问者列表和组列表。

配置描述

登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索acl，在对应的JobHistory，JDBCServer，SparkResource和Spark界面修改以下参数。

表 12-405 参数说明

参数	说明	默认值
spark.history.ui.acls.enable	配置JobHistory是否支持单一任务的权限校验。	true
spark.acls.enable	配置是否开启spark权限管理。 如果开启，将会检查用户是否有权限访问和修改任务信息。	true
spark.admin.acls	配置spark管理员列表，列表中成员有权限管理所有spark任务，此处可以配置多个管理员用户，使用“，”分隔。	admin
spark.admin.acls.groups	配置spark管理组列表，列表中的组有权限管理所有spark任务，此处可以配置多个管理组，使用“，”分隔。	-
spark.modify.acls	配置有权限修改spark任务的成员列表。启动任务的用户默认有此权限，此处可以配置多个用户，使用“，”分隔。	-
spark.modify.acls.groups	配置有权限修改spark任务的组列表，此处可以配置多个组，使用“，”分隔。	-
spark.ui.view.acls	配置有权限访问spark任务的成员列表。启动任务的用户默认有此权限，此处可以配置多个用户，使用“，”分隔。	-
spark.ui.view.acls.groups	配置有权限访问spark任务的组列表，此处可以配置多个组，使用“，”分隔。	-

说明

若使用客户端提交任务，“spark.admin.acls”、“spark.admin.acls.groups”、“spark.modify.acls”、“spark.modify.acls.groups”、“spark.ui.view.acls”和“spark.ui.view.acls.groups”参数修改后需要重新下载客户端。

12.23.2.7.19 配置矢量化读取 ORC 数据

配置场景

ORC文件格式是一种Hadoop生态圈中的列式存储格式，它最初产生自Apache Hive，用于降低Hadoop数据存储空间和加速Hive查询速度。和Parquet文件格式类似，它并不是一个单纯的列式存储格式，仍然是首先根据行组分割整个表，在每一个行组内按列进行存储，并且文件中的数据尽可能的压缩来降低存储空间的消耗。矢量化读取ORC格式的数据能够大幅提升ORC数据读取性能。在Spark2.3版本中，SparkSQL支持矢量化读取ORC数据（这个特性在Hive的历史版本中已经得到支持）。矢量化读取ORC格式的数据能够获得比传统读取方式数倍的性能提升。

该特性可以通过下面的配置项开启：

- “spark.sql.orc.enableVectorizedReader”：指定是否支持矢量化方式读取ORC格式的数据，默认为true。

- “spark.sql.codegen.wholeStage”：指定是否需要将多个操作的所有stage编译为一个java方法，默认为true。
- “spark.sql.codegen.maxFields”：指定codegen的所有stage所支持的最大字段数（包括嵌套字段），默认为100。
- “spark.sql.orc.impl”：指定使用Hive还是Spark SQL native作为SQL执行引擎来读取ORC数据，默认为hive。

配置参数

登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值	取值范围
spark.sql.orc.enableVectorizedReader	指定是否支持矢量化方式读取ORC格式的数据，默认为true。	true	[true,false]
spark.sql.codegen.wholeStage	指定是否需要将多个操作的所有stage编译为一个java方法，默认为true。	true	[true,false]
spark.sql.codegen.maxFields	指定codegen的所有stage所支持的最大字段数（包括嵌套字段），默认为100.	100	大于0
spark.sql.orc.impl	指定使用Hive还是Spark SQL native作为SQL执行引擎来读取ORC数据，默认为hive。	hive	[hive,native]

说明

1. 使用SparkSQL内置的矢量化方式读取ORC数据需要满足下面的条件：
   - spark.sql.orc.enableVectorizedReader : true，默认是true，一般不做修改。
   - spark.sql.codegen.wholeStage : true，默认为true，一般不做修改。
   - spark.sql.codegen.maxFields不小于scheme的列数。
   - 所有的数据类型均为AtomicType类型；所谓Atomic Type表示非NULL、UDTs、arrays，maps类型。如果列中存在这几种类型的任意一种，都无法获得预期的性能。
   - spark.sql.orc.impl : native ,默认为hive。
2. 若使用客户端提交任务，“spark.sql.orc.enableVectorizedReader”、“spark.sql.codegen.wholeStage”、“spark.sql.codegen.maxFields”、“spark.sql.orc.impl”、参数修改后需要重新下载客户端才能生效。

12.23.2.7.20 Hive 分区修剪的谓词下推增强

配置场景

在旧版本中，对Hive表的分区修剪的谓词下推，只支持列名与整数或者字符串的比较表达式的下推，在2.3版本中，增加了对null、in、and、or表达式的下推支持。

配置参数

登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值	取值范围
spark.sql.hive.advancedPartitionPredicatePushdown.enabled	用于配置是否开启Hive表的分区谓词下推增强功能。	true	[true,false]

12.23.2.7.21 支持 Hive 动态分区覆盖语义

配置场景

在旧版本中，使用insert overwrite语法覆写分区表时，只支持对指定的分区表达式进行匹配，未指定表达式的分区将被全部删除。在spark2.3版本中，增加了对未指定表达式的分区动态匹配的支持，此种语法与Hive的动态分区匹配语法行为一致。

配置参数

登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值	取值范围
spark.sql.sources.partitionOverwriteMode	当前执行insert overwrite 命令插入数据到分区表时，支持两种模式：STATIC模式和DYNAMIC模式。STATIC模式下，Spark会按照匹配条件删除所有分区。在DYNAMIC模式下，Spark按照匹配条件匹配分区，并动态匹配没有指定匹配条件的分区。	STATIC	[STATIC,DYNAMIC]

12.23.2.7.22 配置列统计值直方图 Histogram 用以增强 CBO 准确度

配置场景

Spark优化sql的执行，一般的优化规则都是启发式的优化规则，启发式的优化规则，仅仅根据逻辑计划本身的特点给出优化，没有考虑数据本身的特点，也就是未考虑算子本身的执行代价。Spark在2.2中引入了基于代价的优化规则（CBO）。CBO会收集表和列的统计信息，结合算子的输入数据集来估计每个算子的输出条数以及字节大小，这些就是执行一个算子的代价。

CBO会调整执行计划，来最小化端到端的查询时间，中心思路2点：

- 尽早过滤不相关的数据。
- 最小化每个算子的代价。

CBO优化过程分为2步：

1. 收集统计信息。
2. 根据输入的数据集估算特定算子的输出数据集。

表级别统计信息包括：记录条数；表数据文件的总大小。

列级别统计信息包括：唯一值个数；最大值；最小值；空值个数；平均长度；最大长度；直方图。

有了统计信息后，就可以估计算子的执行代价了。常见的算子包括过滤条件Filter算子和Join算子。

直方图为列统计值的一种，可以直观的描述列数据的分布情况，将列的数据从最小值到最大值划分为事先指定数量的槽位（bin），计算各个槽位的上下界的值，使得全部数据都确定槽位后，所有槽位中的数据数量相同（等高直方图）。有了数据的详细分布后，各个算子的代价估计能更加准确，优化效果更好。

该特性可以通过下面的配置项开启：

spark.sql.statistics.histogram.enabled：指定是否开启直方图功能，默认为false。

配置参数

登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值	取值范围
spark.sql.cbo.enabled	开启CBO来估计执行计划的统计值。	false	[true,false]
spark.sql.cbo.joinReorder.enabled	开启CBO连接重排序。	false	[true,false]
spark.sql.cbo.joinReorder.dp.threshold	动态规划算法中允许的最大的join节点数量。	12	>=1
spark.sql.cbo.joinReorder.card.weight	在重连接执行计划代价比较中维度（行数）所占的比重：行数 * 比重 + 文件大小 *（1 - 比重）。	0.7	0-1
spark.sql.statistics.size.autoUpdate.enabled	开启当表的数据发生变化时，自动更新表的大小信息。注意如果表的数据文件总数量非常多时，这个操作会非常耗费资源，减慢对数据的操作速度。	false	[true,false]

参数	说明	默认值	取值范围
spark.sql.statistics.histogram.enabled	开启后，当统计列信息时，会生成直方图。直方图可以提高估计准确度，但是收集直方图信息会有额外工作量。	false	[true,false]
spark.sql.statistics.histogram.numBins	生成的直方图的槽位数。	254	>=2
spark.sql.statistics.ndv.maxError	在生成列级别统计信息时，HyperLogLog++算法允许的最大估计误差。	0.05	0-1
spark.sql.statistics.percentile.accuracy	在生成等高直方图时百分位估计的准确率。该值越大意味着越准确。估计错误值可以通过（1.0 / 百分位估计的准确率）来得到。	10000	>=1

说明

- 如果希望直方图可以在CBO中生效，需要满足下面的条件：
  - spark.sql.statistics.histogram.enabled : true，默认是false，修改为true开启直方图功能。
  - spark.sql.cbo.enabled : true，默认为false，修改为true开启CBO。
  - spark.sql.cbo.joinReorder.enabled : true，默认为false，修改为true开启连接重排序。
- 若使用客户端提交任务，“spark.sql.cbo.enabled”、“spark.sql.cbo.joinReorder.enabled”、“spark.sql.cbo.joinReorder.dp.threshold”、“spark.sql.cbo.joinReorder.card.weight”、“spark.sql.statistics.size.autoUpdate.enabled”、“spark.sql.statistics.histogram.enabled”、“spark.sql.statistics.histogram.numBins”、“spark.sql.statistics.ndv.maxError”、“spark.sql.statistics.percentile.accuracy”参数修改后需要重新下载客户端才能生效。

12.23.2.7.23 配置 JobHistory 本地磁盘缓存

配置场景

JobHistory可使用本地磁盘缓存spark应用的历史数据，以防止JobHistory内存中加载大量应用数据，减少内存压力，同时该部分缓存数据可以复用以提高后续对相同应用的访问速度。

配置参数

登录FusionInsight Manager系统，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值
spark.history.store.path	JobHistory缓存历史信息的本地目录，如果设置了此配置，则JobHistory会将历史应用数据缓存在本地磁盘而不是内存中	${BIGDATA_HOME}/tmp/spark2x_JobHistory
spark.history.store.maxDiskUsage	JobHistory本地磁盘缓存的最大可用空间	10g

12.23.2.7.24 配置 Spark SQL 开启 Adaptive Execution 特性

配置场景

Spark SQL Adaptive Execution特性用于使Spark SQL在运行过程中，根据中间结果优化后续执行流程，提高整体执行效率。当前已实现的特性如下：

1. 自动设置shuffle partition数

   在启用Adaptive Execution特性前，Spark SQL根据spark.sql.shuffle.partitions配置指定shuffle时的partition个数。此种方法在一个应用中执行多种SQL查询时缺乏灵活性，无法保证所有场景下的性能合适。开启Adaptive Execution后，Spark SQL将自动为每个shuffle过程动态设置partition个数，而不是使用通用配置，使每次shuffle过程自动使用最合理的partition数。

2. 动态调整执行计划

   在启用Adaptive Execution特性前，Spark SQL根据RBO和CBO的优化结果创建执行计划，此种方法忽略了数据在运行过程中的结果集变化。比如基于某个大表创建的视图，与其他大表join时，即便视图的结果集很小，也无法将执行计划调整为BroadcastJoin。启用Adaptive Execution特性后，Spark SQL能够在运行过程中根据前面stage的运行结果动态调整后续的执行计划，从而获得更好的执行性能。

3. 自动处理数据倾斜

   在执行SQL语句时，若存在数据倾斜，可能导致单个executor内存溢出、任务执行缓慢等问题。启动Adaptive Execution特性后，Spark SQL能自动处理数据倾斜场景，对倾斜的分区，启动多个task进行处理，每个task读取若干个shuffle输出文件，再对这部分任务的Join结果进行Union操作，以达到消除数据倾斜的效果

配置参数

登录FusionInsight Manager系统，选择“集群 > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值
spark.sql.adaptive.enabled	配置是否启用自适应执行功能。 注意：AQE特性与DPP（动态分区裁剪）特性同时开启时，SparkSQL任务执行中会优先执行DPP特性，从而使得AQE特性不生效。	false

参数	说明	默认值
spark.sql.optimizer.dynamicPartitionPruning.enabled	动态分区裁剪功能的开关。	true
spark.sql.adaptive.coalescePartitions.enabled	如果配置为true并且“spark.sql.adaptive.enabled”为true，Spark将根据目标大小（由spark.sql.adaptive.advisoryPartitionSizeInBytes指定）合并连续的随机播放分区，以避免执行过多的小任务。	true
spark.sql.adaptive.coalescePartitions.initialPartitionNum	合并之前的shuffle分区的初始数量，默认等于spark.sql.shuffle.partitions。只有当spark.sql.adaptive.enabled和spark.sql.adaptive.coalescePartitions.enabled都为true时，该配置才有效。创建时可选，初始分区数必须为正数。	200
spark.sql.adaptive.coalescePartitions.minPartitionNum	合并后的最小shuffle分区数。如果不设置，默认为Spark集群的默认并行度。只有当spark.sql.adaptive.enabled 和spark.sql.adaptive.coalescePartitions.enable都为true时，该配置才有效。创建时可选，最小分区数必须为正数。	1
spark.sql.adaptive.shuffle.targetPostShuffleInputSize	shuffle后单个分区的目标大小，从Spark3.0开始不再支持。	64MB
spark.sql.adaptive.advisoryPartitionSizeInBytes	自适应优化时（spark.sql.adaptive.enabled为true时）shuffle分区的咨询大小（单位：字节），在Spark聚合小shuffle分区或拆分倾斜的shuffle分区时生效。	64MB
spark.sql.adaptive.fetchShuffleBlocksInBatch	是否批量取连续的shuffle块。对于同一个map任务，批量读取连续的shuffle块可以减少IO，提高性能，而不是逐个读取块。注意，只有当spark.sql.adaptive.enabled和spark.sql.adaptive.coalescePartitions.enabled都为true时，单次读取请求中存在多个连续块。这个特性还依赖于一个可重定位的序列化器，使用的级联支持编解码器和新版本的shuffle提取协议。	true
spark.sql.adaptive.localShuffleReader.enabled	当“true”且spark.sql.adaptive.enabled为“true”时，Spark在不需要进行shuffle分区时，会尝试使用本地shuffle reader读取shuffle数据，例如：将sort-merge join转换为broadcast-hash join后。	true

参数	说明	默认值
spark.sql.adaptive.skewJoin.enabled	当此配置为true且spark.sql.adaptive.enabled设置为true时，启用运行时自动处理join运算中的数据倾斜功能	true
spark.sql.adaptive.skewJoin.skewedPartitionFactor	此配置为一个倍数因子，用于判定分区是否为数据倾斜分区。单个分区被判定为数据倾斜分区的条件为：当一个分区的数据大小超过除此分区外其他所有分区大小的中值与该配置的乘积，并且大小超过spark.sql.adaptive.skewJoin.skewedPartitionThresholdInBytes配置值时，此分区被判定为数据倾斜分区	5
spark.sql.adaptive.skewJoin.skewedPartitionThresholdInBytes	分区大小（单位：字节）大于该阈值且大于spark.sql.adaptive.skewJoin.skewedPartitionFactor与分区中值的乘积，则认为该分区存在倾斜。理想情况下，此配置应大于spark.sql.adaptive.advisoryPartitionSizeInBytes.	256MB
spark.sql.adaptive.nonEmptyPartitionRatioForBroadcastJoin	两表进行join操作的时候，当非空分区比率低于此配置时，无论其大小如何，都不会被视为自适应执行中广播哈希连接的生成端。只有当spark.sql.adaptive.enabled为true时，此配置才有效。	0.2

12.23.2.7.25 配置 eventlog 日志回滚

配置场景

当Spark开启事件日志模式，即设置“spark.eventLog.enabled”为“true”时，就会往配置的一个日志文件中写事件，记录程序的运行过程。当程序运行很久，job很多，task很多时就会造成日志文件很大，如JDBCServer、Spark Streaming程序。

而日志回滚功能是指在写事件日志时，将元数据事件（EnviromentUpdate，BlockManagerAdded，BlockManagerRemoved，UnpersistRDD，ExecutorAdded，ExecutorRemoved，MetricsUpdate，ApplicationStart，ApplicationEnd，LogStart）写入日志文件中，Job事件（StageSubmitted， StageCompleted， TaskResubmit，TaskStart，TaskEnd， TaskGettingResult， JobStart，JobEnd）按文件的大小进行决定是否写入新的日志文件。对于Spark SQL的应用，Job事件还包含ExecutionStart、ExecutionEnd。

Spark中有个HistoryServer服务，其UI页面就是通过读取解析这些日志文件获得的。在启动HistoryServer进程时，内存大小就已经定了。因此当日志文件很大时，加载解析这些文件就可能会造成内存不足，driver gc等问题。

所以为了在小内存模式下能加载较大日志文件，需要对大应用开启日志滚动功能。一般情况下，长时间运行的应用建议打开该功能。

配置参数

登录FusionInsight Manager系统，选择“集群 > 服务 > Spark2x > 配置”，单击“全部配置”，搜索以下参数。

参数	说明	默认值
spark.eventLog.rolling.enabled	是否启用滚动event log文件。如果设置为true，则会将每个event log文件缩减到配置的大小。	true
spark.eventLog.rolling.maxFileSize	当spark.eventlog.rolling.enabled=true时，指定要滚动的event log文件的最大大小。	128M
spark.eventLog.compression.codec	用于压缩事件日志的编码解码器。默认情况下，spark提供四种编码解码器：lz4、lzf、snappy和zstd。如果没有给出，将使用spark.io.compression.codec。	无
spark.eventLog.logStageExecutorMetrics	是否将executor metrics的每个stage峰值（针对每个executor）写入event log。	false

12.23.2.8 使用 Ranger 时适配第三方 JDK

配置场景

当使用Ranger作为spark sql的权限管理服务时，访问RangerAdmin需要使用集群中的证书。若用户未使用集群中的JDK或者JRE，而是使用第三方JDK时，会出现访问RangerAdmin失败，进而spark应用程序启动失败的问题。

在这个场景下，需要进行以下操作，将集群中的证书导入第三方JDK或者JRE中。

配置方法

步骤1 导出集群中的证书：

1. 安装集群客户端，例如安装路径为“/opt/client”。
2. 执行以下命令，切换到客户端安装目录。

   cd /opt/client
3. 执行以下命令配置环境变量。

   source bigdata_env
4. 生成证书文件

keytool -export -alias fusioninsightsubroot -storepass changeit -keystore /opt/client/JRE/jre/lib/security/cacerts -file fusioninsightsubroot.crt

步骤2 将集群中的证书导入第三方JDK或者JRE中

将**步骤1**中生成的**fusioninsightsubroot.crt**文件拷贝到第三方JRE节点上，设置好该节点的JAVA_HOME环境变量后，执行以下命令导入证书：

keytool -import -trustcacerts -alias fusioninsightsubroot -storepass changeit -file fusioninsightsubroot.crt -keystore MY_JRE/lib/security/cacerts

说明

'MY_JRE'表示第三方JRE安装路径，请自行修改。

----结束

12.23.3 Spark2x 日志介绍

日志描述

日志存储路径：

- Executor运行日志：“${BIGDATA_DATA_HOME}/hadoop/data${i}/nm/containerlogs/application_${appid}/container_{$contid}”

  说明

  运行中的任务日志存储在以上路径中，运行结束后会基于Yarn的配置确定是否汇聚到HDFS目录中，详情请参见**Yarn常用参数**。

- 其他日志：“/var/log/Bigdata/spark2x”

日志归档规则：

- 使用yarn-client或yarn-cluster模式提交任务时，Executor日志默认50MB滚动存储一次，最多保留10个文件，不压缩。
- JobHistory2x日志默认100MB滚动存储一次，最多保留100个文件，压缩存储。
- JDBCServer2x日志默认100MB滚动存储一次，最多保留100个文件，压缩存储。
- IndexServer2x日志默认100MB滚动存储一次，最多保留100个文件，压缩存储。
- JDBCServer2x审计日志默认20MB滚动存储一次，最多保留20个文件，压缩存储。
- 日志大小和压缩文件保留个数可以在FusionInsight Manager界面中配置。

表 12-406 Spark2x 日志列表

日志类型	日志文件名	描述
SparkResource2x日志	spark.log	Spark2x服务初始化日志。
	prestart.log	prestart脚本日志。
	cleanup.log	安装卸载实例时的清理日志。

日志类型	日志文件名	描述
	spark-availability-check.log	Spark2x服务健康检查日志。
	spark-service-check.log	Spark2x服务检查日志
JDBCServer2x日志	JDBCServer-start.log	JDBCServer2x启动日志。
	JDBCServer-stop.log	JDBCServer2x停止日志。
	JDBCServer.log	JDBCServer2x运行时，Driver端日志。
	jdbc-state-check.log	JDBCServer2x健康检查日志。
	jdbcserver-omm-pid***-gc.log.*.current	JDBCServer2x进程gc日志。
	spark-omm-org.apache.spark.sql.hive.thriftserver.HiveThriftProxyServer2-***.out*	JDBCServer2x进程启动信息日志。若进程停止，会打印jstack信息。
JobHistory2x日志	jobHistory-start.log	JobHistory2x启动日志。
	jobHistory-stop.log	JobHistory2x停止日志。
	JobHistory.log	JobHistory2x运行过程日志。
	jobhistory-omm-pid***-gc.log.*.current	JobHistory2x进程gc日志。
	spark-omm-org.apache.spark.deploy.history.HistoryServer-***.out*	JobHistory2x进程启动信息日志。若进程停止，会打印jstack信息。
IndexServer2x日志	IndexServer-start.log	IndexServer2x启动日志。
	IndexServer-stop.log	IndexServer2x停止日志。
	IndexServer.log	IndexServer2x运行时，Driver端日志。
	indexserver-state-check.log	IndexServer2x健康检查日志。
	indexserver-omm-pid***-gc.log.*.current	IndexServer2x进程gc日志。
	spark-omm-org.apache.spark.sql.hive.thriftserver.IndexServerProxy-***.out*	IndexServer2x进程启动信息日志。若进程停止，会打印jstack信息。
审计日志	jdbcserver-audit.log ranger-audit.log	JDBCServer2x审计日志。

日志级别

Spark2x中提供了如表12-407所示的日志级别。日志级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-407 日志级别

级别	描述
ERROR	ERROR表示当前时间处理存在错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

说明

默认情况下配置Spark2x日志级别不需要重启服务。

步骤1 登录FusionInsight Manager系统。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”。

步骤3 单击“全部配置”。

步骤4 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤5 选择所需修改的日志级别。

步骤6 单击“保存”，然后单击“确定”，成功后配置生效。

----结束

日志格式

表 12-408 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2014-09-22 11:16:23,980 INFO DAGScheduler: Final stage: Stage 0(reduce at SparkPi.scala:35)

12.23.4 获取运行中 Spark 应用的 Container 日志

运行中Spark应用的Container日志分散在多个节点中，本章节用于说明如何快速获取Container日志。

场景说明

可以通过yarn logs命令获取运行在Yarn上的应用的日志，针对不同的场景，可以使用以下命令获取需要的日志：

1. 获取application的完整日志：**yarn logs --applicationId** *<appId>* **-out** *<outputDir>*

   例如：**yarn logs --applicationId application_1574856994802_0016 -out /opt/test**

   执行结果：

   a. 若该application处于运行状态，则无法获取dead状态的container日志

   b. 若该application处于结束状态，则可以获取全部归档的container日志

2. 获取指定Container日志：**yarn logs -applicationId** *<appId>* **-containerId** *<containerId>*

   例如：**yarn logs -applicationId application_1574856994802_0018 -containerId container_e01_1574856994802_0018_01_000003**

   执行结果：

   a. 若该application处于运行状态，则无法获取dead状态的Container日志

   b. 若该application处于结束状态，则可获取任意Container的日志

3. 获取任意状态的Container日志：**yarn logs -applicationId** *<appId>* **-containerId** *<containerId>* **-nodeAddress** *<nodeAddress>*

   例如：**yarn logs -applicationId application_1574856994802_0019 -containerId container_e01_1574856994802_0019_01_000003 -nodeAddress 192-168-1-1:**8041

   执行结果：可获取任意Container的日志

   说明

   此命令的参数中需要填入nodeAddress，可通过以下命令获取：

   yarn node -list -all

12.23.5 小文件合并工具

工具介绍

在Hadoop大规模生产集群中，由于HDFS的元数据都保存在NameNode的内存中，集群规模受制于NameNode单点的内存限制。如果HDFS中有大量的小文件，会消耗NameNode大量内存，还会大幅降低读写性能，延长作业运行时间。因此，小文件问题是制约Hadoop集群规模扩展的关键问题。

本工具主要有如下两个功能：

1. 扫描表中有多少低于用户设定阈值的小文件，返回该表目录中所有数据文件的平均大小。

2. 对表文件提供合并功能，用户可设置合并后的平均文件大小。

支持的表类型

Spark：Parquet、ORC、CSV、Text、Json。

Hive：Parquet、ORC、CSV、Text、RCFile、Sequence、Bucket。

说明

1. 数据有压缩的表在执行合并后会采用Spark默认的压缩格式-Snappy。可以通过在客户端设置“spark.sql.parquet.compression.codec”（可选：uncompressed, gzip, lzo, snappy）和"spark.sql.orc.compression.codec"（可选：uncompressed, zlib, lzo, snappy）来选择Parquet和Orc表的压缩格式；由于Hive和Spark表在可选的压缩格式上有区别，除以上列出的压缩格式外，其他的压缩格式不支持。
2. 合并桶表数据，需要先在Spark2x客户端的hive-site.xml里加上配置：

```
<property>
<name>hive.enforce.bucketing</name>
<value>false</value>
</property>
<property>
<name>hive.enforce.sorting</name>
<value>false</value>
</property>
```

3. Spark暂不支持Hive的加密列特性。

工具使用

下载安装客户端，例如安装目录为“/opt/client”。进入“/opt/client/Spark2x/spark/bin”，执行mergetool.sh脚本。

加载环境变量

source /opt/client/bigdata_env

source /opt/client/Spark2x/component_env

扫描功能

命令形式： **sh mergetool.sh scan <db.table> <filesize>**

db.table的形式是“数据库名.表名”，filesize为用户自定义的小文件阈值（单位MB），返回结果为小于该阈值的文件个数，及整个表目录数据文件的平均大小。

例如：**sh mergetool.sh scan default.table1 128**

合并功能

命令形式： **sh mergetool.sh merge <db.table> <filesize> <shuffle>**

db.table的形式是“数据库名.表名”，filesize为用户自定义的合并后平均文件大小（单位MB），shuffle是一个boolean值，取值true/false，作用是设置合并过程中是否允许数据进行shuffle。

例如：**sh mergetool.sh merge default.table1 128 false**

提示如下，则操作成功：

```
SUCCESS: Merge succeeded
```

说明

1. 请确保当前用户对合并的表具有owner权限。
2. 合并前请确保HDFS上有足够的存储空间，至少需要被合并表大小的一倍以上。
3. 合并表数据的操作需要单独进行，在此过程中读表，可能临时出现找不到文件的问题，合并完成后会恢复正常；另外在合并过程中请注意不要对相应的表进行写操作，否则可能会产生数据一致性问题。
4. 若合并完成后，在一直处于连接状态的spark-beeline/spark-sql session中查询分区表的数据，出现文件不存在的问题，根据提示可以执行"refresh table 表名"后再重新查询。
5. 请依据实际情况合理设置filesize值，例如可以在scan得到表中平均文件大小值average后，在merge时将filesize设置一个比average更大的值；否则，执行合并后可能出现文件数变得更多的情况。
6. 合并过程中，会将原表数据放入回收站，再填入已合并的数据。若在此过程中发生异常，根据工具提示，可将trash目录中的数据通过hdfs的mv命令恢复。
7. 在HDFS router联邦场景下，如果表的根路径与根路径“/user”的目标NameService不同，在二次合并时需要手动清理放入回收站的原表文件，否则会导致合并失败。
8. 此工具应用客户端配置，需要做性能调优可修改客户端配置文件的相关配置。

shuffle设置

对于合并功能，可粗略估计合并前后分区数的变化：

一般来说，旧分区数>新分区数，可设置shuffle为false；但如果旧分区远大于新分区数，例如高于100倍以上，可以考虑设置shuffle为true，增加并行度，提高合并的速度。

须知

- 设置shuffle为true（repartition），会有性能上的提升；但是由于Parquet和Orc存储方式的特殊性，repartition会使压缩率变小，直接表现是hdfs上表的总大小会增大到1.3倍。
- 设置shuffle为false（coalesce），合并后的大小不会非常平均，可能会分布在设置的filesize左右。

日志存放位置

默认日志存放位置为/tmp/SmallFilesLog.log4j，如需自定义日志存放位置，可在/opt/client/Spark2x/spark/tool/log4j.properties中配置log4j.appender.logfile.File。

12.23.6 CarbonData 首查优化工具

工具介绍

CarbonData 的首次查询较慢，对于实时性要求较高的节点可能会造成一定的时延。

本工具主要提供以下功能：

- 对查询时延要求较高的表进行首次查询预热。

工具使用

下载安装客户端，例如安装目录为“/opt/client”。进入 目录“/opt/client/Spark2x/spark/bin”，执行**start-prequery.sh**。

参考表12-409，配置prequeryParams.properties。

表 12-409 参数列表

参数	说明	示例
spark.prequery.period.max.minute	预热的最大时长，单位分钟	60
spark.prequery.tables	表名配置database.table:int，表名支持通配符*，int代表预热多长时间内有更新的表，单位为天。	default.test*:10
spark.prequery.maxThreads	预热时并发的最大线程数	50
spark.prequery.sslEnable	集群安全模式为true，非安全模式为false	true
spark.prequery.driver	JDBCServer的地址ip:port，如需要预热多个Server则需填写多个Server的IP,多个IP:port用逗号隔开。	192.168.0.2:22550
spark.prequery.sql	预热的sql语句，不同语句冒号隔开	SELECT COUNT(*) FROM %s;SELECT * FROM %s LIMIT 1
spark.security.url	安全模式下jdbc所需url	;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.hadoop.com@HADOOP.COM;

说明

spark.prequery.sql 配置的语句在每个所预热的表中都会执行，表名用%s代替。

脚本使用

命令形式：**sh start-prequery.sh**

执行此条命令需要：将user.keytab或jaas.conf（二选一），krb5.conf（必须）放入conf目录中。

说明

- 此工具暂时只支持Carbon表。
- 此工具会初始化Carbon环境和预读取表的元数据到JDBCServer，所以更适合在多主实例、静态分配模式下使用。

12.23.7 Spark2x 性能调优

12.23.7.1 Spark Core 调优

12.23.7.1.1 数据序列化

操作场景

Spark支持两种方式的序列化：

- Java原生序列化JavaSerializer
- Kryo序列化KryoSerializer

序列化对于Spark应用的性能来说，具有很大的影响。在特定的数据格式的情况下，KryoSerializer的性能可以达到JavaSerializer的10倍以上，而对于一些Int之类的基本类型数据，性能的提升就几乎可以忽略。

KryoSerializer依赖Twitter的Chill库来实现，相对于JavaSerializer，主要的问题在于不是所有的Java Serializable对象都能支持，兼容性不好，所以需要手动注册类。

序列化功能用在两个地方：序列化任务和序列化数据。Spark任务序列化只支持JavaSerializer，数据序列化支持JavaSerializer和KryoSerializer。

操作步骤

Spark程序运行时，在shuffle和RDD Cache等过程中，会有大量的数据需要序列化，默认使用JavaSerializer，通过配置让KryoSerializer作为数据序列化器来提升序列化性能。

在开发应用程序时，添加如下代码来使用KryoSerializer作为数据序列化器。

- 实现类注册器并手动注册类。

```
package com.etl.common;

import com.esotericsoftware.kryo.Kryo;
import org.apache.spark.serializer.KryoRegistrator;

public class DemoRegistrator implements KryoRegistrator
{
   @Override
   public void registerClasses(Kryo kryo)
   {
      //以下为示例类，请注册自定义的类
      kryo.register(AggrateKey.class);
      kryo.register(AggrateValue.class);
   }
}
```

  您可以在Spark客户端对spark.kryo.registrationRequired参数进行配置，设置是否需要Kryo注册序列化。

  当参数设置为true时，如果工程中存在未被序列化的类，则会抛出异常。如果设置为false（默认值），Kryo会自动将未注册的类名写到对应的对象中。此操作会对系统性能造成影响。设置为true时，用户需手动注册类，针对未序列化的类，系统不会自动写入类名，而是抛出异常，相对比false，其性能较好。

- 配置KryoSerializer作为数据序列化器和类注册器。

```
val conf = new SparkConf()
conf.set("spark.serializer", "org.apache.spark.serializer.KryoSerializer")
.set("spark.kryo.registrator", "com.etl.common.DemoRegistrator")
```

12.23.7.1.2 配置内存

操作场景

Spark是内存计算框架，计算过程中内存不够对Spark的执行效率影响很大。可以通过监控GC（Garbage Collection），评估内存中RDD的大小来判断内存是否变成性能瓶颈，并根据情况优化。

监控节点进程的GC情况（在客户端的conf/spark-default.conf配置文件中，在spark.driver.extraJavaOptions和spark.executor.extraJavaOptions配置项中添加参数："-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCTimeStamps"

），如果频繁出现Full GC，需要优化GC。把RDD做Cache操作，通过日志查看RDD在内存中的大小，如果数据太大，需要改变RDD的存储级别来优化。

操作步骤

- 优化GC，调整老年代和新生代的大小和比例。在客户端的conf/spark-default.conf配置文件中，在spark.driver.extraJavaOptions和spark.executor.extraJavaOptions配置项中添加参数：-XX:NewRatio。如，" -XX:NewRatio=2"，则新生代占整个堆空间的1/3，老年代占2/3。
- 开发Spark应用程序时，优化RDD的数据结构。
  - 使用原始类型数组替代集合类，如可使用fastutil库。
  - 避免嵌套结构。
  - Key尽量不要使用String。
- 开发Spark应用程序时，建议序列化RDD。

  RDD做cache时默认是不序列化数据的，可以通过设置存储级别来序列化RDD减小内存。例如：

  ```
  testRDD.persist(StorageLevel.MEMORY_ONLY_SER)
  ```

12.23.7.1.3 设置并行度

操作场景

并行度控制任务的数量，影响shuffle操作后数据被切分成的块数。调整并行度让任务的数量和每个任务处理的数据与机器的处理能力达到合适。

查看CPU使用情况和内存占用情况，当任务和数据不是平均分布在各节点，而是集中在个别节点时，可以增大并行度使任务和数据更均匀的分布在各个节点。增加任务的并行度，充分利用集群机器的计算能力，一般并行度设置为集群CPU总和的2-3倍。

操作步骤

并行度可以通过如下三种方式来设置，用户可以根据实际的内存、CPU、数据以及应用程序逻辑的情况调整并行度参数。

- 在会产生shuffle的操作函数内设置并行度参数，优先级最高。
  ```
  testRDD.groupByKey(24)
  ```
- 在代码中配置“spark.default.parallelism”设置并行度，优先级次之。

```
val conf = new SparkConf()
conf.set("spark.default.parallelism", 24)
```

- 在“$SPARK_HOME/conf/spark-defaults.conf”文件中配置“spark.default.parallelism”的值，优先级最低。

```
spark.default.parallelism    24
```

12.23.7.1.4 使用广播变量

操作场景

Broadcast（广播）可以把数据集合分发到每一个节点上，Spark任务在执行过程中要使用这个数据集合时，就会在本地查找Broadcast过来的数据集合。如果不使用Broadcast，每次任务需要数据集合时，都会把数据序列化到任务里面，不但耗时，还使任务变得很大。

1. 每个任务分片在执行中都需要同一份数据集合时，就可以把公共数据集Broadcast到每个节点，让每个节点在本地都保存一份。
2. 大表和小表做join操作时可以把小表Broadcast到各个节点，从而就可以把join操作转变成普通的操作，减少了shuffle操作。

操作步骤

在开发应用程序时，添加如下代码，将“testArr”数据广播到各个节点。

```
def main(args: Array[String]) {
  ...
  val testArr: Array[Long] = new Array[Long](200)
  val testBroadcast: Broadcast[Array[Long]] = sc.broadcast(testArr)
  val resultRdd: RDD[Long] = inpputRdd.map(input => handleData(testBroadcast, input))
  ...
}

def handleData(broadcast: Broadcast[Array[Long]], input: String) {
  val value = broadcast.value
  ...
}
```

12.23.7.1.5 使用 External Shuffle Service 提升性能

操作场景

Spark系统在运行含shuffle过程的应用时，Executor进程除了运行task，还要负责写shuffle数据以及给其他Executor提供shuffle数据。当Executor进程任务过重，导致触发GC（Garbage Collection）而不能为其他Executor提供shuffle数据时，会影响任务运行。

External shuffle Service是长期存在于NodeManager进程中的一个辅助服务。通过该服务来抓取shuffle数据，减少了Executor的压力，在Executor GC的时候也不会影响其他Executor的任务运行。

操作步骤

步骤1 登录FusionInsight Manager系统。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”。单击“全部配置”。

步骤3 选择“SparkResource2x > 默认”，修改以下参数：

表 12-410 参数列表

参数	默认值	修改结果
spark.shuffle.service.enabled	false	true

步骤4 重启Spark2x服务，配置生效。

说明

如果需要在Spark2x客户端用External Shuffle Service功能，需要重新下载并安装Spark2x客户端。

----结束

12.23.7.1.6 Yarn 模式下动态资源调度

操作场景

对于Spark应用来说，资源是影响Spark应用执行效率的一个重要因素。当一个长期运行的服务（比如JDBCServer），若分配给它多个Executor，可是却没有任何任务分配给它，而此时有其他的应用却资源紧张，这就造成了很大的资源浪费和资源不合理的调度。

动态资源调度就是为了解决这种场景，根据当前应用任务的负载情况，实时的增减Executor个数，从而实现动态分配资源，使整个Spark系统更加健康。

操作步骤

步骤1 需要先配置External shuffle service。

步骤2 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置 > 全部配置”。在搜索框中输入“spark.dynamicAllocation.enabled”参数名称，将参数的值设置为“true”，表示开启动态资源调度功能。默认情况下关闭此功能。

----结束

下面是一些可选配置，如表12-411所示。

表 12-411 动态资源调度参数

配置项	说明	默认值
spark.dynamicAllocation.minExecutors	最小Executor个数。	0
spark.dynamicAllocation.initialExecutors	初始Executor个数。	0
spark.dynamicAllocation.maxExecutors	最大Executor个数。	2048

配置项	说明	默认值
spark.dynamicAllocation.schedulerBacklogTimeout	调度第一次超时时间。	1s
spark.dynamicAllocation.sustainedSchedulerBacklogTimeout	调度第二次及之后超时时间。	1s
spark.dynamicAllocation.executorIdleTimeout	普通Executor空闲超时时间。	60s
spark.dynamicAllocation.cachedExecutorIdleTimeout	含有cached blocks的Executor空闲超时时间。	• JDBCServer2x：2147483647s • IndexServer2x：2147483647s • SparkResource2x：120

说明

使用动态资源调度功能，必须配置External Shuffle Service。

12.23.7.1.7 配置进程参数

操作场景

Spark on Yarn模式下，有Driver、ApplicationMaster、Executor三种进程。在任务调度和运行的过程中，Driver和Executor承担了很大的责任，而ApplicationMaster主要负责container的启停。

因而Driver和Executor的参数配置对Spark应用的执行有着很大的影响意义。用户可通过如下操作对Spark集群性能做优化。

操作步骤

步骤1 配置Driver内存。

Driver负责任务的调度，和Executor、AM之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。

您可以根据实际任务数量的多少，为Driver设置一个合适的内存。

- 将“spark-defaults.conf”中的“spark.driver.memory”配置项设置为合适大小。
- 在使用spark-submit命令时，添加“--driver-memory MEM”参数设置内存。

步骤2 配置Executor个数。

每个Executor每个核同时能跑一个task，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。

- 将“spark-defaults.conf”中的“spark.executor.instance”配置项或者“spark-env.sh”中的“SPARK_EXECUTOR_INSTANCES”配置项设置为合适大小。
- 在使用spark-submit命令时，添加“--num-executors NUM”参数设置Executor个数。

步骤3 配置Executor核数。

每个Executor多个核同时能跑多个task，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。

- 将“spark-defaults.conf”中的“spark.executor.cores”配置项或者“spark-env.sh”中的“SPARK_EXECUTOR_CORES”配置项设置为合适大小。
- 在使用spark-submit命令时，添加“--executor-cores NUM”参数设置核数。

步骤4 配置Executor内存。

Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。

- 将“spark-defaults.conf”中的“spark.executor.memory”配置项或者“spark-env.sh”中的“SPARK_EXECUTOR_MEMORY”配置项设置为合适大小。
- 在使用spark-submit命令时，添加“--executor-memory MEM”参数设置内存。

----结束

示例

- 在执行spark wordcount计算中。1.6T数据，250个executor。

  在默认参数下执行失败，出现Futures timed out和OOM错误。

  因为数据量大，task数多，而wordcount每个task都比较小，完成速度快。当task数多时driver端相应的一些对象就变大了，而且每个task完成时executor和driver都要通信，这就会导致由于内存不足，进程之间通信断连等问题。

  当把Driver的内存设置到4g时，应用成功跑完。
- 使用JDBCServer执行TPC-DS测试套，默认参数配置下也报了很多错误：Executor Lost等。而当配置Driver内存为30g，executor核数为2，executor个数为125，executor内存为6g时，所有任务才执行成功。

12.23.7.1.8 设计 DAG

操作场景

合理的设计程序结构，可以优化执行效率。在程序编写过程中要尽量减少shuffle操作，合并窄依赖操作。

操作步骤

以“同行车判断”例子讲解DAG设计的思路。

- **数据格式：**通过收费站时间、车牌号、收费站编号.....
- **逻辑：**以下两种情况下判定这两辆车是同行车：
  - 如果两辆车都通过相同序列的收费站，

- 通过同一收费站之间的时间差小于一个特定的值。

该例子有两种实现模式，其中实现1的逻辑如图12-55所示，实现2的逻辑如图12-56所示。

图 12-55 实现 1 逻辑

实现1的逻辑说明：

1. 根据车牌号聚合该车通过的所有收费站并排序，处理后数据如下：

   车牌号1，[（通过时间，收费站3），（通过时间，收费站2），（通过时间，收费站4），（通过时间，收费站5）]
2. 标识该收费站是这辆车通过的第几个收费站。

   （收费站3，（车牌号1，通过时间，通过的第1个收费站））

   （收费站2，（车牌号1，通过时间，通过的第2个收费站））

   （收费站4，（车牌号1，通过时间，通过的第3个收费站））

   （收费站5，（车牌号1，通过时间，通过的第4个收费站））
3. 根据收费站聚合数据。

   收费站1，[（车牌号1，通过时间，通过的第1个收费站），（车牌号2，通过时间，通过的第5个收费站），（车牌号3，通过时间，通过的第2个收费站）]
4. 判断两辆车通过该收费站的时间差是否满足同行车的要求，如果满足则取出这两辆车。

   （车牌号1，车牌号2），（通过的第1个收费站，通过的第5个收费站）

   （车牌号1，车牌号3），（通过的第1个收费站，通过的第2个收费站）
5. 根据通过相同收费站的两辆车的车牌号聚合数据，如下：

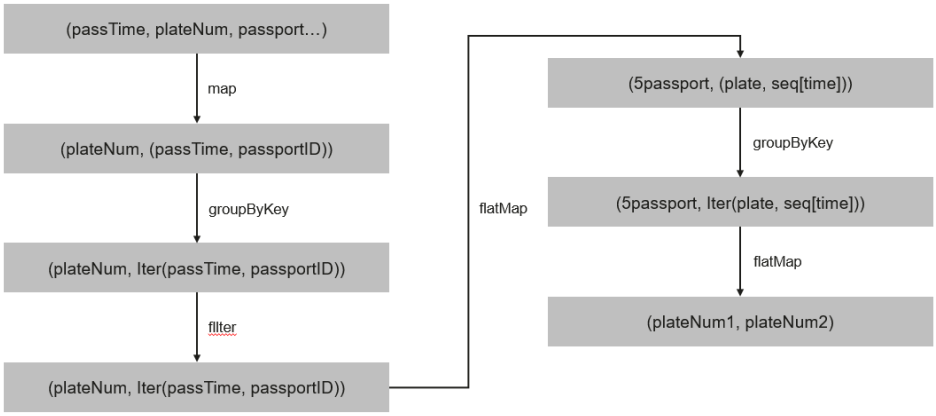
   （车牌号1，车牌号2），[（通过的第1个收费站，通过的第5个收费站），（通过的第2个收费站，通过的第6个收费站），（通过的第1个收费站，通过的第7个收费站），（通过的第3个收费站，通过的第8个收费站）]
6. 如果车牌号1和车牌号2通过相同收费站是顺序排列的（比如收费站3、4、5是车牌1通过的第1、2、3个收费站，是车牌2通过的第6、7、8个收费站）且数量大于同行车要求的数量则这两辆车是同行车。

实现1逻辑的缺点：

- 逻辑复杂

- 实现过程中shuffle操作过多，对性能影响较大。

图 12-56 实现 2 逻辑



实现2的逻辑说明：

1. 根据车牌号聚合该车通过的所有收费站并排序，处理后数据如下：

   车牌号1，[（通过时间，收费站3），（通过时间，收费站2），（通过时间，收费站4），（通过时间，收费站5）]
2. 根据同行车要通过的收费站数量（例子里为3）分段该车通过的收费站序列，如上面的数据被分解成：

   收费站3->收费站2->收费站4，（车牌号1，[收费站3时间，收费站2时间，收费站4时间]）

   收费站2->收费站4->收费站5，（车牌号1，[收费站2时间，收费站4时间，收费站5时间]）
3. 把通过相同收费站序列的车辆聚合，如下：

   收费站3->收费站2->收费站4，[（车牌号1，[收费站3时间，收费站2时间，收费站4时间]），（车牌号2，[收费站3时间，收费站2时间，收费站4时间]），（车牌号3，[收费站3时间，收费站2时间，收费站4时间]）]
4. 判断通过相同序列收费站的车辆通过相同收费站的时间差是不是满足同行车的要求，如果满足则说明是同行车。

实现2的优点如下：

- 简化了实现逻辑。
- 减少了一个groupByKey，也就减少了一次shuffle操作，提升了性能。

12.23.7.1.9 经验总结

使用 mapPartitions，按每个分区计算结果

如果每条记录的开销太大，例：

```
rdd.map{x=>conn=getDBConn;conn.write(x.toString);conn.close}
```

则可以使用MapPartitions，按每个分区计算结果，如

```
rdd.mapPartitions(records => conn.getDBConn;for(item <- records)
write(item.toString); conn.close)
```

使用mapPartitions可以更灵活地操作数据，例如对一个很大的数据求TopN，当N不是很大时，可以先使用mapPartitions对每个partition求TopN，collect结果到本地之后再做排序取TopN。这样相比直接对全量数据做排序取TopN效率要高很多。

使用 coalesce 调整分片的数量

coalesce可以调整分片的数量。coalesce函数有两个参数：

```
coalesce(numPartitions: Int, shuffle: Boolean = false)
```

当shuffle为true的时候，函数作用与repartition(numPartitions: Int)相同，会将数据通过Shuffle的方式重新分区；当shuffle为false的时候，则只是简单的将父RDD的多个partition合并到同一个task进行计算，shuffle为false时，如果numPartitions大于父RDD的切片数，那么分区不会重新调整。

遇到下列场景，可选择使用coalesce算子：

- 当之前的操作有很多filter时，使用coalesce减少空运行的任务数量。此时使用coalesce(numPartitions, false)，numPartitions小于父RDD切片数。
- 当输入切片个数太大，导致程序无法正常运行时使用。
- 当任务数过大时候Shuffle压力太大导致程序挂住不动，或者出现linux资源受限的问题。此时需要对数据重新进行分区，使用coalesce(numPartitions, true)。

localDir 配置

Spark的Shuffle过程需要写本地磁盘，Shuffle是Spark性能的瓶颈，I/O是Shuffle的瓶颈。配置多个磁盘则可以并行的把数据写入磁盘。如果节点中挂载多个磁盘，则在每个磁盘配置一个Spark的localDir，这将有效分散Shuffle文件的存放，提高磁盘I/O的效率。如果只有一个磁盘，配置了多个目录，性能提升效果不明显。

Collect 小数据

大数据量不适用collect操作。

collect操作会将Executor的数据发送到Driver端，因此使用collect前需要确保Driver端内存足够，以免Driver进程发生OutOfMemory异常。当不确定数据量大小时，可使用saveAsTextFile等操作把数据写入HDFS中。只有在能够大致确定数据大小且driver内存充足的时候，才能使用collect。

使用 reduceByKey

reduceByKey会在Map端做本地聚合，使得Shuffle过程更加平缓，而groupByKey等Shuffle操作不会在Map端做聚合。因此能使用reduceByKey的地方尽量使用该算子，避免出现groupByKey().map(x=>(x._1,x._2.size))这类实现方式。

广播 map 代替数组

当每条记录需要查表，如果是Driver端用广播方式传递的数据，数据结构优先采用set/map而不是Iterator，因为Set/Map的查询速率接近O(1)，而Iterator是O(n)。

数据倾斜

当数据发生倾斜（某一部分数据量特别大），虽然没有GC（Gabage Collection，垃圾回收），但是task执行时间严重不一致。

- 需要重新设计key，以更小粒度的key使得task大小合理化。
- 修改并行度。

优化数据结构

- 把数据按列存放，读取数据时就可以只扫描需要的列。
- 使用Hash Shuffle时，通过设置spark.shuffle.consolidateFiles为true，来合并shuffle中间文件，减少shuffle文件的数量，减少文件IO操作以提升性能。最终文件数为reduce tasks数目。

12.23.7.2 SQL 和 DataFrame 调优

12.23.7.2.1 Spark SQL join 优化

操作场景

Spark SQL中，当对两个表进行join操作时，利用Broadcast特性（见“使用广播变量”章节），将被广播的表BroadCast到各个节点上，从而转变成非shuffle操作，提高任务执行性能。

说明

这里join操作，只指inner join。

操作步骤

在Spark SQL中进行Join操作时，可以按照以下步骤进行优化。为了方便说明，设表A和表B，且A、B表都有个名为name的列。对A、B表进行join操作。

1. 估计表的大小。

   根据每次加载数据的大小，来估计表大小。

   也可以在Hive的数据库存储路径下直接查看表的大小。首先在Spark的配置文件“hive-site.xml”中，查看Hive的数据库路径的配置，默认为“/user/hive/warehouse”。Spark服务多实例默认数据库路径为“/user/*hive*/warehouse”，例如“/user/hive1/warehouse”。

   ```
   <property>
     <name>hive.metastore.warehouse.dir</name>
    <value>${test.warehouse.dir}</value>
    <description></description>
   </property>
   ```

   然后通过hadoop命令查看对应表的大小。如查看表A的大小命令为：

   ```
   hadoop fs -du -s -h ${test.warehouse.dir}/a
   ```

   说明

   进行广播操作，需要至少有一个表不是空表。

2. 配置自动广播的阈值。

   Spark中，判断表是否广播的阈值为10485760（即10M）。如果两个表的大小至少有一个小于10M时，可以跳过该步骤。

   自动广播阈值的配置参数介绍，见**表12-412**。

表 12-412 参数介绍

参数	默认值	描述
spark.sql.autoBroadcastJoinThreshold	10485760	当进行join操作时，配置广播的最大值。 ● 当SQL语句中涉及的表中相应字段的大小小于该值时，进行广播。 ● 配置为-1时，将不进行广播。 参见**https://archive.apache.org/dist/spark/docs/3.1.1/sql-programming-guide.html**

配置自动广播阈值的方法：

- 在Spark的配置文件“spark-defaults.conf”中，设置“spark.sql.autoBroadcastJoinThreshold”的值。
  ```
  spark.sql.autoBroadcastJoinThreshold = <size>
  ```
- 利用Hive CLI命令，设置阈值。在运行Join操作时，提前运行下面语句：
  ```
  SET spark.sql.autoBroadcastJoinThreshold=<size>;
  ```

3. 进行join操作。
   - 两个表的大小都小于阈值。
     - A表的字节数小于B表，则运行B join A，如
       ```
       SELECT A.name FROM B JOIN A ON A.name = B.name;
       ```
     - 否则运行A join B。
       ```
       SELECT A.name FROM A JOIN B ON A.name = B.name;
       ```
   - 一个表大于阈值一个表小于阈值。

     将小表进行BroadCast操作。
   - 两个表的大小都大于阈值。

     比较查询所涉及的字段大小与阈值的大小。
     - 若某表中涉及字段的大小小于阈值，将该表相应数据进行广播。
     - 若两表中涉及字段的大小都大于阈值，则不进行广播。
4. （可选）如下两种场景，需要执行Analyze命令（***ANALYZE TABLE tableName COMPUTE STATISTICS noscan;***）更新表元数据后进行广播。
   - 需要广播的表是分区表，新建表且文件类型为非Parquet文件类型。
   - 需要广播的表是分区表，更新表数据后。

参考信息

被广播的表执行超时，导致任务结束。

默认情况下，BroadCastJoin只允许被广播的表计算5分钟，超过5分钟该任务会出现超时异常，而这个时候被广播的表的broadcast任务依然在执行，造成资源浪费。

这种情况下，有两种方式处理：

- 调整“spark.sql.broadcastTimeout”的数值，加大超时的时间限制。
- 降低“spark.sql.autoBroadcastJoinThreshold”的数值，不使用BroadCastJoin的优化。

12.23.7.2.2 优化数据倾斜场景下的 Spark SQL 性能

配置场景

在Spark SQL多表Join的场景下，会存在关联键严重倾斜的情况，导致Hash分桶后，部分桶中的数据远高于其它分桶。最终导致部分Task过重，跑得很慢；其它Task过轻，跑得很快。一方面，数据量大Task运行慢，使得计算性能低；另一方面，数据量少的Task在运行完成后，导致很多CPU空闲，造成CPU资源浪费。

通过如下配置项可开启自动进行数据倾斜处理功能，通过将Hash分桶后数据量很大的、且超过数据倾斜阈值的分桶拆散，变成多个task处理一个桶的数据机制，提高CPU资源利用率，提高系统性能。

说明

未产生倾斜的数据，将采用原有方式进行分桶并运行。

使用约束：

- 只支持两表Join的场景。
- 不支持FULL OUTER JOIN的数据倾斜处理。
  示例：执行下面SQL语句，a表倾斜或b表倾斜都无法触发该优化。
  select aid FROM a FULL OUTER JOIN b ON aid=bid;
- 不支持LEFT OUTER JOIN的右表倾斜处理。
  示例：执行下面SQL语句，b表倾斜无法触发该优化。
  select aid FROM a LEFT OUTER JOIN b ON aid=bid;
- 不支持RIGHT OUTER JOIN的左表倾斜处理。
  示例：执行下面SQL语句，a表倾斜无法触发该优化。
  select aid FROM a RIGHT OUTER JOIN b ON aid=bid;

配置描述

在Spark Driver端的“spark-defaults.conf”配置文件中添加如下表格中的参数。

表 12-413 参数说明

参数	描述	默认值
spark.sql.adaptive.enabled	自适应执行特性的总开关。 注意：AQE特性与DPP（动态分区裁剪）特性同时开启时，SparkSQL任务执行中会优先执行DPP特性，从而使得AQE特性不生效。集群中DPP特性是默认开启的，因此开启AQE特性的同时，需要将DPP特性关闭。	false

参数	描述	默认值
spark.sql.optimizer.dynamicPartitionPruning.enabled	动态分区裁剪功能的开关。	true
spark.sql.adaptive.skewJoin.enabled	当此配置为true且spark.sql.adaptive.enabled设置为true时，启用运行时自动处理join运算中的数据倾斜功能。	true
spark.sql.adaptive.skewJoin.skewedPartitionFactor	此配置为一个倍数因子，用于判定分区是否为数据倾斜分区。单个分区被判定为数据倾斜分区的条件为：当一个分区的数据大小超过除此分区外其他所有分区大小的中值与该配置的乘积，并且大小超过spark.sql.adaptive.skewJoin.skewedPartitionThresholdInBytes配置值时，此分区被判定为数据倾斜分区	5
spark.sql.adaptive.skewjoin.skewedPartitionThresholdInBytes	分区大小（单位：字节）大于该阈值且大于spark.sql.adaptive.skewJoin.skewedPartitionFactor与分区中值的乘积，则认为该分区存在倾斜。理想情况下，此配置应大于spark.sql.adaptive.advisoryPartitionSizeInBytes.	256MB
spark.sql.adaptive.shuffle.targetPostShuffleInputSize	每个task处理的shuffle数据的最小数据量。单位：Byte。	67108864

12.23.7.2.3 优化小文件场景下的 Spark SQL 性能

配置场景

Spark SQL的表中，经常会存在很多小文件（大小远小于HDFS块大小），每个小文件默认对应Spark中的一个Partition，也就是一个Task。在很多小文件场景下，Spark会起很多Task。当SQL逻辑中存在Shuffle操作时，会大大增加hash分桶数，严重影响性能。

在小文件场景下，您可以通过如下配置手动指定每个Task的数据量（Split Size），确保不会产生过多的Task，提高性能。

说明

当SQL逻辑中不包含Shuffle操作时，设置此配置项，不会有明显的性能提升。

配置描述

要启动小文件优化，在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-414 参数说明

参数	描述	默认值
spark.sql.files.maxPartitionBytes	在读取文件时，将单个分区打包的最大字节数。 单位：byte。	134217728（即128M）
spark.files.openCostInBytes	打开文件的预估成本， 按照同一时间能够扫描的字节数来测量。当一个分区写入多个文件时使用。高估更好，这样小文件分区将比大文件分区更先被调度。	4M

12.23.7.2.4 INSERT...SELECT 操作调优

操作场景

在以下几种情况下，执行INSERT...SELECT操作可以进行一定的调优操作。

- 查询的数据是大量的小文件。
- 查询的数据是较多的大文件。
- 在Beeline/JDBCServer模式下使用非Spark用户操作。

操作步骤

可对INSERT...SELECT操作做如下的调优操作。

- 如果建的是Hive表，将存储类型设为Parquet，从而减少执行INSERT...SELECT语句的时间。
- 建议使用spark-sql或者在Beeline/JDBCServer模式下使用spark用户来执行INSERT...SELECT操作，避免执行更改文件owner的操作，从而减少执行INSERT...SELECT语句的时间。

  说明

  在Beeline/JDBCServer模式下，executor的用户跟driver是一致的，driver是JDBCServer服务的一部分，是由spark用户启动的，因此其用户也是spark用户，且当前无法实现在运行时将Beeline端的用户透传到executor，因此使用非spark用户时需要对文件进行更改owner为Beeline端的用户，即实际用户。

- 如果查询的数据是大量的小文件将会产生大量map操作，从而导致输出存在大量的小文件，在执行重命名文件操作时将会耗费较多时间，此时可以通过设置“spark.sql.files.maxPartitionBytes”与“spark.files.openCostInBytes”来设置一个partiton读取的最大字节，在一个partition中合并多个小文件来减少输出文件数及执行重命名文件操作的时间，从而减少执行INSERT...SELECT语句的时间。

说明

上述优化操作并不能解决全部的性能问题，对于以下场景仍然需要较多时间：

对于动态分区表，如果其分区数非常多，那么也需要执行较长的时间。

12.23.7.2.5 多并发 JDBC 客户端连接 JDBCServer

操作场景

JDBCServer支持多用户多并发接入，但当并发任务数量较高的时候，默认的JDBCServer配置将无法支持，因此需要进行优化来支持该场景。

操作步骤

1. 设置JDBCServer的公平调度策略。

   Spark默认使用FIFO（First In First Out）的调度策略，但对于多并发的场景，使用FIFO策略容易导致短任务执行失败。因此在多并发的场景下，需要使用公平调度策略，防止任务执行失败。

   a. 在Spark中设置公平调度，具体请参考**http://archive.apache.org/dist/spark/docs/3.1.1/job-scheduling.html#scheduling-within-an-application**

   b. 在JDBC客户端中设置公平调度。

      i. 在BeeLine命令行客户端或者JDBC自定义代码中，执行以下语句，其中PoolName是公平调度的某一个调度池。

      ```
      SET spark.sql.thriftserver.scheduler.pool=PoolName;
      ```

      ii. 执行相应的SQL命令，Spark任务将会在上面的调度池中运行。

2. 设置BroadCastHashJoin的超时时间。

   BroadCastHashJoin有超时参数，一旦超过预设的时间，该查询任务直接失败，在多并发场景下，由于计算任务抢占资源，可能会导致BroadCastHashJoin的Spark任务无法执行，导致超时出现。因此需要在JDBCServer的“spark-defaults.conf”配置文件中调整超时时间。

   表 12-415 参数描述

参数	描述	默认值
spark.sql.broadcastTimeout	BroadcastHashJoin中广播表的超时时间，当任务并发数较高的时候，可以调高该参数值。	-1（数值类型，实际为五分钟）

12.23.7.2.6 动态分区插入场景内存优化

操作场景

SparkSQL在往动态分区表中插入数据时，分区数越多，单个Task生成的HDFS文件越多，则元数据占用的内存也越多。这就导致程序GC（Gabage Collection）严重，甚至发生OOM（Out of Memory）。

经测试证明：10240个Task，2000个分区，在执行HDFS文件从临时目录rename到目标目录动作前，FileStatus元数据大小约29G。为避免以上问题，可修改SQL语句对数据进行重分区，以减少HDFS文件个数。

操作步骤

在动态分区语句中加入**distribute by**，by值为分区字段。

示例如下：

```
insert into table store_returns partition (sr_returned_date_sk) select
sr_return_time_sk,sr_item_sk,sr_customer_sk,sr_cdemo_sk,sr_hdemo_sk,sr_addr_sk,
sr_store_sk,sr_reason_sk,sr_ticket_number,sr_return_quantity,sr_return_amt,sr_retur
n_tax,sr_return_amt_inc_tax,sr_fee,sr_return_ship_cost,sr_refunded_cash,sr_reversed
_charge,sr_store_credit,sr_net_loss,sr_returned_date_sk from $
{SOURCE}.store_returns distribute by sr_returned_date_sk;
```

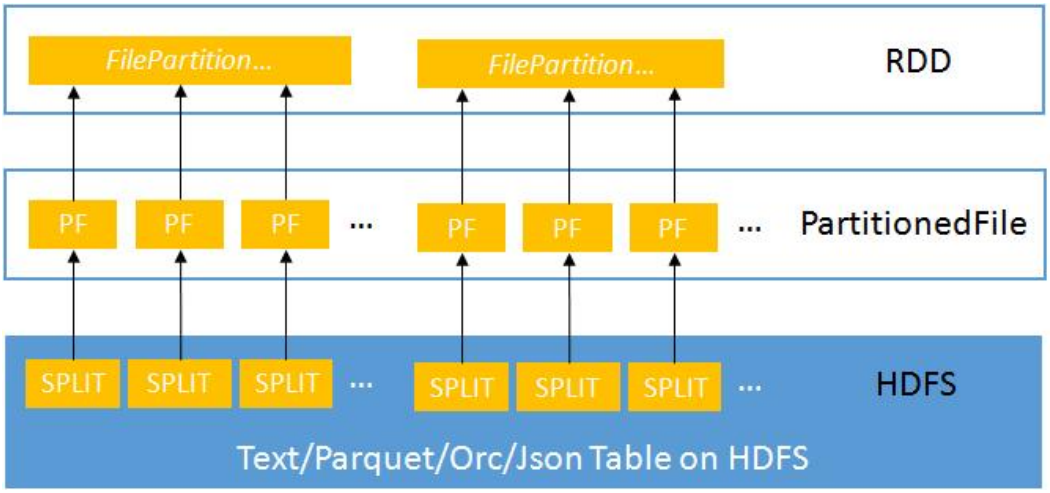
12.23.7.2.7 小文件优化

操作场景

Spark SQL表中，经常会存在很多小文件（大小远小于HDFS的块大小），每个小文件默认对应Spark中的一个Partition，即一个Task。在有很多小文件时，Spark会启动很多Task，此时当SQL逻辑中存在Shuffle操作时，会大大增加hash分桶数，严重影响系统性能。

针对小文件很多的场景，DataSource在创建RDD时，先将Table中的split生成PartitionedFile，再将这些PartitionedFile进行合并。即将多个PartitionedFile组成一个partition，从而减少partition数量，避免在Shuffle操作时生成过多的hash分桶，如图12-57所示。

图 12-57 小文件合并



操作步骤

要启动小文件优化，在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-416 参数介绍

参数	描述	默认值
spark.sql.files.maxPartitionBytes	在读取文件时，将单个分区打包的最大字节数。 单位：byte。	134217728（即128M）

参数	描述	默认值
spark.files.openCostInBytes	打开文件的预估成本， 按照同一时间能够扫描的字节数来测量。当一个分区写入多个文件时使用。高估更好，这样小文件分区将比大文件分区更先被调度。	4M

12.23.7.2.8 聚合算法优化

操作场景

在Spark SQL中支持基于行的哈希聚合算法，即使用快速聚合hashmap作为缓存，以提高聚合性能。hashmap替代了之前的ColumnarBatch支持，从而避免拥有聚合表的宽模式（大量key字段或value字段）时产生的性能问题。

操作步骤

要启动聚合算法优化，在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-417 参数介绍

参数	描述	默认值
spark.sql.codegen.aggregate.map.twolevel.enabled	是否开启聚合算法优化： ● true：开启 ● false：不开启	true

12.23.7.2.9 Datasource 表优化

操作场景

将datasource表的分区消息存储到Metastore中，并在Metastore中对分区消息进行处理。

- 优化datasource表，支持对表中分区执行增加、删除和修改等语法，从而增加与Hive的兼容性。
- 支持在查询语句中，把分区裁剪并下压到Metastore上，从而过滤掉不匹配的分区。

  示例如下：

  ```
  select count(*) from table where partCol=1;  //partCol列为分区列
  ```

  此时，在物理计划中执行TableScan操作时，只处理分区(partCol=1)对应的数据。

操作步骤

要启动Datasource表优化，在Spark客户端的“spark-defaults.conf”配置文件中进行设置。

表 12-418 参数介绍

参数	描述	默认值
spark.sql.hive.manageFilesourcePartitions	是否启用Metastore分区管理（包括数据源表和转换的Hive表）。 ● true：启用Metastore分区管理，即数据源表存储分区在Hive中，并在查询语句中使用Metastore修剪分区。 ● false：不启用Metastore分区管理。	true
spark.sql.hive.metastorePartitionPruning	是否支持将predicate下压到Hive Metastore中。 ● true：支持，目前仅支持Hive表的predicate下压。 ● false：不支持	true
spark.sql.hive.filesourcePartitionFileCacheSize	启用内存中分区文件元数据的缓存大小。 所有表共享一个可以使用指定的num字节进行文件元数据的缓存。 只有当“spark.sql.hive.manageFilesourcePartitions”配置为“true”时，该配置项才会生效。	250 * 1024 * 1024
spark.sql.hive.convertMetastoreOrc	设置ORC表的处理方式： ● false：Spark SQL使用Hive SerDe处理ORC表。 ● true：Spark SQL使用Spark内置的机制处理ORC表。	true

12.23.7.2.10 合并 CBO 优化

操作场景

Spark SQL默认支持基于规则的优化，但仅仅基于规则优化不能保证Spark选择合适的查询计划。CBO（Cost-Bsed Optimizer）是一种为SQL智能选择查询计划的技术。通过配置开启CBO后，CBO优化器可以基于表和列的统计信息，进行一系列的估算，最终选择出合适的查询计划。

操作步骤

要使用CBO优化，可以按照以下步骤进行优化。

1. 需要先执行特定的SQL语句来收集所需的表和列的统计信息。
   SQL命令如下（根据具体情况选择需要执行的SQL命令）：
   - 生成表级别统计信息（扫表）：
     ANALYZE TABLE src COMPUTE STATISTICS
     生成sizeInBytes和rowCount。

使用ANALYZE语句收集统计信息时，无法计算非HDFS数据源的表的文件大小。

- 生成表级别统计信息（不扫表）：

  ANALYZE TABLE src COMPUTE STATISTICS NOSCAN

  只生成sizeInBytes，如果原来已经生成过sizeInBytes和rowCount，而本次生成的sizeInBytes和原来的大小一样，则保留rowCount（若存在），否则清除rowCount。

- 生成列级别统计信息

  ANALYZE TABLE src COMPUTE STATISTICS FOR COLUMNS a, b, c

  生成列统计信息，为保证一致性，会同步更新表统计信息。目前不支持复杂数据类型（如Seq, Map等）和HiveStringType的统计信息生成。

- 显示统计信息

  DESC FORMATTED src

  在Statistics中会显示“xxx bytes, xxx rows”分别表示表级别的统计信息。也可以通过如下命令显示列统计信息：

  DESC FORMATTED src a

使用限制：当前统计信息收集不支持针对分区表的分区级别的统计信息。

2. 在Spark客户端的“spark-defaults.conf”配置文件中进行**表12-419**设置。

表 12-419 参数介绍

参数	描述	默认值
spark.sql.cbo.enabled	CBO总开关。 ● true表示打开， ● false表示关闭。 要使用该功能，需确保相关表和列的统计信息已经生成。	false
spark.sql.cbo.joinReorder.enabled	使用CBO来自动调整连续的inner join的顺序。 ● true：表示打开 ● false：表示关闭 要使用该功能，需确保相关表和列的统计信息已经生成，且CBO总开关打开。	false
spark.sql.cbo.joinReorder.dp.threshold	使用CBO来自动调整连续inner join的表的个数阈值。 如果超出该阈值，则不会调整join顺序。	12

12.23.7.2.11 跨源复杂数据的 SQL 查询优化

操作场景

本章节介绍如何打开或关闭跨源复杂数据的SQL查询优化功能。

操作步骤

- （可选）连接MPPDB数据源的准备

  如果连接的数据源为MPPDB，由于MPPDB Driver文件“gsjdbc4.jar”和Spark中的jar包“gsjdbc4-VXXXRXXXCXXSPCXXX.jar”包含了相同的类名，存在类名冲突的问题。因此在连接MPPDB数据库之前，需要执行以下步骤：

  a. 移除Spark中的“gsjdbc4-VXXXRXXXCXXSPCXXX.jar”，由于Spark运行不依赖该jar包，因此将该jar包移动到其他目录（例如，移动到“/tmp”目录，不建议直接删除）不会影响Spark正常运行。

     i. 登录Spark服务端主机，移除“${BIGDATA_HOME}/FusionInsight_Spark2x_8.1.0.1/install/FusionInsight-Spark2x-3.1.1/spark/jars”路径下的“gsjdbc4-VXXXRXXXCXXSPCXXX.jar”。

     ii. 登录Spark客户端主机，移除“/opt/client/Spark2x/spark/jars”路径下的“gsjdbc4-VXXXRXXXCXXSPCXXX.jar”。

  b. 在MPPDB的安装包中获取MPPDB Driver文件“gsjdbc4.jar”，并将该文件分别上传到以下位置：

     - Spark服务端的“/${BIGDATA_HOME}/FusionInsight_Spark2x_8.1.0.1/install/FusionInsight-Spark2x-3.1.1/spark/jars”路径下。
     - Spark客户端的“/opt/client/Spark2x/spark/jars”路径下。

  c. 更新存储在HDFS中的“/user/spark2x/jars/8.1.0.1/spark-archive-2x.zip”压缩包。

     说明

     此处版本号8.1.0.1为示例，具体以实际环境的版本号为准。

     i. 使用客户端安装用户登录客户端所在节点。执行命令切换到客户端安装目录，例如“/opt/client”。

     cd /opt/client

     ii. 执行以下命令配置环境变量。

     source bigdata_env

     iii. 如果集群为安全模式，执行以下命令获得认证。

     kinit *组件业务用户*

     iv. 新建临时文件./tmp，并从HDFS获取“spark-archive-2x.zip”并解压到tmp目录，命令如下：

     mkdir tmp

     hdfs dfs -get /user/spark2x/jars/8.1.0.1/spark-archive-2x.zip **./**

     unzip spark-archive-2x.zip -d ./tmp

     v. 切换到tmp目录，删除“gsjdbc4-VXXXRXXXCXXSPCXXX.jar”文件，并将MPPDB Driver文件“gsjdbc4.jar”上传到tmp目录中，然后执行以下命令重新打包。

     zip -r spark-archive-2x.zip *.jar

     vi. 删除HDFS上的“spark-archive-2x.zip”，将步骤**c.v**中新生成的压缩包“spark-archive-2x.zip”更新至HDFS的“/user/spark2x/jars/8.1.0.1/”路径下。

     hdfs dfs -rm /user/spark2x/jars/8.1.0.1/spark-archive-2x.zip

     hdfs dfs -put ./spark-archive-2x.zip /user/spark2x/jars/8.1.0.1

d. 重启Spark服务，等重启成功后，重新启动Spark客户端。

- 打开优化开关

  对所有支持查询下推的模块，可以通过在spark-beeline客户端中执行SET命令打开跨源查询优化功能，默认均为关闭状态。

  可以从全局、数据源、表这三个维度进行下推开关控制。打开方法如下：

  - 全局（对所有数据源生效）：

    SET spark.sql.datasource.jdbc = project,aggregate,orderby-limit

  - 数据源：

    SET spark.sql.datasource.${url} = project,aggregate,orderby-limit

  - 表：

    SET spark.sql.datasource.${url}.${table} = project,aggregate,orderby-limit

  执行SET命令设置上述参数时，允许一次设置多个下推模块，中间以逗号分隔。各个下推模块对应的参数值如下所示：

  表 12-420 各模块对应的参数值

模块名称	SET命令的参数值
project	project
aggregate	aggregate
order by，limit over project or aggregate	orderby-limit

  示例：创建一个MySQL的外表的语句为：

  create table if not exists pdmysql using org.apache.spark.sql.jdbc options(driver "com.mysql.jdbc.Driver", url "jdbc:mysql://ip2:3306/test", user "hive", password "*xxx*", dbtable "mysqldata");

  则其中：

  - ${url} = jdbc:mysql://ip2:3306/test
  - ${table} = mysqldata

  说明

  - “=”后即设置可以下推打开的算子，以“,”隔开。
  - 优先级：table开关>数据源开关>全局开关。即若设置了table开关，则数据源开关全局开关对该表失效；若配置了数据源开关，则全局开关对该数据源失效。
  - url中不能包含“=”，若包含，set时直接删掉“=”。
  - 可多次执行set，key不同不会相互覆盖。

- 新增支持查询下推的函数

  除了支持对abs()、month()、length()等数学、时间、字符串函数进行查询下推外，用户还可以通过SET命令新增数据源支持查询下推的函数。在spark-beeline客户端中执行如下命令：

  SET spark.sql.datasource.${datasource}.functions = fun1,fun2

- 取消开关设置及取消新增的下推函数

  当前只能通过在spark-beeline客户端中执行***RESET***命令取消所有SET的内容。由于执行***RESET***后所有SET的参数值都将被清除，请谨慎使用。

  控制开关的设置仅在客户端当前的会话中生效，当客户端关闭后，SET内容就失效了。

  或者修改客户端配置文件spark-defaults.conf中的spark.sql.locale.support参数为true。

注意事项

数据源只支持MySQL和MPPDB，Hive，oracle， postgresql。

12.23.7.2.12 多级嵌套子查询以及混合 Join 的 SQL 调优

操作场景

本章节介绍在多级嵌套以及混合Join SQL查询的调优建议。

前提条件

例如有一个复杂的查询样例如下：

```
select
s_name,
count(1) as numwait
from (
select s_name from (
select
s_name,
t2.l_orderkey,
l_suppkey,
count_suppkey,
max_suppkey
from
test2 t2 right outer join (
select
s_name,
l_orderkey,
l_suppkey from (
select
s_name,
t1.l_orderkey,
l_suppkey,
count_suppkey,
max_suppkey
from
test1 t1 join (
select
s_name,
l_orderkey,
l_suppkey
from
orders o join (
select
s_name,
l_orderkey,
l_suppkey
from
nation n join supplier s
on
s.s_nationkey = n.n_nationkey
and n.n_name = 'SAUDI ARABIA'
```

```
join lineitem l
on
s.s_suppkey = l.l_suppkey
where
l.l_receiptdate > l.l_commitdate
and l.l_orderkey is not null
) l1 on o.o_orderkey = l1.l_orderkey and o.o_orderstatus = 'F'
) l2 on l2.l_orderkey = t1.l_orderkey
) a
where
(count_suppkey > 1)
or ((count_suppkey=1)
and (l_suppkey <> max_suppkey))
) l3 on l3.l_orderkey = t2.l_orderkey
) b
where
(count_suppkey is null)
or ((count_suppkey=1)
and (l_suppkey = max_suppkey))
) c
group by
s_name
order by
numwait desc,
s_name
limit 100;
```

操作步骤

步骤1 分析业务。

从业务入手分析是否可以简化SQL，例如可以通过合并表去减少嵌套的层级和Join的次数。

步骤2 如果业务需求对应的SQL无法简化，则需要配置DRIVER内存：

- 使用**spark-submit**或者**spark-sql**运行SQL语句，执行**步骤3**。
- 使用**spark-beeline**运行SQL语句，执行**步骤4**。

步骤3 执行SQL语句时，需要添加参数“--driver-memory”，设置内存大小，例如：

/spark-sql --master=local[4] --driver-memory=512M -f /tpch.sql

步骤4 在执行SQL语句前，请使用管理员用户修改内存大小配置。

1. 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”。
2. 单击“全部配置”，并搜索“SPARK_DRIVER_MEMORY”。
3. 修改参数值适当增加内存大小。仅支持整数值，且需要输入单位M或者G。例如输入512M。

----结束

参考信息

DRIVER内存不足时，查询操作可能遇到以下错误提示信息：

```
2018-02-11 09:13:14,683 | WARN  | Executor task launch worker for task 5 | Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0. |
org.apache.spark.sql.catalyst.expressions.RowBasedKeyValueBatch.spill(RowBasedKeyValueBatch.java:173)
2018-02-11 09:13:14,682 | WARN  | Executor task launch worker for task 3 | Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0. |
org.apache.spark.sql.catalyst.expressions.RowBasedKeyValueBatch.spill(RowBasedKeyValueBatch.java:173)
```

```
2018-02-11 09:13:14,704 | ERROR | Executor task launch worker for task 2 | Exception in task 2.0 in stage 1.0 (TID 2) | org.apache.spark.internal.Logging$class.logError(Logging.scala:91)
java.lang.OutOfMemoryError: Unable to acquire 262144 bytes of memory, got 0
        at org.apache.spark.memory.MemoryConsumer.allocateArray(MemoryConsumer.java:100)
        at org.apache.spark.unsafe.map.BytesToBytesMap.allocate(BytesToBytesMap.java:791)
        at org.apache.spark.unsafe.map.BytesToBytesMap.<init>(BytesToBytesMap.java:208)
        at org.apache.spark.unsafe.map.BytesToBytesMap.<init>(BytesToBytesMap.java:223)
        at
org.apache.spark.sql.execution.UnsafeFixedWidthAggregationMap.<init>(UnsafeFixedWidthAggregationMap.java:104)
        at
org.apache.spark.sql.execution.aggregate.HashAggregateExec.createHashMap(HashAggregateExec.scala:307)
        at org.apache.spark.sql.catalyst.expressions.GeneratedClass
$GeneratedIterator.agg_doAggregateWithKeys$(Unknown Source)
        at org.apache.spark.sql.catalyst.expressions.GeneratedClass$GeneratedIterator.processNext(Unknown Source)
        at org.apache.spark.sql.execution.BufferedRowIterator.hasNext(BufferedRowIterator.java:43)
        at org.apache.spark.sql.execution.WholeStageCodegenExec$$anonfun$8$$anon
$1.hasNext(WholeStageCodegenExec.scala:381)
        at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:408)
        at
org.apache.spark.shuffle.sort.BypassMergeSortShuffleWriter.write(BypassMergeSortShuffleWriter.java:126)
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:96)
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:53)
        at org.apache.spark.scheduler.Task.run(Task.scala:99)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:325)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
```

12.23.7.3 Spark Streaming 调优

操作场景

Streaming作为一种mini-batch方式的流式处理框架，它主要的特点是：秒级时延和高吞吐量。因此Streaming调优的目标：在秒级延迟的情景下，提高Streaming的吞吐能力，在单位时间处理尽可能多的数据。

说明

本章节适用于输入数据源为Kafka的使用场景。

操作步骤

一个简单的流处理系统由以下三部分组件组成：数据源 + 接收器 + 处理器。数据源为Kafka，接受器为Streaming中的Kafka数据源接收器，处理器为Streaming。

对Streaming调优，就必须使该三个部件的性能都合适。

- **数据源调优**

  在实际的应用场景中，数据源为了保证数据的容错性，会将数据保存在本地磁盘中，而Streaming的计算结果全部在内存中完成，数据源很有可能成为流式系统的最大瓶颈点。

  对Kafka的性能调优，有以下几个点：

  - 使用Kafka-0.8.2以后版本，可以使用异步模式的新Producer接口。
  - 配置多个Broker的目录，设置多个IO线程，配置Topic合理的Partition个数。

  详情请参见Kafka开源文档中的“性能调优”部分：**http://kafka.apache.org/documentation.html**

- **接收器调优**

  Streaming中已有多种数据源的接收器，例如Kafka、Flume、MQTT、ZeroMQ等，其中Kafka的接收器类型最多，也是最成熟一套接收器。

  Kafka包括三种模式的接收器API：

  - KafkaReceiver：直接接收Kafka数据，进程异常后，可能出现数据丢失。
  - ReliableKafkaReceiver：通过ZooKeeper记录接收数据位移。
  - DirectKafka：直接通过RDD读取Kafka每个Partition中的数据，数据高可靠。

  从实现上来看，DirectKafka的性能更好，实际测试上来看，DirectKafka也确实比其他两个API性能好了不少。因此推荐使用DirectKafka的API实现接收器。

  数据接收器作为一个Kafka的消费者，对于它的配置优化，请参见Kafka开源文档：**http://kafka.apache.org/documentation.html**

- **处理器调优**

  Spark Streaming的底层由Spark执行，因此大部分对于Spark的调优措施，都可以应用在Spark Streaming之中，例如：

  - 数据序列化
  - 配置内存
  - 设置并行度
  - 使用External Shuffle Service提升性能

  说明

  在做Spark Streaming的性能优化时需注意一点，越追求性能上的优化，Spark Streaming整体的可靠性会越差。例如：

  “spark.streaming.receiver.writeAheadLog.enable”配置为“false”的时候，会明显减少磁盘的操作，提高性能，但由于缺少WAL机制，会出现异常恢复时，数据丢失。

  因此，在调优Spark Streaming的时候，这些保证数据可靠性的配置项，在生产环境中是不能关闭的。

- **日志归档调优**

  参数“spark.eventLog.group.size”用来设置一个应用的JobHistory日志按照指定job个数分组，每个分组会单独创建一个文件记录日志，从而避免应用长期运行时形成单个过大日志造成JobHistory无法读取的问题，设置为“0”时表示不分组。

  大部分Spark Streaming任务属于小型job，而且产生速度较快，会导致频繁的分组，产生大量日志小文件消耗磁盘I/O。建议增大此值，例如改为“1000”或更大值。

12.23.8 Spark2x 常见问题

12.23.8.1 Spark Core

12.23.8.1.1 日志聚合下，如何查看 Spark 已完成应用日志

问题

当YARN开启了日志聚合功能时，如何在页面看到聚合后的container日志?

回答

请参考**配置WebUI上查看聚合后的container日志**。

12.23.8.1.2 Driver 返回码和 RM WebUI 上应用状态显示不一致

问题

ApplicationMaster与ResourceManager之间通信发生长时间异常时，为什么Driver返回码和RM WebUI上应用状态显示不一致?

回答

在yarn-client模式下，Spark的Driver和ApplicationMaster作为两个独立的进程在运行。当Driver完成任务退出时，会通知ApplicationMaster向ResourceManager注销自身，即调用unregister方法。

由于是远程调用，则存在发生网络故障的可能性。当发生网络故障时，ApplicationMaster会使用Yarn客户端的重试机制进行重试。在达到最大重试次数之前网络恢复正常，则ApplicationMaster会正常退出。

若超过重试次数和重试时长，则ApplicationMaster注销失败，ResourceManager会认为ApplicationMaster异常退出并尝试重新启动ApplicationMaster。新启动的ApplicationMaster在尝试连接已经退出的Driver失败后，会在ResourceManager页面上标记此次Application为FAILED状态。

这种情况为小概率事件且不影响Spark SQL对外展现的应用完成状态。也可以通过增大Yarn客户端连接次数和连接时长的方式减少此事件发生的概率。配置详情请参见：**http://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-common/yarn-default.xml**

12.23.8.1.3 为什么 Driver 进程不能退出

问题

运行Spark Streaming任务，然后使用**yarn application -kill applicationID**命令停止任务，为什么Driver进程不能退出?

回答

使用**yarn application -kill applicationID**命令后Spark只会停掉任务对应的SparkContext，而不是退出当前进程。如果当前进程中存在其他常驻的线程（类似spark-shell需要不断检测命令输入，Spark Streaming不断在从数据源读取数据），SparkContext被停止并不会终止整个进程。

如果需要退出Driver进程，建议使用**kill -9 pid**命令手动退出当前Driver。

12.23.8.1.4 网络连接超时导致 FetchFailedException

问题

在380节点的大集群上，运行29T数据量的HiBench测试套中ScalaSort测试用例，使用以下关键配置（**--executor-cores 4**）出现如下异常：

```
org.apache.spark.shuffle.FetchFailedException: Failed to connect to /192.168.114.12:23242
  at
```

```
org.apache.spark.storage.ShuffleBlockFetcherIterator.throwFetchFailedException(ShuffleBlockFetcherIterator.s
cala:321)
    at org.apache.spark.storage.ShuffleBlockFetcherIterator.next(ShuffleBlockFetcherIterator.scala:306)
    at org.apache.spark.storage.ShuffleBlockFetcherIterator.next(ShuffleBlockFetcherIterator.scala:51)
    at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
    at scala.collection.Iterator$$anon$13.hasNext(Iterator.scala:371)
    at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:327)
    at org.apache.spark.util.CompletionIterator.hasNext(CompletionIterator.scala:32)
    at org.apache.spark.InterruptibleIterator.hasNext(InterruptibleIterator.scala:39)
    at org.apache.spark.util.collection.ExternalSorter.insertAll(ExternalSorter.scala:217)
    at org.apache.spark.shuffle.hash.HashShuffleReader.read(HashShuffleReader.scala:102)
    at org.apache.spark.rdd.ShuffledRDD.compute(ShuffledRDD.scala:90)
    at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:301)
    at org.apache.spark.rdd.RDD.iterator(RDD.scala:265)
    at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:38)
    at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:301)
    at org.apache.spark.rdd.RDD.iterator(RDD.scala:265)
    at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:38)
    at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:301)
    at org.apache.spark.rdd.RDD.iterator(RDD.scala:265)
    at org.apache.spark.rdd.UnionRDD.compute(UnionRDD.scala:87)
    at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:301)
    at org.apache.spark.rdd.RDD.iterator(RDD.scala:265)
    at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:73)
    at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:41)
    at org.apache.spark.scheduler.Task.run(Task.scala:87)
    at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:213)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)
Caused by: java.io.IOException: Failed to connect to /192.168.114.12:23242
    at org.apache.spark.network.client.TransportClientFactory.createClient(TransportClientFactory.java:214)
    at org.apache.spark.network.client.TransportClientFactory.createClient(TransportClientFactory.java:167)
    at org.apache.spark.network.netty.NettyBlockTransferService$$anon
$1.createAndStart(NettyBlockTransferService.scala:91)
    at
org.apache.spark.network.shuffle.RetryingBlockFetcher.fetchAllOutstanding(RetryingBlockFetcher.java:140)
    at org.apache.spark.network.shuffle.RetryingBlockFetcher.access$200(RetryingBlockFetcher.java:43)
    at org.apache.spark.network.shuffle.RetryingBlockFetcher$1.run(RetryingBlockFetcher.java:170)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    ... 3 more
Caused by: java.net.ConnectException: Connection timed out: /192.168.114.12:23242
    at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
    at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
    at io.netty.channel.socket.nio.NioSocketChannel.doFinishConnect(NioSocketChannel.java:224)
    at io.netty.channel.nio.AbstractNioChannel
$AbstractNioUnsafe.finishConnect(AbstractNioChannel.java:289)
    at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:528)
    at io.netty.channel.nio.NioEventLoop.processSelectedKeysOptimized(NioEventLoop.java:468)
    at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:382)
    at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:354)
    at io.netty.util.concurrent.SingleThreadEventExecutor$2.run(SingleThreadEventExecutor.java:111)
    ... 1 more
```

回答

在运行应用程序时，使用Executor参数“--executor-cores 4”，单进程中并行度高导致IO非常繁忙，以至于任务运行缓慢。

```
16/02/26 10:04:53 INFO TaskSetManager: Finished task 2139.0 in stage 1.0 (TID 151149) in 376455 ms on
10-196-115-2 (694/153378)
```

单个任务运行时间超过6分钟，从而导致连接超时问题，最终使得任务失败。

将参数中的核数设置为1，“--executor-cores 1”，任务正常完成，单个任务处理时间在合理范围之内(15秒左右)。

```
16/02/29 02:24:46 INFO TaskSetManager: Finished task 59564.0 in stage 1.0 (TID 208574) in 15088 ms on
10-196-115-6 (59515/153378)
```

因此，处理这类网络超时任务，可以减少单个Executor的核数来规避该类问题。

12.23.8.1.5 当事件队列溢出时如何配置事件队列的大小

问题

当Driver日志中出现如下的日志时，表示事件队列溢出了。当事件队列溢出时如何配置事件队列的大小?

- 普通应用

  ```
  Dropping SparkListenerEvent because no remaining room in event queue.
  This likely means one of the SparkListeners is too slow and cannot keep
  up with the rate at which tasks are being started by the scheduler.
  ```

- Spark Streaming应用

  ```
  Dropping StreamingListenerEvent because no remaining room in event queue.
  This likely means one of the StreamingListeners is too slow and cannot keep
  up with the rate at which events are being started by the scheduler.
  ```

回答

1. 停止应用，在Spark的配置文件“spark-defaults.conf”中将配置项“spark.event.listener.logEnable”配置为“true”。并把配置项“spark.eventQueue.size”配置为1000W。如果需要控制打印频率（默认为1000毫秒打印1条日志），请根据需要修改配置项“spark.event.listener.logRate”，该配置项的单位为毫秒。
2. 启动应用，可以发现如下的日志信息（消费者速率、生产者速率、当前队列中的消息数量和队列中消息数量的最大值）。

   ```
   INFO LiveListenerBus: [SparkListenerBus]:16044 events are consumed in 5000 ms.
   INFO LiveListenerBus: [SparkListenerBus]:51381 events are produced in 5000 ms, eventQueue still has
   86417 events, MaxSize: 171764.
   ```

3. 用户可以根据日志信息【队列中消息数量的最大值MaxSize】，在配置文件“spark-defaults.conf”中将配置项“spark.eventQueue.size”配置成合适的队列大小。比如【队列中消息数量的最大值】为250000，那么配置合适的队列大小为300000。

12.23.8.1.6 Spark 应用执行过程中，日志中一直打印 getApplicationReport 异常且应用较长时间不退出

问题

Spark应用执行过程中，当driver连接RM失败时，会报下面的错误，且较长时间不退出。

```
16/04/23 15:31:44 INFO RetryInvocationHandler: Exception while invoking getApplicationReport of class
ApplicationClientProtocolPBClientImpl over 37 after 1 fail over attempts. Trying to fail over after sleeping
for 44160ms.
java.net.ConnectException: Call From vm1/192.168.39.30 to vm1:8032 failed on connection exception:
java.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/hadoop/
ConnectionRefused
```

回答

在Spark中有个定期线程，通过连接RM监听AM的状态。由于连接RM超时，就会报上面的错误，且一直重试。RM中对重试次数有限制，默认是30次，每次间隔默认为30秒左右，每次重试时都会报上面的错误。超过次数后，driver才会退出。

RM中关于重试相关的配置项如表12-421所示。

表 12-421 参数说明

参数	描述	默认值
yarn.resourcemanager.connect.max-wait.ms	连接RM的等待时间最大值。	900000
yarn.resourcemanager.connect.retry-interval.ms	重试连接RM的时间频率。	30000

重试次数=yarn.resourcemanager.connect.max-wait.ms/yarn.resourcemanager.connect.retry-interval.ms，即重试次数=连接RM的等待时间最大值/重试连接RM的时间频率。

在Spark客户端机器中，通过修改“conf/yarn-site.xml”文件，添加并配置“yarn.resourcemanager.connect.max-wait.ms”和“yarn.resourcemanager.connect.retry-interval.ms”，这样可以更改重试次数，Spark应用可以提早退出。

12.23.8.1.7 Spark 执行应用时上报“Connection to ip:port has been quiet for xxx ms while there are outstanding requests”并导致应用结束

问题

Spark执行应用时上报如下类似错误并导致应用结束。

```
2016-04-20 10:42:00,557 | ERROR | [shuffle-server-2] | Connection to 10-91-8-208/10.18.0.115:57959 has been quiet for 180000 ms while there are outstanding requests. Assuming connection is dead; please adjust spark.network.timeout if this is wrong. | org.apache.spark.network.server.TransportChannelHandler.userEventTriggered(TransportChannelHandler.java:128)
2016-04-20 10:42:00,558 | ERROR | [shuffle-server-2] | Still have 1 requests outstanding when connection from 10-91-8-208/10.18.0.115:57959 is closed | org.apache.spark.network.client.TransportResponseHandler.channelUnregistered(TransportResponseHandler.java:102)
2016-04-20 10:42:00,562 | WARN  | [yarn-scheduler-ask-am-thread-pool-160] | Error sending message [message = DoShuffleClean(application_1459995017785_0108,319)] in 1 attempts | org.apache.spark.Logging$class
s.logWarning(Logging.scala:92)
java.io.IOException: Connection from 10-91-8-208/10.18.0.115:57959 closed
        at org.apache.spark.network.client.TransportResponseHandler.channelUnregistered(TransportResponseHandler.java:104)
        at org.apache.spark.network.server.TransportChannelHandler.channelUnregistered(TransportChannelHandler.java:94)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelUnregistered(AbstractChannelHandlerContext.java:158)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelUnregistered(AbstractChannelHandlerContext.java:144)
```

```
        at
io.netty.channel.ChannelInboundHandlerAdapter.channelUnregistered(ChannelInboundHandlerAdapter.java:5
3)
        at
io.netty.channel.AbstractChannelHandlerContext.invokeChannelUnregistered(AbstractChannelHandlerContex
t.java:158)
        at
io.netty.channel.AbstractChannelHandlerContext.fireChannelUnregistered(AbstractChannelHandlerContext.ja
va:144)
        at
io.netty.channel.ChannelInboundHandlerAdapter.channelUnregistered(ChannelInboundHandlerAdapter.java:5
3)
        at
io.netty.channel.AbstractChannelHandlerContext.invokeChannelUnregistered(AbstractChannelHandlerContex
t.java:158)
        at
io.netty.channel.AbstractChannelHandlerContext.fireChannelUnregistered(AbstractChannelHandlerContext.ja
va:144)
        at
io.netty.channel.ChannelInboundHandlerAdapter.channelUnregistered(ChannelInboundHandlerAdapter.java:5
3)
        at
io.netty.channel.AbstractChannelHandlerContext.invokeChannelUnregistered(AbstractChannelHandlerContex
t.java:158)
        at
io.netty.channel.AbstractChannelHandlerContext.fireChannelUnregistered(AbstractChannelHandlerContext.ja
va:144)
        at io.netty.channel.DefaultChannelPipeline.fireChannelUnregistered(DefaultChannelPipeline.java:739)
        at io.netty.channel.AbstractChannel$AbstractUnsafe$8.run(AbstractChannel.java:659)
        at io.netty.util.concurrent.SingleThreadEventExecutor.runAllTasks(SingleThreadEventExecutor.java:357)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:357)
        at io.netty.util.concurrent.SingleThreadEventExecutor$2.run(SingleThreadEventExecutor.java:111)
        at java.lang.Thread.run(Thread.java:745)
2016-04-20 10:42:00,573 | INFO  | [dispatcher-event-loop-14] | Starting task 177.0 in stage 1492.0 (TID
1996351, linux-254, PROCESS_LOCAL, 2106 bytes) | org.apache.spark.Logging$class.logInfo(Logging.scala:
59)
2016-04-20 10:42:00,574 | INFO  | [task-result-getter-0] | Finished task 85.0 in stage 1492.0 (TID 1996259)
in 191336 ms on linux-254 (106/3000) | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
2016-04-20 10:42:00,811 | ERROR | [Yarn application state monitor] | Yarn application has already exited
with state FINISHED! | org.apache.spark.Logging$class.logError(Logging.scala:75)
```

回答

当配置channel过期时间（spark.rpc.io.connectionTimeout） < RPC响应超时时间（spark.rpc.askTimeout），在特殊条件下（Full GC，网络延时等）消息响应时间较长，消息还没有反馈，channel又达到了过期时间，该channel就被终止了，AM端感知到channel被终止后认为driver失联，然后整个应用停止。

解决办法：在Spark客户端的“spark-defaults.conf”文件中或通过set命令行进行设置。参数配置时要保证channel过期时间（spark.rpc.io.connectionTimeout）大于或等于RPC响应超时时间（spark.rpc.askTimeout）。

表 12-422 参数说明

参数	描述	默认值
spark.rpc.askTimeout	RPC响应超时时间，不配置的话默认使用spark.network.timeout的值。	120s

12.23.8.1.8 NodeManager 关闭导致 Executor(s)未移除

问题

在Executor动态分配打开的情况下，如果在任务执行过程中，执行NodeManager关闭动作，NodeManager关闭节点上的Executor(s)在空闲超时之后，在driver页面上未被移除。

回答

这是因为ResourceManager感知到NodeManager关闭时，Executor(s)已经因空闲超时而被driver请求kill掉，但因NodeManager已经关闭，这些Executor(s)实际上并不能被kill掉，因此driver不能感知到这些Executor(s)的LOST事件，所以并未从自身的Executor list中移除，从而导致在driver页面上还能看到这些Executor(s)，这是YARN NodeManager关闭之后的正常现象，NodeManager再次启动后，这些Executor(s)会被移除。

12.23.8.1.9 Password cannot be null if SASL is enabled 异常

问题

运行Spark的应用启用了ExternalShuffle，应用出现了Task任务丢失，原因是由于java.lang.NullPointerException: Password cannot be null if SASL is enabled异常，部分关键日志如下图所示：

```
2016-05-13 12:05:27,093 | WARN  | [task-result-getter-2] | Lost task 98.0 in stage 22.1 (TID 193603, linux-173, 2): FetchFailed(BlockManagerId(13, 172.168.100.13, 27337),
org.apache.spark.shuffle.FetchFailedException: java.lang.NullPointerException: Password cannot be null if SASL is enabled
        at org.spark-project.guava.base.Preconditions.checkNotNull(Preconditions.java:208)
        at org.apache.spark.network.sasl.SparkSaslServer.encodePassword(SparkSaslServer.java:196)
        at org.apache.spark.network.sasl.SparkSaslServer$DigestCallbackHandler.handle(SparkSaslServer.java:166)
        at com.sun.security.sasl.digest.DigestMD5Server.validateClientResponse(DigestMD5Server.java:589)
        at com.sun.security.sasl.digest.DigestMD5Server.evaluateResponse(DigestMD5Server.java:244)
        at org.apache.spark.network.sasl.SparkSaslServer.response(SparkSaslServer.java:119)
        at org.apache.spark.network.sasl.SaslRpcHandler.receive(SaslRpcHandler.java:100)
        at org.apache.spark.network.server.TransportRequestHandler.processRpcRequest(TransportRequestHandler.java:128)
        at org.apache.spark.network.server.TransportRequestHandler.handle(TransportRequestHandler.java:99)
        at org.apache.spark.network.server.TransportChannelHandler.channelRead0(TransportChannelHandler.java:104)
```

回答

造成该现象的原因是NodeManager重启。使用ExternalShuffle的时候，Spark将借用NodeManager传输Shuffle数据，因此NodeManager的内存将成为瓶颈。

在当前版本的FusionInsight中，NodeManager的默认内存只有1G，在数据量比较大（1T以上）的Spark任务下，内存严重不足，消息响应缓慢，导致FusionInsight健康检查认为NodeManager进程退出，强制重启NodeManager，导致上述问题产生。

解决方式：

调整NodeManager的内存，数据量比较大（1T以上）的情况下，NodeManager的内存至少在4G以上。

12.23.8.1.10 向动态分区表中插入数据时，在重试的 task 中出现"Failed to CREATE_FILE"异常

问题

向动态分区表中插入数据时，shuffle过程中大面积shuffle文件损坏（磁盘掉线、节点故障等）后，为什么会在重试的task中出现"Failed to CREATE_FILE"异常?

```
2016-06-25 15:11:31,323 | ERROR | [Executor task launch worker-0] | Exception in task 15.0 in stage 10.1 (TID 1258) | org.apache.spark.Logging$class.logError(Logging.scala:96)
```

```
org.apache.hadoop.hive.ql.metadata.HiveException:
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.protocol.AlreadyBeingCreatedException):
Failed to CREATE_FILE /user/hive/warehouse/testdb.db/we
b_sales/.hive-staging_hive_2016-06-25_15-09-16_999_8137121701603617850-1/-ext-10000/_temporary/0/
_temporary/attempt_201606251509_0010_m_000015_0/ws_sold_date=1999-12-17/part-00015 for
DFSClient_attempt_2016
06251509_0010_m_000015_0_353134803_151 on 10.1.1.5 because this file lease is currently owned by
DFSClient_attempt_201606251509_0010_m_000015_0_-848353830_156 on 10.1.1.6
```

回答

动态分区表插入数据的最后一步是读取shuffle文件的数据，再写入到表对应的分区文件中。

当大面积shuffle文件损坏后，会引起大批量task失败，然后进行job重试。重试前Spark会将写表分区文件的句柄关闭，大批量task关闭句柄时HDFS无法及时处理。在task进行下一次重试时，句柄在NameNode端未被及时释放，即会抛出"Failed to CREATE_FILE"异常。

这种现象仅会在大面积shuffle文件损坏时发生，出现异常后task会重试，重试耗时在毫秒级，影响较小，可以忽略不计。

12.23.8.1.11 使用 Hash shuffle 出现任务失败

问题

使用Hash shuffle运行1000000（map个数）*100000（reduce个数）的任务，运行日志中出现大量的消息发送失败和Executor心跳超时，从而导致任务失败。

回答

对于Hash shuffle，在shuffle的过程中写数据时不做排序操作，只是将数据根据Hash的结果，将各个reduce分区的数据写到各自的磁盘文件中。

这样带来的问题是如果reduce分区的数量比较大的话，将会产生大量的磁盘文件（比如：该问题中将产生1000000 * 100000 = 10^11个shuffle文件）。如果磁盘文件数量特别巨大，对文件读写的性能会带来比较大的影响，此外由于同时打开的文件句柄数量多，序列化以及压缩等操作需要占用非常大的临时内存空间，对内存的使用和GC带来很大的压力，从而容易造成Executor无法响应Driver。

因此，建议使用Sort shuffle，而不使用Hash shuffle。

12.23.8.1.12 访问 Spark 应用的聚合日志页面报“DNS 查找失败”错误

问题

采用http(s)://<spark ip>:<spark port>的方式直接访问Spark JobHistory页面时，如果当前跳转的Spark JobHistory页面不是FusionInsight代理的页面（FusionInsight代理的URL地址类似于：https://<oms ip>:20026/Spark2x/JobHistory2x/xx/），单击某个应用，再单击“AggregatedLogs”，然后单击需要查看的其中一个Executor的“logs”，此时会报如**图12-58**所示的错误。

图 12-58 聚合日志失败页面

回答

原因：弹出的URL地址（如https://<hostname>:20026/Spark2x/JobHistory2x/xx/history/application_xxx/jobs/），其中的<hostname>没有在Windows系统的hosts文件中添加域名信息，导致DNS查找失败无法显示此网页。

解决措施：

- 建议用户使用FusionInsight代理去访问Spark JobHistory页面。
- 如果用户需要不通过FusionInsight Manager访问Spark JobHistory页面，则需要将URL地址中的<hostname>更改为IP地址进行访问，或者在Windows系统的hosts文件中添加该域名信息。

12.23.8.1.13 由于 Timeout waiting for task 异常导致 Shuffle FetchFailed

问题

使用JDBCServer模式执行100T的TPCDS测试套，出现Timeout waiting for task异常导致Shuffle FetchFailed，Stage一直重试，任务无法正常完成。

回答

JDBCServer方式使用了ShuffleService功能，Reduce阶段所有的Executor会从NodeManager中获取数据，当数据量达到一个级别（10T级别），会出现NodeManager单点瓶颈（ShuffleService服务在NodeManager进程中），就会出现某些Task获取数据超时，从而出现该问题。

因此，当数据量达到10T级别以上的Spark任务，建议用户关闭ShuffleService功能，即在“Spark-defaults.conf”配置文件中将配置项“spark.shuffle.service.enabled”配置为“false”。

12.23.8.1.14 Executor 进程 Crash 导致 Stage 重试

问题

在执行大数据量的Spark任务（如100T的TPCDS测试套）过程中，有时会出现Executor丢失从而导致Stage重试的现象。查看Executor的日志，出现“Executor 532 is lost rpc with driver,but is still alive, going to kill it”所示信息，表明Executor丢失是由于JVM Crash导致的。

JVM的关键Crash错误日志，如下：

```
#
# A fatal error has been detected by the Java Runtime Environment:
#
#  Internal Error (sharedRuntime.cpp:834), pid=241075, tid=140476258551552
#  fatal error: exception happened outside interpreter, nmethods and vtable stubs at pc
0x00007fcda9eb8eb1
```

回答

上述问题在Oracle官网上有类似的情况，该问题现象是Oracle JVM的缺陷，并不是平台代码引入的问题，且Spark中有对Executor的容错机制，Executor Crash之后，Stage会进入重试，可以保证任务最终可以执行完成，不会对业务产生影响。

12.23.8.1.15 执行大数据量的 shuffle 过程时 Executor 注册 shuffle service 失败

问题

执行超过50T数据的shuffle过程时，出现部分Executor注册shuffle service超时然后丢失从而导致任务失败的问题。错误日志如下所示：

```
2016-10-19 01:33:34,030 | WARN | ContainersLauncher #14 | Exception from container-launch with
container ID: container_e1452_1476801295027_2003_01_004512 and exit code: 1 |
LinuxContainerExecutor.java:397
ExitCodeException exitCode=1:
at org.apache.hadoop.util.Shell.runCommand(Shell.java:561)
at org.apache.hadoop.util.Shell.run(Shell.java:472)
at org.apache.hadoop.util.Shell$ShellCommandExecutor.execute(Shell.java:738)
at
org.apache.hadoop.yarn.server.nodemanager.LinuxContainerExecutor.launchContainer(LinuxContainerExecuto
r.java:381)
at
org.apache.hadoop.yarn.server.nodemanager.containermanager.launcher.ContainerLaunch.call(ContainerLaun
ch.java:312)
at
org.apache.hadoop.yarn.server.nodemanager.containermanager.launcher.ContainerLaunch.call(ContainerLaun
ch.java:88)
at java.util.concurrent.FutureTask.run(FutureTask.java:266)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
2016-10-19 01:33:34,031 | INFO | ContainersLauncher #14 | Exception from container-launch. |
ContainerExecutor.java:300
2016-10-19 01:33:34,031 | INFO | ContainersLauncher #14 | Container id:
container_e1452_1476801295027_2003_01_004512 | ContainerExecutor.java:300
2016-10-19 01:33:34,031 | INFO | ContainersLauncher #14 | Exit code: 1 | ContainerExecutor.java:300
2016-10-19 01:33:34,031 | INFO | ContainersLauncher #14 | Stack trace: ExitCodeException exitCode=1: |
ContainerExecutor.java:300
```

回答

由于当前数据量较大，有50T数据导入，超过了shuffle的规格，shuffle负载过高，shuffle service服务处于过载状态，可能无法及时响应Executor的注册请求，从而出现上面的问题。

Executor注册shuffle service的超时时间是5秒，最多重试3次，该参数目前不可配。

建议适当调大task retry次数和Executor失败次数。

在客户端的“spark-defaults.conf”配置文件中配置如下参数。
“spark.yarn.max.executor.failures”若不存在，则手动添加该参数项。

表 12-423 参数说明

参数	描述	默认值
spark.task.maxFailures	task retry次数。	4
spark.yarn.max.executor.failures	Executor失败次数。 关闭Executor个数动态分配功能的场景即“spark.dynamicAllocation.enabled”参数设为“false”时。	numExecutors * 2, with minimum of 3
	Executor失败次数。 开启Executor个数动态分配功能的场景即“spark.dynamicAllocation.enabled”参数设为“true”时。	3

12.23.8.1.16 在 Spark 应用执行过程中 NodeManager 出现 OOM 异常

问题

当开启Yarn External Shuffle服务时，在Spark应用执行过程中，如果当前shuffle连接过多，Yarn External Shuffle会出现“java.lang.OutofMemoryError: Direct buffer Memory”的异常，该异常说明内存不足。错误日志如下：

```
2016-12-06 02:01:00,768 | WARN  | shuffle-server-38 | Exception in connection from /192.168.101.95:53680 | TransportChannelHandler.java:79
io.netty.handler.codec.DecoderException: java.lang.OutOfMemoryError: Direct buffer memory
        at io.netty.handler.codec.ByteToMessageDecoder.channelRead(ByteToMessageDecoder.java:153)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:333)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelRead(AbstractChannelHandlerContext.java:319)
        at io.netty.channel.DefaultChannelPipeline.fireChannelRead(DefaultChannelPipeline.java:787)
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:130)
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:511)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysOptimized(NioEventLoop.java:468)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:382)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:354)
        at io.netty.util.concurrent.SingleThreadEventExecutor$2.run(SingleThreadEventExecutor.java:116)
        at java.lang.Thread.run(Thread.java:745)
Caused by: java.lang.OutOfMemoryError: Direct buffer memory
```

```
at java.nio.Bits.reserveMemory(Bits.java:693)
at java.nio.DirectByteBuffer.<init>(DirectByteBuffer.java:123)
at java.nio.ByteBuffer.allocateDirect(ByteBuffer.java:311)
at io.netty.buffer.PoolArena$DirectArena.newChunk(PoolArena.java:434)
at io.netty.buffer.PoolArena.allocateNormal(PoolArena.java:179)
at io.netty.buffer.PoolArena.allocate(PoolArena.java:168)
at io.netty.buffer.PoolArena.reallocate(PoolArena.java:277)
at io.netty.buffer.PooledByteBuf.capacity(PooledByteBuf.java:108)
at io.netty.buffer.AbstractByteBuf.ensureWritable(AbstractByteBuf.java:251)
at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:849)
at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:841)
at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:831)
at io.netty.handler.codec.ByteToMessageDecoder.channelRead(ByteToMessageDecoder.java:146)
... 10 more
```

回答

对于Yarn的Shuffle Service，其启动的线程数为机器可用CPU核数的两倍，而默认配置的Direct buffer Memory为128M，因此当有较多shuffle同时连接时，平均分配到各线程所能使用的Direct buffer Memory将较低（例如，当机器的CPU为40核，Yarn的Shuffle Service启动的线程数为80，80个线程共享进程里的Direct buffer Memory，这种场景下每个线程分配到的内存将不足2MB）。

因此建议根据集群中的NodeManager节点的CPU核数适当调整Direct buffer Memory，例如在CPU核数为40时，将Direct buffer Memory配置为512M。即配置集群NodeManger的“GC_OPTS”参数，如：

-XX:MaxDirectMemorySize=512M

说明

GC_OPTS参数中-XX:MaxDirectMemorySize默认没有配置，如需配置，用户可在GC_OPTS参数中自定义添加。

具体的配置方法如下：

用户可登录FusionInsight Manager，单击“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，单击“全部配置”，单击“NodeManager > 系统”，在“GC_OPTS”参数中修改配置。

表 12-424 参数说明

参数	描述	默认值
GC_OPTS	Yarn NodeManger的GC参数。	128M

12.23.8.1.17 安全集群使用 HiBench 工具运行 sparkbench 获取不到 realm

问题

运行HiBench6的sparkbench任务，如Wordcount，任务执行失败，bench.log显示Yarn任务执行失败，登录Yarn UI，查看对应application的失败信息，显示如下：

```
Exception in thread "main" org.apache.spark.SparkException: Unable to load YARN support
  at org.apache.spark.deploy.SparkHadoopUtil$.liftedTree1$1(SparkHadoopUtil.scala:390)
  at org.apache.spark.deploy.SparkHadoopUtil$.yarn$lzycompute(SparkHadoopUtil.scala:385)
  at org.apache.spark.deploy.SparkHadoopUtil$.yarn(SparkHadoopUtil.scala:385)
  at org.apache.spark.deploy.SparkHadoopUtil$.get(SparkHadoopUtil.scala:410)
```

```
 at org.apache.spark.deploy.yarn.ApplicationMaster$.main(ApplicationMaster.scala:796)
 at org.apache.spark.deploy.yarn.ExecutorLauncher$.main(ApplicationMaster.scala:821)
 at org.apache.spark.deploy.yarn.ExecutorLauncher.main(ApplicationMaster.scala)
Caused by: java.lang.IllegalArgumentException: Can't get Kerberos realm
 at org.apache.hadoop.security.HadoopKerberosName.setConfiguration(HadoopKerberosName.java:65)
 at org.apache.hadoop.security.UserGroupInformation.initialize(UserGroupInformation.java:288)
 at org.apache.hadoop.security.UserGroupInformation.setConfiguration(UserGroupInformation.java:336)
 at org.apache.spark.deploy.SparkHadoopUtil.<init>(SparkHadoopUtil.scala:51)
 at org.apache.spark.deploy.yarn.YarnSparkHadoopUtil.<init>(YarnSparkHadoopUtil.scala:49)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
 at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
 at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
 at java.lang.Class.newInstance(Class.java:442)
 at org.apache.spark.deploy.SparkHadoopUtil$.liftedTree1$1(SparkHadoopUtil.scala:387)
 ... 6 more
Caused by: java.lang.reflect.InvocationTargetException
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:498)
 at org.apache.hadoop.security.authentication.util.KerberosUtil.getDefaultRealm(KerberosUtil.java:88)
 at org.apache.hadoop.security.HadoopKerberosName.setConfiguration(HadoopKerberosName.java:63)
 ... 16 more
Caused by: KrbException: Cannot locate default realm
 at sun.security.krb5.Config.getDefaultRealm(Config.java:1029)
 ... 22 more
```

回答

失败原因是C80SPC200版本开始，安装集群不再替换/etc/krb5.conf文件，改为通过配置参数指定到客户端内krb5路径，而HiBench并不引用客户端配置文件。解决方案：将客户端/opt/client/KrbClient/kerberos/var/krb5kdc/krb5.conf，copy覆盖集群内所有节点的/etc/krb5.conf，注意替换前需要备份。

12.23.8.2 SQL 和 DataFrame

12.23.8.2.1 Spark SQL ROLLUP 和 CUBE 使用的注意事项

问题

假设有表src(d1, d2, m)，其数据如下：

```
1 a 1
1 b 1
2 b 2
```

对于语句select d1, sum(d1) from src group by d1, d2 with rollup其结果如下：

```
NULL 0
1    2
2    2
1    1
1    1
2    2
```

对于以上结果的第一条为什么是(NULL,0)而不是(NULL,4)。

回答

在进行rollup和cube操作时，用户通常是基于维度进行分析，需要的是度量的结果，因此不会对维度进行聚合操作。

例如当前有表src(d1, d2, m)，那么语句1“select d1, sum(m) from src group by d1, d2 with rollup”就是对维度d1和d2进行上卷操作计算度量m的结果，因此有实际业务意义，而其结果也跟预期是一致的。但语句2“select d1, sum(d1) from src group by d1, d2 with rollup”则从业务上无法解释。当前对于语句2所有聚合（sum/avg/max/min）结果均为0。

说明

只有在rollup和cube操作中对出现在group by中的字段进行聚合结果才是0，非rollup和cube操作其结果跟预期一致。

12.23.8.2.2 Spark SQL 在不同 DB 都可以显示临时表

问题

切换数据库之后，为什么还能看到之前数据库的临时表?

1. 创建一个DataSource的临时表，例如以下建表语句。

```
create temporary table ds_parquet
using org.apache.spark.sql.parquet
options(path '/tmp/users.parquet');
```

2. 切换到另外一个数据库，执行***show tables***，依然可以看到上个步骤创建的临时表。

```
0: jdbc:hive2://192.168.169.84:22550/default> show tables;
+-----------------+--------------+--+
|    tableName    | isTemporary  |
+-----------------+--------------+--+
| ds_parquet      | true         |
| cmb_tbl_carbon  | false        |
+-----------------+--------------+--+
2 rows selected (0.109 seconds)
0: jdbc:hive2://192.168.169.84:22550/default>
```

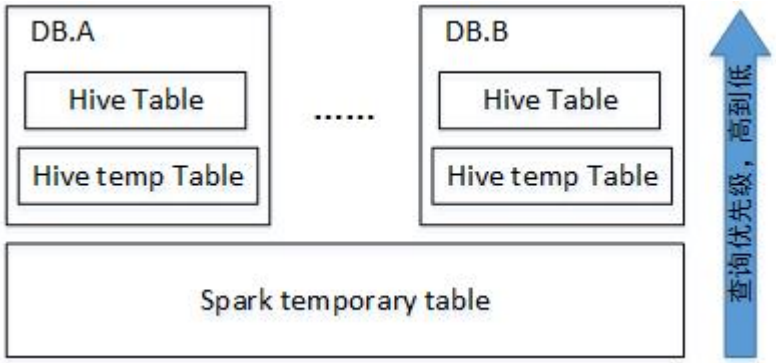
回答

Spark的表管理层次如图12-59所示，最底层是Spark的临时表，存储着使用DataSource方式的临时表，在这一个层面中没有数据库的概念，因此对于这种类型表，表名在各个数据库中都是可见的。

上层为Hive的MetaStore，该层有了各个DB之分。在每个DB中，又有Hive的临时表与Hive的持久化表，因此在Spark中允许三个层次的同名数据表。

查询的时候，Spark SQL优先查看是否有Spark的临时表，再查找当前DB的Hive临时表，最后查找当前DB的Hive持久化表。

图 12-59 Spark 表管理层次



当Session退出时，用户操作相关的临时表将自动删除。建议用户不要手动删除临时表。

删除临时表时，其优先级与查询相同，从高到低为Spark临时表、Hive临时表、Hive持久化表。如果想直接删除Hive表，不删除Spark临时表，您可以直接使用***drop table DbName.TableName***命令。

12.23.8.2.3 如何在 Spark 命令中指定参数值

问题

如果用户不希望在界面上或配置文件设置参数值，如何在Spark命令中指定参数值?

回答

Spark的配置项，不仅可以在配置文件中设置，也可以在命令中指定参数值。

在Spark客户端，应用执行命令添加如下内容设置参数值，命令执行完成后立即生效。在--conf后添加参数名称及其参数值，例如：

--conf spark.eventQueue.size=50000

12.23.8.2.4 SparkSQL 建表时的目录权限

问题

新建的用户，使用SparkSQL建表时出现类似如下错误：

```
0: jdbc:hive2://192.168.169.84:22550/default> create table testACL(c string);
Error: org.apache.spark.sql.execution.QueryExecutionException: FAILED: Execution Error, return code 1 from
org.apache.hadoop.hive.ql.exec.DDLTask. MetaException(message:Got exception:
org.apache.hadoop.security.AccessControlException
Permission denied: user=testACL, access=EXECUTE, inode="/user/hive/warehouse/
testacl":spark:hadoop:drwxrwx---
    at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkAccessAcl(FSPermissionChecker.java:403
)
    at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:306)
    at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:259)
    at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:20
5)
    at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:19
0)
    at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1710)
    at
org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getFileInfo(FSDirStatAndListingOp.java:109)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getFileInfo(FSNamesystem.java:3762)
    at
org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getFileInfo(NameNodeRpcServer.java:1014)
    at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getFileInfo(ClientNamen
odeProtocolServerSideTranslatorPB.java:853)
    at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol
$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:616)
    at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:973)
    at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2089)
    at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2085)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
```

```
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1675)
    at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2083)
) (state=,code=0)
```

回答

Spark SQL建表底层调用的是Hive的接口，其建表时会在“/user/hive/warehouse”目录下新建一个以表名命名的目录，因此要求用户具备“/user/hive/warehouse”目录的读写、执行权限或具有Hive的group权限。

“/user/hive/warehouse”目录可通过hive.metastore.warehouse.dir参数指定。

12.23.8.2.5 为什么不同服务之间互相删除 UDF 失败

问题

不同服务之间互相删除UDF失败，例如，Spark SQL无法删除Hive创建的UDF。

回答

当前可以通过以下3种方式创建UDF：

1. 在Hive端创建UDF。
2. 通过JDBCServer接口创建UDF。用户可以通过Spark Beeline或者JDBC客户端代码来连接JDBCServer，从而执行SQL命令，创建UDF。
3. 通过spark-sql创建UDF。

删除UDF失败，存在以下两种场景：

- 在Spark Beeline中，对于其他方式创建的UDF，需要重新启动Spark服务端的JDBCServer后，才能将此类UDF删除成功，否则删除失败。在spark-sql中，对于其他方式创建的UDF，需要重新启动spark-sql后，才能将此类UDF删除成功，否则删除失败。

  原因：创建UDF后，Spark服务端的JDBCServer未重启或者spark-sql未重新启动的场景，Spark所在线程的FunctionRegistry对象未保存新创建的UDF，那么删除UDF时就会出现错误。

  解决方法：重启Spark服务端的JDBCServer和spark-sql，再删除此类UDF。
- 在Hive端创建UDF时未在创建语句中指定jar包路径，而是通过**add jar**命令添加UDF的jar包如**add jar /opt/test/two_udfs.jar**，这种场景下，在其他服务中删除UDF时就会出现ClassNotfound的错误，从而导致删除失败。

  原因：在删除UDF时，会先获取该UDF，此时会去加载该UDF对应的类，由于创建UDF时是通过**add jar**命令指定jar包路径的，其他服务进程的classpath不存在这些jar包，因此会出现ClassNotfound的错误从而导致删除失败。

  解决方法：该方式创建的UDF不支持通过其他方式删除，只能通过与创建时一致的方式删除。

12.23.8.2.6 Spark SQL 无法查询到 Parquet 类型的 Hive 表的新插入数据

问题

为什么通过Spark SQL无法查询到存储类型为Parquet的Hive表的新插入数据？主要有以下两种场景存在这个问题：

1. 对于分区表和非分区表，在Hive客户端中执行插入数据的操作后，会出现Spark SQL无法查询到最新插入的数据的问题。
2. 对于分区表，在Spark SQL中执行插入数据的操作后，如果分区信息未改变，会出现Spark SQL无法查询到最新插入的数据的问题。

回答

由于Spark存在一个机制，为了提高性能会缓存Parquet的元数据信息。当通过Hive或其他方式更新了Parquet表时，缓存的元数据信息未更新，导致Spark SQL查询不到新插入的数据。

对于存储类型为Parquet的Hive分区表，在执行插入数据操作后，如果分区信息未改变，则缓存的元数据信息未更新，导致Spark SQL查询不到新插入的数据。

解决措施：在使用Spark SQL查询之前，需执行Refresh操作更新元数据信息。

REFRESH TABLE table_name;

*table_name*为刷新的表名，该表必须存在，否则会出错。

执行查询语句时，即可获取到最新插入的数据。

Spark官网提供了此机制的描述，详情请参见：**https://archive.apache.org/dist/spark/docs/3.1.1/sql-programming-guide.html#metadata-refreshing**

12.23.8.2.7 cache table 使用指导

问题

cache table的作用是什么？cache table时需要注意哪些方面？

回答

Spark SQL可以将表cache到内存中，并且使用压缩存储来尽量减少内存压力。通过将表cache，查询可以直接从内存中读取数据，从而减少读取磁盘带来的内存开销。

但需要注意的是，被cache的表会占用executor的内存。尽管在Spark SQL采用压缩存储的方式来尽量减少内存开销、缓解GC压力，但当缓存的表较大或者缓存表数量较多时，将不可避免的影响executor的稳定性。

此时的最佳实践是，当不需要将表cache来实现查询加速时，应及时将表进行uncache以释放内存。可以执行命令***uncache table*** *table_name*来uncache表。

说明

被cache的表也可以在Spark Driver UI的Storage标签里查看。

12.23.8.2.8 Repartition 时有部分 Partition 没数据

问题

在repartition操作时，分块数“spark.sql.shuffle.partitions”设置为4500，repartition用到的key列中有超过4000个的不同key值。期望不同key对应的数据能分到不同的partition，实际上却只有2000个partition里有数据，不同key对应的数据也被分到相同的partition里。

回答

这是正常现象。

数据分到哪个partition是通过对key的hashcode取模得到的，不同的hashcode取模后的结果有可能是一样的，那样数据就会被分到相同的partition里面，因此出现有些partition没有数据而有些partition里面有多个key对应的数据。

通过调整“spark.sql.shuffle.partitions”参数值可以调整取模时的基数，改善数据分块不均匀的情况，多次验证发现配置为质数或者奇数效果比较好。

在Driver端的“spark-defaults.conf”配置文件中调整如下参数。

表 12-425 参数说明

参数	描述	默认值
spark.sql.shuffle.partitions	shuffle操作时，shuffle数据的分块数。	200

12.23.8.2.9 16T 的文本数据转成 4T Parquet 数据失败

问题

使用默认配置时，16T的文本数据转成4T Parquet数据失败，报如下错误信息。

```
Job aborted due to stage failure: Task 2866 in stage 11.0 failed 4 times, most recent failure: Lost task 2866.6 in stage 11.0 (TID 54863, linux-161, 2): java.io.IOException: Failed to connect to /10.16.1.11:23124
at org.apache.spark.network.client.TransportClientFactory.createClient(TransportClientFactory.java:214)
at org.apache.spark.network.client.TransportClientFactory.createClient(TransportClientFactory.java:167)
at org.apache.spark.network.netty.NettyBlockTransferService$$anon
$1.createAndStart(NettyBlockTransferService.scala:92)
```

使用的默认配置如表12-426所示。

表 12-426 参数说明

参数	描述	默认值
spark.sql.shuffle.partitions	shuffle操作时，shuffle数据的分块数。	200
spark.shuffle.sasl.timeout	shuffle操作时SASL认证的超时时间。单位：秒。	120s
spark.shuffle.io.connectionTimeout	shuffle操作时连接远程节点的超时时间。单位：秒。	120s
spark.network.timeout	所有涉及网络连接操作的超时时间。单位：秒。	360s

回答

由于当前数据量较大，有16T，而分区数只有200，造成每个task任务过重，才会出现上面的问题。

为了解决上面问题，需要对参数进行调整。

- 增大partition数，把任务切分的更小。
- 增大任务执行过程中的超时时间。

在客户端的“spark-defaults.conf”配置文件中配置如下参数。

表 12-427 参数说明

参数	描述	建议值
spark.sql.shuffle.partitions	shuffle操作时，shuffle数据的分块数。	4501
spark.shuffle.sasl.timeout	shuffle操作时SASL认证的超时时间。单位：秒。	2000s
spark.shuffle.io.connectionTimeout	shuffle操作时连接远程节点的超时时间。单位：秒。	3000s
spark.network.timeout	所有涉及网络连接操作的超时时间。单位：秒。	360s

12.23.8.2.10 当表名为 table 时，执行相关操作时出现异常

问题

当创建了表名为table的表后，执行**drop table table**上报以下错误，或者执行其他操作也会出现类似错误。

```
16/07/12 18:56:29 ERROR SparkSQLDriver: Failed in [drop table table]
java.lang.RuntimeException: [1.1] failure: identifier expected
table
^
at scala.sys.package$.error(package.scala:27)
at org.apache.spark.sql.catalyst.SqlParserTrait$class.parseTableIdentifier(SqlParser.scala:56)
at org.apache.spark.sql.catalyst.SqlParser$.parseTableIdentifier(SqlParser.scala:485)
```

回答

这是因为table为Spark SQL的关键词，不能作为表名使用。建议用户不要使用table作为表的名字。

12.23.8.2.11 执行 analyze table 语句，因资源不足出现任务卡住

问题

使用spark-sql执行***analyze table***语句，任务一直卡住，打印的信息如下：

```
spark-sql> analyze table hivetable2 compute statistics;
Query ID = root_20160716174218_90f55869-000a-40b4-a908-533f63866fed
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
16/07/20 17:40:56 WARN JobResourceUploader: Hadoop command-line option parsing not performed.
Implement the Tool interface and execute your application with ToolRunner to remedy this.
Starting Job = job_1468982600676_0002, Tracking URL = http://10-120-175-107:8088/proxy/
application_1468982600676_0002/
Kill Command = /opt/hadoopclient/HDFS/hadoop/bin/hadoop job  -kill job_1468982600676_0002
```

回答

执行***analyze table hivetable2 compute statistics***语句时，由于该sql语句会启动MapReduce任务。从YARN的ResourceManager Web UI页面看到，该任务由于资源不足导致任务没有被执行，表现出任务卡住的现象。

图 12-60 ResourceManager Web UI 页面

建议用户执行***analyze table***语句时加上noscan，其功能与***analyze table hivetable2 compute statistics***语句相同，具体命令如下：

```
spark-sql> analyze table hivetable2 compute statistics noscan
```

该命令不用启动MapReduce任务，不会占用YARN资源，从而任务可以被执行。

12.23.8.2.12 为什么有时访问没有权限的 parquet 表时，在上报“Missing Privileges”错误提示之前，会运行一个 Job?

问题

为什么有时访问没有权限的parquet表时，在上报“Missing Privileges”错误提示之前，会运行一个Job?

回答

Spark SQL对用户SQL语句的执行逻辑是：首先解析出语句中包含的表，再获取表的元数据信息，然后对权限进行检查。

当表是parquet表时，元数据信息包括文件的Split信息。Split信息需要调用HDFS的接口去读取，当表包含的文件数量很多时，串行读取Split信息变得缓慢，影响性能。故对此做了优化，当表包含的文件大于一定阈值（即spark.sql.sources.parallelSplitDiscovery.threshold参数值）时，会生成一个Job，利用Executor的并行能力去读取，从而提升执行效率。

由于权限检查在获取表元数据之后，因此当读取的parquet表包含的文件数量很多时，会在报“Missing Privileges”之前，运行一个Job来并行读取元数据信息。

12.23.8.2.13 执行 Hive 命令修改元数据时失败或不生效

问题

对于datasource表和Spark on HBase表，执行Hive相关命令修改元数据时，出现失败或者不生效情况。

回答

当前版本不支持执行Hive修改元数据的相关命令操作datasource表和Spark on HBase表。

12.23.8.2.14 spark-sql 退出时打印 RejectedExecutionException 异常栈

问题

执行大数据量的Spark任务（如2T的TPCDS测试套），任务运行成功后，在spark-sql退出时概率性出现RejectedExecutionException的异常栈信息，相关日志如下所示：

```
16/07/16 10:19:56 ERROR TransportResponseHandler: Still have 2 requests outstanding when connection from linux-192/10.1.1.5:59250 is closed
java.util.concurrent.RejectedExecutionException: Task scala.concurrent.impl.CallbackRunnable@5fc1ab rejected from java.util.concurrent.ThreadPoolExecutor@52fa7e19[Terminated, pool size = 0, active threads = 0, queued tasks = 0, completed tasks = 3025]
```

回答

出现上述问题的原因是：当spark-sql退出时，应用退出关闭消息通道，如果当前还有消息未处理，需要做连接关闭异常的处理，此时，如果scala内部的线程池已经关闭，就会打印RejectEdExecutionException的异常栈，如果scala内部的线程池尚未关闭就不会打印该异常栈。

因为该问题出现在应用退出时，此时任务已经运行成功，所以不会对业务产生影响。

12.23.8.2.15 健康检查时，误将 JDBCServer Kill

问题

健康检查方案中，在并发执行的语句达到线程池上限后依然会导致健康检查命令无法执行，从而导致健康检查程序超时，然后把Spark JDBCServer进程Kill。

回答

当前JDBCServer中存在两个线程池HiveServer2-Handler-Pool和HiveServer2-Background-Pool，其中HiveServer2-Handler-Pool用于处理session连接，HiveServer2-Background-Pool用于处理SQL语句的执行。

当前的健康检查机制是通过新增一个session连接，并在该session所在的线程中执行健康检查命令***HEALTHCHECK***来判断Spark JDBCServer的健康状况，因此HiveServer2-Handler-Pool必须保留一个线程，用于处理健康检查的session连接和健康检查命令执行，否则将导致无法建立健康检查的session连接或健康检查命令无法执行，从而认为Spark JDBCServer不健康而被Kill。即如果当前HiveServer2-Handler-Pool的线程池数为100，那么最多支持连接99个session。

12.23.8.2.16 日期类型的字段作为过滤条件时匹配'2016-6-30'时没有查询结果

问题

为什么日期类型的字段作为过滤条件时匹配'2016-6-30'时没有查询结果，匹配'2016-06-30'时有查询结果。

如下图所示：“select count(*)from trxfintrx2012 a where trx_dte_par='2016-6-30'”，其中trx_dte_par为日期类型的字段，当过滤条件为“where trx_dte_par='2016-6-30'”时没有查询结果，当过滤条件为“where trx_dte_par='2016-06-30'”时有查询结果。

图 12-61 示例

```
0: jdbc:hive2://ha-cluster/default> select count(*)
0: jdbc:hive2://ha-cluster/default>   from TRXFINTRX2012 a
0: jdbc:hive2://ha-cluster/default>   where trx_dte_par = '2016-6-30';
+------+--+
| _c0  |
+------+--+
| 0    |
+------+--+
1 row selected (0.498 seconds)
0: jdbc:hive2://ha-cluster/default> select count(*)
0: jdbc:hive2://ha-cluster/default>   from TRXFINTRX2012 a
0: jdbc:hive2://ha-cluster/default>   where trx_dte_par = '2016-06-30';
+----------+--+
|   _c0    |
+----------+--+
| 8520808  |
+----------+--+
1 row selected (15.788 seconds)
```

回答

在Spark SQL查询语句中，当查询条件中含有日期格式的字符串时，Spark SQL不会对它做日期格式的检查，就是把它当做普通的字符串进行匹配。以上面的例子为例，如果数据格式为"yyyy-mm-dd"，那么字符串'2016-6-30'就是不正确的数据格式。

12.23.8.2.17 为什么在启动 spark-beeline 的命令中指定“--hivevar”选项无效

问题

为什么在启动spark-beeline的命令中指定“--hivevar”选项无效?

从V100R002C60版本开始，在启动spark-beeline的命令中如果使用了“--hivevar <VAR_NAME>=<var_value>”选项自定义一个变量，在启动spark-beeline时不会报错，但在SQL语句中用到变量<VAR_NAME>时会报无法解析<VAR_NAME>的错误。

举例说明，场景如下：

1. 执行以下命令启动spark-beeline：

   spark-beeline --hivevar <VAR_NAME>=<var_value>

2. 启动成功后，在spark-beeline中执行SQL语句，如“DROP TABLE ${VAR_NAME}”，报无法解析VAR_NAME的错误。

回答

从V100R002C60版本开始，因新增多session管理功能，Hive的特性“--hivevar <VAR_NAME>=<var_value>”在Spark中已不再支持，因此在spark-beeline的启动命令中使用“--hivevar”选项无效。

12.23.8.2.18 在 spark-beeline 中创建临时表/视图时，报 HDFS 目录无权限操作的错误

问题

在普通模式下，用户在spark-beeline中创建临时表或创建视图时，报“Permission denied”的错误，这个错误表明HDFS目录无权限操作。错误日志信息如下：

```
org.apache.hadoop.security.AccessControlException Permission denied: user=root, access=EXECUTE,
inode="/tmp/spark/sparkhive-scratch/omm/e579a76f-43ed-4014-8a54-1072c07ceeff/_tmp_space.db/
52db1561-60b0-4e7d-8a25-c2eaa44850a9":omm:hadoop:drwx------
```

回答

在普通模式下，当使用非omm用户（例如root用户）执行启动spark-beeline的命令时，在未指定“-n”时用户为root，而启动spark-beeline后，JDBCServer会创建一个HDFS新目录，由于当前版本启动JDBCServer的用户是omm，而在DataSightV100R002C30以前的版本是root用户，因此当前该HDFS目录的owner为omm、group为hadoop。在spark-beeline中创建临时表或视图时会使用该HDFS目录，此时是root用户，但是root用户在HDFS中是一个普通用户，因此没有权限操作omm用户的HDFS目录，从而报“Permission denied”的错误。

综上所述，在普通模式下，只有omm用户可以创建临时表或视图，如果用户需要创建临时表或视图，可通过在启动spark-beeline时带“-n omm”选项指定操作用户为omm，这样便有权限操作成功。

12.23.8.2.19 执行复杂 SQL 语句时报“Code of method ... grows beyond 64 KB”的错误

问题

当执行一个很复杂的SQL语句时，例如有多层语句嵌套，且单层语句中对字段有大量的逻辑处理（如多层嵌套的case when语句），此时执行该语句会报如下所示的错误日志，该错误表明某个方法的代码超出了64KB。

```
java.util.concurrent.ExecutionException: java.lang.Exception: failed to compile:
org.codehaus.janino.JaninoRuntimeException: Code of method "(Lorg/apache/spark/sql/catalyst/expressions/
GeneratedClass$SpecificUnsafeProjection;Lorg/apache/spark/sql/catalyst/InternalRow;)V" of class
"org.apache.spark.sql.catalyst.expressions.GeneratedClass$SpecificUnsafeProjection" grows beyond 64 KB
```

回答

在开启钨丝计划（即tungsten功能）后，Spark对于部分执行计划会使用codegen的方式来生成Java代码，但JDK编译时要求Java代码中的每个函数的长度不能超过64KB。当执行一个很复杂的SQL语句时，例如有多层语句嵌套，且单层语句中对字段有大量的逻辑处理（如多层嵌套的case when语句），这种情况下，通过codegen生成的Java代码中函数的大小就可能会超过64KB，从而导致编译失败。

规避措施：

当出现上述问题时，用户可以通过关闭钨丝计划，关闭使用codegen的方式来生成Java代码的功能，从而确保语句的正常执行。即在客户端的“spark-defaults.conf”配置文件中将“spark.sql.codegen.wholeStage”配置为“false”。

12.23.8.2.20 在 Beeline/JDBCServer 模式下连续运行 10T 的 TPCDS 测试套会出现内存不足的现象

问题

在Driver内存配置为10G时，Beeline/JDBCServer模式下连续运行10T的TPCDS测试套，会出现因为Driver内存不足导致SQL语句执行失败的现象。

回答

当前在默认配置下，在内存中保留的Job和Stage的UI数据个数为1000个。

当前大集群优化已增加将UI数据溢出到磁盘的优化，其溢出条件是每个Stage中的UI数据大小达到最小阈值5MB。如果每个Stage的task数较小，那么其UI数据大小可能达不到该阈值，从而导致该Stage的UI数据一直缓存在内存中，直到UI数据个数到达保留的上限值（当前默认值为1000个），旧的UI数据才会在内存中被清除。

因此，在将旧的UI数据从内存中清除之前，UI数据会占用大量内存，从而导致执行10T的TPCDS测试套时出现Driver内存不足的现象。

规避措施：

- 根据业务需要，配置合适的需要保留的Job和Stage的UI数据个数，即配置“spark.ui.retainedJobs”和“spark.ui.retainedStages”参数。详细信息请参考**常用参数**中的**表12-359**。
- 如果需要保留的Job和Stage的UI数据个数较多，可通过配置“spark.driver.memory”参数，适当增大Driver的内存。详细信息请参考**常用参数**中的**表12-356**。

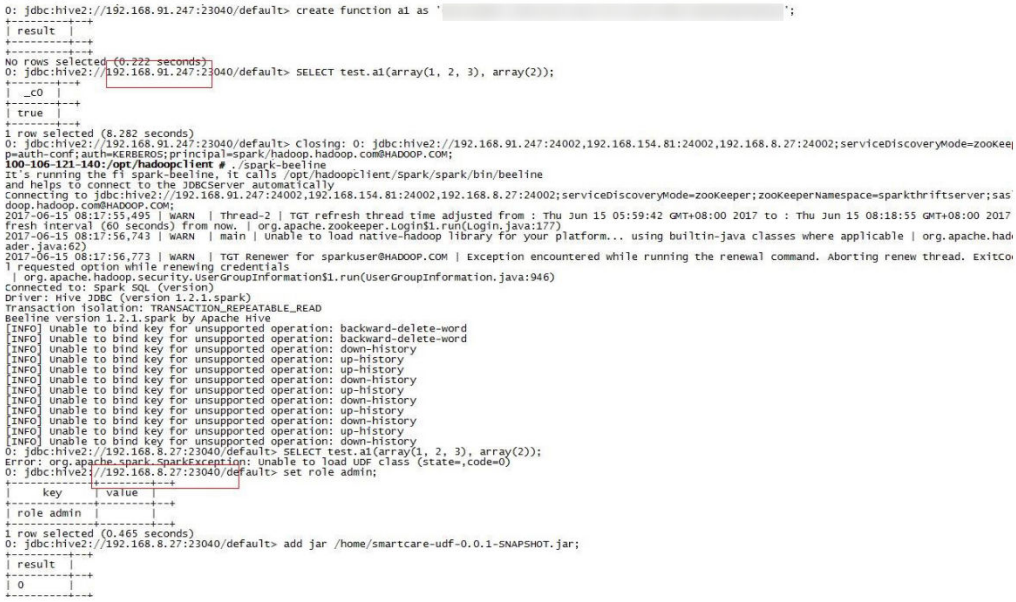
12.23.8.2.21 连上不同的 JDBCServer，function 不能正常使用

问题

场景一：

通过add jar的方式建立永久函数，当Beeline连上不同的JDBCServer或者JDBCServer重启后都需要重新add jar。

图 12-62 场景一异常信息



场景二：

show functions能够查到相应的函数，但是无法使用，这是由于连接上的JDBC节点上没有相应路径的jar包，添加上相应的jar包能够查询成功。

图 12-63 场景二异常信息

回答

场景一：

add jar语句只会将jar加载到当前连接的JDBCServer的jarClassLoader，不同JDBCServer不会共用。JDBCServer重启后会创建新的jarClassLoader，所以需要重新add jar。

添加jar包有两种方式：可以在启动spark-sql的时候添加jar包，如**spark-sql --jars /opt/test/two_udfs.jar**；也可在spark-sql启动后再添加jar包，如**add jar /opt/test/two_udfs.jar**。add jar所指定的路径可以是本地路径也可以是HDFS上的路径。

场景二：

show functions会从外部的Catalog获取当前database中所有的function。SQL中使用function时，JDBCServer会加载该function对应的jar。

若jar不存在，则该function无法使用，需要重新执行***add jar***命令。

12.23.8.2.22 Spark2x 无法访问 Spark1.5 创建的 DataSource 表

问题

在Spark2x中访问Spark1.5创建的DataSource表时，报无法获取schema信息，导致无法访问表。

回答

- 原因分析：

  这是由于Spark2x与Spark1.5存储DataSoure表信息的格式不一致导致的。Spark1.5会将schema信息分成多个part，使用path.park.0作为key进行存储，读取时再将各个part都读取出来，重新拼成完整的信息。而Spark2x直接使用相应的key获取对应的信息。这样在Spark2x中去读取Spark1.5创建的DataSource表时，就无法成功读取到key对应的信息，导致解析DataSource表信息失败。

  而在处理Hive格式的表时，Spark2x与Spark1.5的存储方式一致，所以Spark2x可以直接读取Spark1.5创建的表，不存在上述问题。

- 规避措施：

  Spark2x可以通过创建外表的方式来创建一张指向Spark1.5表实际数据的表，这样可以实现在Spark2x中读取Spark1.5创建的DataSource表。同时，Spark1.5更新过数据后，Spark2x中访问也能感知到变化 ，反过来一样。这样即可实现Spark2x对Spark1.5创建的DataSource表的访问。

12.23.8.2.23 为什么 spark-beeline 运行失败报“Failed to create ThriftService instance”的错误

问题

为什么spark-beeline运行失败报“Failed to create ThriftService instance”的错误?

Beeline日志如下所示：

```
Error: Failed to create ThriftService instance (state=,code=0)
Beeline version 1.2.1.spark by Apache Hive
[INFO] Unable to bind key for unsupported operation: backward-delete-word
[INFO] Unable to bind key for unsupported operation: backward-delete-word
[INFO] Unable to bind key for unsupported operation: down-history
[INFO] Unable to bind key for unsupported operation: up-history
[INFO] Unable to bind key for unsupported operation: up-history
[INFO] Unable to bind key for unsupported operation: down-history
[INFO] Unable to bind key for unsupported operation: up-history
[INFO] Unable to bind key for unsupported operation: down-history
[INFO] Unable to bind key for unsupported operation: up-history
[INFO] Unable to bind key for unsupported operation: down-history
[INFO] Unable to bind key for unsupported operation: up-history
```

```
[INFO] Unable to bind key for unsupported operation: down-history
beeline>
```

同时，在JDBCServer端出现“Timed out waiting for client to connect”的错误日志，关键日志如下所示：

```
2017-07-12 17:35:11,284 | INFO  | [main] | Will try to open client transport with JDBC Uri:
jdbc:hive2://192.168.101.97:23040/default;principal=spark/hadoop.<系统域名>@<系统域名
>;healthcheck=true;saslQop=auth-conf;auth=KERBEROS;user.principal=spark/hadoop.<系统域名>@<系统域名
>;user.keytab=${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-Spark-3.1.1/keytab/spark/
JDBCServer/spark.keytab | org.apache.hive.jdbc.HiveConnection.openTransport(HiveConnection.java:317)
2017-07-12 17:35:11,326 | INFO  | [HiveServer2-Handler-Pool: Thread-92] | Client protocol version:
HIVE_CLI_SERVICE_PROTOCOL_V8 |
org.apache.proxy.service.ThriftCLIProxyService.OpenSession(ThriftCLIProxyService.java:554)
2017-07-12 17:35:49,790 | ERROR | [HiveServer2-Handler-Pool: Thread-113] | Timed out waiting for client
to connect.
Possible reasons include network issues, errors in remote driver or the cluster has no available resources, etc.
Please check YARN or Spark driver's logs for further information. |
org.apache.proxy.service.client.SparkClientImpl.<init>(SparkClientImpl.java:90)
java.util.concurrent.ExecutionException: java.util.concurrent.TimeoutException: Timed out waiting for
client connection.
 at io.netty.util.concurrent.AbstractFuture.get(AbstractFuture.java:37)
 at org.apache.proxy.service.client.SparkClientImpl.<init>(SparkClientImpl.java:87)
 at org.apache.proxy.service.client.SparkClientFactory.createClient(SparkClientFactory.java:79)
 at org.apache.proxy.service.SparkClientManager.createSparkClient(SparkClientManager.java:145)
 at org.apache.proxy.service.SparkClientManager.createThriftServerInstance(SparkClientManager.java:160)
 at org.apache.proxy.service.ThriftServiceManager.getOrCreateThriftServer(ThriftServiceManager.java:182)
 at org.apache.proxy.service.ThriftCLIProxyService.OpenSession(ThriftCLIProxyService.java:596)
 at org.apache.hive.service.cli.thrift.TCLIService$Processor$OpenSession.getResult(TCLIService.java:1257)
 at org.apache.hive.service.cli.thrift.TCLIService$Processor$OpenSession.getResult(TCLIService.java:1242)
 at org.apache.thrift.ProcessFunction.process(ProcessFunction.java:39)
 at org.apache.thrift.TBaseProcessor.process(TBaseProcessor.java:39)
 at org.apache.hadoop.hive.thrift.HadoopThriftAuthBridge$Server
$TUGIAssumingProcessor.process(HadoopThriftAuthBridge.java:696)
 at org.apache.thrift.server.TThreadPoolServer$WorkerProcess.run(TThreadPoolServer.java:286)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
 at java.lang.Thread.run(Thread.java:748)
Caused by: java.util.concurrent.TimeoutException: Timed out waiting for client connection.
```

回答

当网络不稳定时，会出现上述问题。当beeline出现timed-out异常时，Spark不会尝试重连。相反，用户需要通过重新启动spark-beeline进行重连。

12.23.8.2.24 Spark SQL 无法查询到 ORC 类型的 Hive 表的新插入数据

问题

为什么通过Spark SQL无法查询到存储类型为ORC的Hive表的新插入数据？主要有以下两种场景存在这个问题：

- 对于分区表和非分区表，在Hive客户端中执行插入数据的操作后，会出现Spark SQL无法查询到最新插入的数据的问题。
- 对于分区表，在Spark SQL中执行插入数据的操作后，如果分区信息未改变，会出现Spark SQL无法查询到最新插入的数据的问题。

回答

由于Spark存在一个机制，为了提高性能会缓存ORC的元数据信息。当通过Hive或其他方式更新了ORC表时，缓存的元数据信息未更新，导致Spark SQL查询不到新插入的数据。

对于存储类型为ORC的Hive分区表，在执行插入数据操作后，如果分区信息未改变，则缓存的元数据信息未更新，导致Spark SQL查询不到新插入的数据。

解决措施：

1. 在使用Spark SQL查询之前，需执行Refresh操作更新元数据信息：

   REFRESH TABLE *table_name*;

   table_name为刷新的表名，该表必须存在，否则会出错。

   执行查询语句时，即可获取到最新插入的数据。
2. 使用sqark时，执行以下命令禁用Spark优化：

   set spark.sql.hive.convertMetastoreOrc=false;

12.23.8.3 Spark Streaming

12.23.8.3.1 Spark Streaming 任务一直阻塞

问题

运行一个Spark Streaming任务，确认有数据输入后，发现没有任何处理的结果。打开Web界面查看Spark Job执行情况，发现如下图所示：有两个Job一直在等待运行，但一直无法成功运行。

图 12-64 Active Jobs

Active Jobs (2)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total
3	print at test2StreamFromKafka.scala:31	2015/05/25 18:28:55	63.7 h	0/3
2	start at test2StreamFromKafka.scala:34	2015/05/25 18:28:55	63.7 h	0/1

继续查看已经完成的Job，发现也只有两个，说明Spark Streaming都没有触发数据计算的任务（ Spark Streaming默认有两个尝试运行的Job，就是图中两个 ）

图 12-65 Completed Jobs

Completed Jobs (2)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total
1	print at test2StreamFromKafka.scala:31	2015/05/25 18:28:55	0.7 s	2/2 (1 skipped)
0	start at test2StreamFromKafka.scala:34	2015/05/25 18:28:54	1 s	2/2

回答

经过定位发现，导致这个问题的原因是：Spark Streaming的计算核数少于Receiver的个数，导致部分Receiver启动以后，系统已经没人资源去运行计算任务，导致第一个任务一直在等待，后续任务一直在排队。从现象上看，就是如问题中的图12-64中所示，会有两个任务一直在等待。

因此，当Web出现两个任务一直在等待的情况，首先检查Spark的核数是否大于Receiver的个数。

说明

Receiver在Spark Streaming中是一个常驻的Spark Job，Receiver对于Spark是一个普通的任务，但它的生命周期和Spark Streaming任务相同，并且占用一个核的计算资源。

在调试和测试等经常使用默认配置的场景下，要时刻注意核数与Receiver个数的关系。

12.23.8.3.2 运行 Spark Streaming 任务参数调优的注意事项

问题

运行Spark Streaming任务时，随着executor个数的增长，数据处理性能没有明显提升，对于参数调优有哪些注意事项?

回答

在executor核数等于1的情况下，遵循以下规则对调优Spark Streaming运行参数有所帮助。

- Spark任务处理速度和Kafka上partition个数有关，当partition个数小于给定executor个数时，实际使用的executor个数和partition个数相同，其余的将会被空闲。所以应该使得executor个数小于或者等于partition个数。
- 当Kafka上不同partition数据有倾斜时，数据较多的partition对应的executor将成为数据处理的瓶颈，所以在执行Producer程序时，数据平均发送到每个partition可以提升处理的速度。
- 在partition数据均匀分布的情况下，同时提高partition和executor个数，将会提升Spark处理速度（当partition个数和executor个数保持一致时，处理速度是最快的）。
- 在partition数据均匀分布的情况下，尽量保持partition个数是executor个数的整数倍，这样将会使资源得到合理利用。

12.23.8.3.3 为什么提交 Spark Streaming 应用超过 token 有效期，应用失败

问题

修改kerberos的票据和HDFS token过期时间为5分钟，设置“dfs.namenode.delegation.token.renew-interval”小于60秒，提交Spark Streaming应用，超过token有效期，提示以下错误，应用失败。

```
token (HDFS_DELEGATION_TOKEN token 17410 for spark2x) is expired
```

回答

- 问题原因：

  ApplicationMaster进程中有1个Credential Refresh Thread会根据*token renew周期 * 0.75*的时间比例上传更新后的Credential文件到HDFS上。

  Executor进程中有1个Credential Refresh Thread会根据*token renew周期 *0.8*的时间比例去HDFS上获取更新后的Credential文件，用来刷新UserGroupInformation中的token，避免token失效。

  当Executor进程的Credential Refresh Thread发现当前时间已经超过Credential文件更新时间（即*token renew周期 *0.8*）时，会等待1分钟再去HDFS上面获取最新的Credential文件，以确保AM端已经将更新后的Credential文件放到HDFS上。

当“dfs.namenode.delegation.token.renew-interval”配置值小于60秒，Executor进程起来时发现当前时间已经超过Credential文件更新时间，等待1分钟再去HDFS上面获取最新的Credential文件，而此时token已经失效，task运行失败，然后在其他Executor上重试，由于重试时间都是在1分钟内完成，所以task在其他Executor上也运行失败，导致运行失败的Executor加入到黑名单，没有可用的Executor，应用退出。

- 修改方案：

  在Spark使用场景下，需设置“dfs.namenode.delegation.token.renew-interval”大于80秒。“dfs.namenode.delegation.token.renew-interval”参数描述请参表12-428考。

  表 12-428 参数说明

参数	描述	默认值
dfs.namenode.delegation.token.renew-interval	该参数为服务器端参数，设置token renew的时间间隔，单位为毫秒。	86400000

12.23.8.3.4 为什么 Spark Streaming 应用创建输入流，但该输入流无输出逻辑时，应用从checkpoint 恢复启动失败

问题

Spark Streaming应用创建1个输入流，但该输入流无输出逻辑。应用从checkpoint恢复启动失败，报错如下：

```
17/04/24 10:13:57 ERROR Utils: Exception encountered
java.lang.NullPointerException
at org.apache.spark.streaming.dstream.DStreamCheckpointData$$anonfun$writeObject$1.apply$mcV
$sp(DStreamCheckpointData.scala:125)
at org.apache.spark.streaming.dstream.DStreamCheckpointData$$anonfun$writeObject
$1.apply(DStreamCheckpointData.scala:123)
at org.apache.spark.streaming.dstream.DStreamCheckpointData$$anonfun$writeObject
$1.apply(DStreamCheckpointData.scala:123)
at org.apache.spark.util.Utils$.tryOrIOException(Utils.scala:1195)
at
org.apache.spark.streaming.dstream.DStreamCheckpointData.writeObject(DStreamCheckpointData.scala:123
)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at java.io.ObjectStreamClass.invokeWriteObject(ObjectStreamClass.java:1028)
at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1496)
at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1432
at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1178)
at java.io.ObjectOutputStream.defaultWriteFields(ObjectOutputStream.java:1548)
at java.io.ObjectOutputStream.defaultWriteObject(ObjectOutputStream.java:441)
at org.apache.spark.streaming.dstream.DStream$$anonfun$writeObject$1.apply$mcV$sp(DStream.scala:515)
at org.apache.spark.streaming.dstream.DStream$$anonfun$writeObject$1.apply(DStream.scala:510)
at org.apache.spark.streaming.dstream.DStream$$anonfun$writeObject$1.apply(DStream.scala:510)
at org.apache.spark.util.Utils$.tryOrIOException(Utils.scala:1195)
at org.apache.spark.streaming.dstream.DStream.writeObject(DStream.scala:510)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
```

```
at java.io.ObjectStreamClass.invokeWriteObject(ObjectStreamClass.java:1028)
at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1496)
at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1432
at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1178)
at java.io.ObjectOutputStream.writeArray(ObjectOutputStream.java:1378)
at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1174)
at java.io.ObjectOutputStream.defaultWriteFields(ObjectOutputStream.java:1548)
at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1509)
at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1432
at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1178)
at java.io.ObjectOutputStream.defaultWriteFields(ObjectOutputStream.java:1548)
at java.io.ObjectOutputStream.defaultWriteObject(ObjectOutputStream.java:441)
at org.apache.spark.streaming.DStreamGraph$$anonfun$writeObject$1.apply$mcV
$sp(DStreamGraph.scala:191)
at org.apache.spark.streaming.DStreamGraph$$anonfun$writeObject$1.apply(DStreamGraph.scala:186)
at org.apache.spark.streaming.DStreamGraph$$anonfun$writeObject$1.apply(DStreamGraph.scala:186)
at org.apache.spark.util.Utils$.tryOrIOException(Utils.scala:1195)
at org.apache.spark.streaming.DStreamGraph.writeObject(DStreamGraph.scala:186
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at java.io.ObjectStreamClass.invokeWriteObject(ObjectStreamClass.java:1028)
at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1496)
at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1432
at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1178)
at java.io.ObjectOutputStream.defaultWriteFields(ObjectOutputStream.java:1548)
at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1509)
at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1432
at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1178)
at java.io.ObjectOutputStream.writeObject(ObjectOutputStream.java:348)
at org.apache.spark.streaming.Checkpoint$$anonfun$serialize$1.apply$mcV$sp(Checkpoint.scala:142)
at org.apache.spark.streaming.Checkpoint$$anonfun$serialize$1.apply(Checkpoint.scala:142)
at org.apache.spark.streaming.Checkpoint$$anonfun$serialize$1.apply(Checkpoint.scala:142)
at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1230)
at org.apache.spark.streaming.Checkpoint$.serialize(Checkpoint.scala:143)
at org.apache.spark.streaming.StreamingContext.validate(StreamingContext.scala:566)
at org.apache.spark.streaming.StreamingContext.liftedTree1$1(StreamingContext.scala:612)
at org.apache.spark.streaming.StreamingContext.start(StreamingContext.scala:611)
at com.spark.test.kafka08LifoTwoInkfk$.main(kafka08LifoTwoInkfk.scala:21)
at com.spark.test.kafka08LifoTwoInkfk.main(kafka08LifoTwoInkfk.scala)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$
$runMain(SparkSubmit.scala:772)
at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:183)
at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:208)
at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:123)
at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
```

回答

Streaming Context启动时，若应用设置了checkpoint，则需要对应用中的DStream checkpoint对象进行序列化，序列化时会用到dstream.context。

dstream.context是Streaming Context启动时从output Streams反向查找所依赖的DStream，逐个设置context。若Spark Streaming应用创建1个输入流，但该输入流无输出逻辑时，则不会给它设置context。所以在序列化时报“NullPointerException”。

解决办法：应用中如果有无输出逻辑的输入流，则在代码中删除该输入流，或添加该输入流的相关输出逻辑。

12.23.8.3.5 Spark Streaming 应用运行过程中重启 Kafka，Web UI 界面部分 batch time 对应 Input Size 为 0 records

问题

在Spark Streaming应用执行过程中重启Kafka时，应用无法从Kafka获取topic offset，从而导致生成Job失败。如图12-66所示，其中2017/05/11 10:57:00~2017/05/11 10:58:00为Kafka重启时间段。2017/05/11 10:58:00重启成功后对应的“Input Size”的值显示为“0 records”。

图 12-66 Web UI 界面部分 batch time 对应 Input Size 为 0 records

Completed Batches (last 9 out of 9)

Batch Time	Input Size	Scheduling Delay (?)	Processing Time (?)	Total Delay (?)	Output Ops: Succeeded/Total
2017/05/11 10:58:50	18 records	0 ms	0.4 s	0.4 s	1/1
2017/05/11 10:58:40	20 records	4 s	0.3 s	4 s	1/1
2017/05/11 10:58:30	20 records	14 s	0.5 s	14 s	1/1
2017/05/11 10:58:20	20 records	23 s	0.4 s	24 s	1/1
2017/05/11 10:58:10	20 records	33 s	0.5 s	33 s	1/1
2017/05/11 10:58:00	0 records	6 ms	43 s	43 s	1/1
2017/05/11 10:57:00	19 records	1 ms	0.9 s	0.9 s	1/1
2017/05/11 10:56:50	20 records	1 ms	0.6 s	0.6 s	1/1
2017/05/11 10:56:40	28 records	13 ms	5 s	5 s	1/1

回答

Kafka重启成功后应用会按照batch时间把2017/05/11 10:57:00~2017/05/11 10:58:00缺失的RDD补上，尽管UI界面上显示读取的数据个数为“0”，但实际上这部分数据在补的RDD中进行了处理，因此，不存在数据丢失。

Kafka重启时间段的数据处理机制如下。

Spark Streaming应用使用了state函数（例如：updateStateByKey），在Kafka重启成功后，Spark Streaming应用生成2017/05/11 10:58:00 batch任务时，会按照batch时间把2017/05/11 10:57:00~2017/05/11 10:58:00缺失的RDD补上（Kafka重启前Kafka上未读取完的数据，属于2017/05/11 10:57:00之前的batch）。

12.23.8.4 访问 Spark 应用获取的 restful 接口信息有误

问题

当Spark应用结束后，访问该应用的restful接口获取job信息，发现job信息中“numActiveTasks”的值是负数，如图12-67所示。

图 12-67 job 信息


```
[ {
  "jobId" : 0,
  "name" : "reduce at SparkPi.scala:36",
  "submissionTime" : "2016-05-28T09:35:34.415GMT",
  "completionTime" : "2016-05-28T09:35:35.686GMT",
  "stageIds" : [ 0 ],
  "status" : "SUCCEEDED",
  "numTasks" : 2,
  "numActiveTasks" : -1,
  "numCompletedTasks" : 2,
  "numSkippedTasks" : 2,
  "numFailedTasks" : 0,
  "numActiveStages" : 0,
  "numCompletedStages" : 1,
  "numSkippedStages" : 0,
  "numFailedStages" : 0
} ]
```


说明

numActiveTasks是指当前正在运行task的个数。

回答

通过下面两种途径获取上面的job信息：

- 配置**spark.history.briefInfo.gather=true**，查看JobHistory的brief信息。
- 使用Spark JobHistory2x页面访问：https://IP:port/api/v1/<appid>/jobs/。

job信息中“numActiveTasks”的值是根据eventlog文件中SparkListenerTaskStart和SparkListenerTaskEnd事件的个数的差值计算得到的。如果eventLog文件中有事件丢失，就可能出现上面的现象。

12.23.8.5 为什么从 Yarn Web UI 页面无法跳转到 Spark Web UI 界面

问题

FusionInsight版本中，在客户端采用yarn-client模式运行Spark应用，然后从Yarn的页面打开该应用的Web UI界面，出现下面的错误：

Error Occurred.

Problem accessing /proxy/application_1468986660719_0045/

Powered by Jetty://

从YARN ResourceManager的日志看到：

2016-07-21 16:35:27,099 | INFO | Socket Reader #1 for port 8032 | Auth successful for mapred/hadoop.*<系统域名>*@*<系统域名>* (auth:KERBEROS) | Server.java:1388
2016-07-21 16:35:27,105 | INFO | 1526016381@qtp-1178290888-1015 | admin is accessing unchecked http://10.120.169.53:23011 which is the app master GUI of application_1468986660719_0045 owned by spark | WebAppProxyServlet.java:393
2016-07-21 16:36:02,843 | INFO | Socket Reader #1 for port 8032 | Auth successful for hive/hadoop.*<系统域名>*@*<系统域名>* (auth:KERBEROS) | Server.java:1388

```
2016-07-21 16:36:02,851 | INFO  | Socket Reader #1 for port 8032 | Auth successful for hive/hadoop.<系统域名>@<系统域名> (auth:KERBEROS) | Server.java:1388
2016-07-21 16:36:12,163 | WARN  | 1526016381@qtp-1178290888-1015 | /proxy/application_1468986660719_0045/: java.net.ConnectException: Connection timed out | Slf4jLog.java:76
2016-07-21 16:37:03,918 | INFO  | Socket Reader #1 for port 8032 | Auth successful for hive/hadoop.<系统域名>@<系统域名> (auth:KERBEROS) | Server.java:1388
2016-07-21 16:37:03,926 | INFO  | Socket Reader #1 for port 8032 | Auth successful for hive/hadoop.<系统域名>@<系统域名> (auth:KERBEROS) | Server.java:1388
2016-07-21 16:37:11,956 | INFO  | AsyncDispatcher event handler | Updating application attempt appattempt_1468986660719_0045_000001 with final state: FINISHING,
and exit status: -1000 | RMAppAttemptImpl.java:1253
```

回答

打开FusionInsight Manager页面，看到Yarn服务的业务IP地址为192网段。

从Yarn的日志看到，Yarn读取的Spark Web UI地址为http://10.120.169.53:23011，是10网段的IP地址。由于192网段的IP和10网段的IP不能互通，所以导致访问Spark Web UI界面失败。

修改方案：

登录10.120.169.53客户端机器，修改/etc/hosts文件，将10.120.169.53更改为相对应的192网段的IP地址。再重新运行Spark应用，这时就可以打开Spark Web UI界面。

12.23.8.6 HistoryServer 缓存的应用被回收，导致此类应用页面访问时出错

问题

在History Server页面中访问某个Spark应用的页面时，发现访问时出错。

查看相应的HistoryServer日志后，发现有“FileNotFound”异常，相关日志如下所示：

```
2016-11-22 23:58:03,694 | WARN  | [qtp55429210-232] | /history/application_1479662594976_0001/stages/stage/ | org.sparkproject.jetty.servlet.ServletHandler.doHandle(ServletHandler.java:628)
java.io.FileNotFoundException: ${BIGDATA_HOME}/tmp/spark/jobHistoryTemp/blockmgr-5f1f6aca-2303-4290-9845-88fa94d78480/09/temp_shuffle_11f82aaf-e226-46dc-b1f0-002751557694 (No such file or directory)
```

回答

在History Server页面加载Task个数较多的Spark应用时，由于无法把全部的数据放入内存中，导致数据溢出到磁盘时，会产生前缀为“temp_shuffle”的文件。

HistoryServer默认会缓存50个Spark应用（由配置项“spark.history.retainedApplications”决定），当内存中的Spark应用个数超过这个数值时，HistoryServer会回收最先缓存的Spark应用，同时会清理掉相应的“temp_shuffle”文件。

当用户正在查看即将被回收的Spark应用时，可能会出现找不到“temp_shuffle”文件的错误，从而导致当前页面无法访问。

如果遇到上述问题，可参考以下两种方法解决。

- 重新访问这个Spark应用的HistoryServer页面，即可查看到正确的页面信息。
- 如果用户场景需要同时访问50个以上的Spark应用时，需要调大“spark.history.retainedApplications”参数的值。

请登录FusionInsight Manager管理界面，单击“集群 > *待操作集群的名称* > 服务 > Spark2x > 配置”，单击“全部配置”，在左侧的导航列表中，单击“JobHistory2x > 界面”，配置如下参数。

表 12-429 参数说明

参数	描述	默认值
spark.history.retainedApplications	HistoryServer缓存的Spark应用数，当需要缓存的应用个数超过此参数值时，HistoryServer会回收最先缓存的Spark应用。	50

12.23.8.7 加载空的 part 文件时，app 无法显示在 JobHistory 的页面上

问题

在分组模式下执行应用，当HDFS上的part文件为空时，发现JobHistory首页面上不显示该part对应的app。

回答

JobHistory服务更新页面上的app时，会根据HDFS上的part文件大小变更与否判断是否刷新首页面的app显示信息。若文件为第一次查看，则将当前文件大小与0作比较，如果大于0则读取该文件。

分组的情况下，如果执行的app没有job处于执行状态，则part文件为空，即JobHistory服务不会读取该文件，此app也不会显示在JobHistory页面上。但若part文件大小之后有更新，JobHistory又会显示该app。

12.23.8.8 Spark2x 导出带有相同字段名的表，结果导出失败

问题

在Spark2x的spark-shell上执行如下代码失败：

```
val acctId = List(("49562", "Amal", "Derry"), ("00000", "Fred", "Xanadu"))
val rddLeft = sc.makeRDD(acctId)
val dfLeft = rddLeft.toDF("Id", "Name", "City")
//dfLeft.show
val acctCustId = List(("Amal", "49562", "CO"), ("Dave", "99999", "ZZ"))
val rddRight = sc.makeRDD(acctCustId)
val dfRight = rddRight.toDF("Name", "CustId", "State")
//dfRight.show
val dfJoin = dfLeft.join(dfRight, dfLeft("Id") === dfRight("CustId"), "outer")
dfJoin.show
dfJoin.repartition(1).write.format("com.databricks.spark.csv").option("delimiter", "\t").option("header",
"true").option("treatEmptyValuesAsNulls", "true").option("nullValue", "").save("/tmp/outputDir")
```

回答

Spark2x中对join语句重名字段做了判断，需要修改代码保证保存的数据中无重复字段。

12.23.8.9 为什么多次运行 Spark 应用程序会引发致命 JRE 错误

问题

为什么多次运行Spark应用程序会引发致命JRE错误?

回答

多次运行Spark应用程序会引发致命的JRE错误，这个错误由Linux内核导致。

升级内核版本到4.13.9-2.ge7d7106-default来解决这个问题。

12.23.8.10 IE 浏览器访问 Spark2x 原生 UI 界面失败，无法显示此页或者页面显示错误

问题

通过IE 9、IE 10和IE 11浏览器访问Spark2x的原生UI界面，出现访问失败情况或者页面显示错误问题。

现象

访问页面失败，浏览器无法显示此页，如下图所示：

原因

IE 9、IE 10、IE 11浏览器的某些版本在处理SSL握手有问题导致访问失败。

解决方法

推荐使用Google Chrome浏览器71及其以上版本和Firefox浏览器62及其以上版本。

12.23.8.11 Spark2x 如何访问外部集群组件

问题

存在两个集群：cluster1 和cluster2，如何使用cluster1中的Spark2x访问cluster2中的HDFS、Hive、HBase和Kafka组件。

回答

1. 可以有条件的实现两个集群间组件互相访问，但是存在以下限制：
   - 仅允许访问一个Hive MetaStore，不支持同时访问cluster1的Hive MetaStore和cluster2的Hive MetaStore。

- 不同集群的用户系统没有同步，因此访问跨集群组件时，用户的权限管理由对端集群的用户配置决定。比如cluster1的userA没有访问本集群HBase meta表权限，但是cluster2的userA有访问该集群HBase meta表权限，则cluster1的userA可以访问cluster2的HBase meta表。
- 跨Manager之间的安全集群间组件互相访问，需要先配置系统互信。

2. 以下分别阐述cluster1上使用userA访问cluster2的Hive、HBase、Kafka组件。

   说明

   以下操作皆以用户使用FusionInsight客户端提交Spark2x应用为基础，若用户使用了自己的配置文件目录，则需要修改本应用配置目录中的对应文件，并注意需要将配置文件上传到executor端。

   由于hdfs和hbase客户端访问服务端时，使用hostname配置服务端地址，因此，客户端的/etc/hosts需要保存有所有需要访问节点的hosts配置。用户可预先将对端集群节点的host添加到客户端节点的/etc/hosts文件中。

   - 访问Hive MetaStore：使用cluster2中的Spark2x客户端下“conf”目录下的hive-site.xml文件，替换到cluster1中的Spark2x客户端下“conf”目录下的hive-site.xml文件。

     如上操作后可以用sparksql访问hive MetaStore，如需访问hive表数据，需要按照**•同时访问两个集群的HDFS：**的操作步骤配置且指定对端集群nameservice为LOCATION后才能访问表数据。
   - 访问对端集群的HBase：
     i. 先将cluster2集群的所有Zookeeper节点和HBase节点的IP和主机名配置到cluster1集群的客户端节点的/etc/hosts文件中。
     ii. 使用cluster2中的Spark2x客户端下“conf”目录的hbase-site.xml文件，替换到cluster1中的Spark2x客户端下“conf”目录hbase-site.xml文件。
   - 访问Kafka，仅需将应用访问的Kafka Broker地址设置为cluster2中的Kafka Broker地址即可。
   - 同时访问两个集群的HDFS：
     - 无法同时获取两个相同nameservice的token，因此两个HDFS的nameservice必须不同，例如：一个为hacluster，一个为test
       1) 从cluster2的hdfs-site.xml中获取以下配置，添加到cluster1的spark2x客户端conf目录的hdfs-site.xml中

          dfs.nameservices.mappings、dfs.nameservices、dfs.namenode.rpc-address.test.*、dfs.ha.namenodes.test、dfs.client.failover.proxy.provider.test

          参考样例如下：

          ```
          <property>
          <name>dfs.nameservices.mappings</name>
          <value>[{"name":"hacluster","roleInstances":["14","15"]},
          {"name":"test","roleInstances":["16","17"]}]</value>
          </property>
          <property>
          <name>dfs.nameservices</name>
          <value>hacluster,test</value>
          </property>
          <property>
          <name>dfs.namenode.rpc-address.test.16</name>
          <value>192.168.0.1:8020</value>
          </property>
          <property>
          <name>dfs.namenode.rpc-address.test.17</name>
          <value>192.168.0.2:8020</value>
          </property>
          ```

```
<property>
<name>dfs.ha.namenodes.test</name>
<value>16,17</value>
</property>
<property>
<name>dfs.client.failover.proxy.provider.test</name>
<value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider
</value>
</property>
```

2) 修改cluster1的spark客户端conf目录下的spark-defaults.conf配置文件中，修改spark.yarn.extra.hadoopFileSystems = hdfs://test，spark.hadoop.hdfs.externalToken.enable = true，如下所示：

```
spark.yarn.extra.hadoopFileSystems = hdfs://test
spark.hadoop.hdfs.externalToken.enable = true
```

3) 应用提交命令中，需要添加--keytab 和 --principal参数，参数配置为cluster1中提交任务的用户。

4) 使用cluster1的spark客户端提交应用，即可同时访问两个hdfs服务

- 同时访问两个集群的HBase：

i. 修改cluster1的spark客户端conf目录下的spark-defaults.conf配置文件中，修改spark.hadoop.hbase.externalToken.enable = true，如下所示：

```
spark.hadoop.hbase.externalToken.enable = true
```

ii. 用户访问HBase时，需要使用对应集群的配置文件创建Configuration对象，用于创建Connection对象。

iii. MRS集群中支持同时获取多个HBase服务的token，以解决Executor中无法访问HBase的问题，使用方式如下：

假设需要访问本集群的HBase和cluster2的HBase，将cluster2的hbase-site.xml文件放到一个压缩包内，压缩包命名为external_hbase_conf***，提交命令时，使用--archives指定这些压缩包。

12.23.8.12 对同一目录创建多个外表，可能导致外表查询失败

问题

假设存在数据文件路径“/test_data_path”，用户userA对该目录创建外表tableA，用户userB对该目录创建外表tableB，当userB对tableB执行insert操作后，userA将查询tableA失败，出现Permission denied异常。

回答

当userB对tableB执行insert操作后，会在外表数据路径下生成新的数据文件，且文件属组是userB，当userA查询tableA时，会读取外表数据目录下的所有的文件，此时会因没有userB生成的文件的读取权限而查询失败。

实际上，不只是查询场景，还有其他场景也会出现问题。例如：inset overwrite操作将会把此目录下的其他表文件也一起复写。

由于Spark SQL当前的实现机制，如果对此种场景添加检查限制，会存在一致性问题和性能问题，因此未对此种场景添加限制，但是用户应避免此种用法，以避免此场景带来的各种问题。

12.23.8.13 访问 Spark2x JobHistory 中某个应用的原生页面时页面显示错误

问题

提交一个Spark应用，包含单个Job 百万个task。应用结束后，在JobHistory中访问该应用的原生页面，浏览器会等待较长时间才跳转到应用原生页面，若10分钟内无法跳转，则页面会显示Proxy Error信息。

图 12-68 错误信息样例

回答

在JobHistory界面中跳转到某个应用的原生页面时，JobHistory需要回放该应用的Event log，若应用包含的事件日志较大，则回放时间较长，浏览器需要较长时间的等待。

当前浏览器访问JobHistory原生页面需经过httpd代理，代理的超时时间是10分钟，因此，如果JobHistory在10分钟内无法完成Event log的解析并返回，httpd会主动向浏览器返回Proxy Error信息。

解决方法

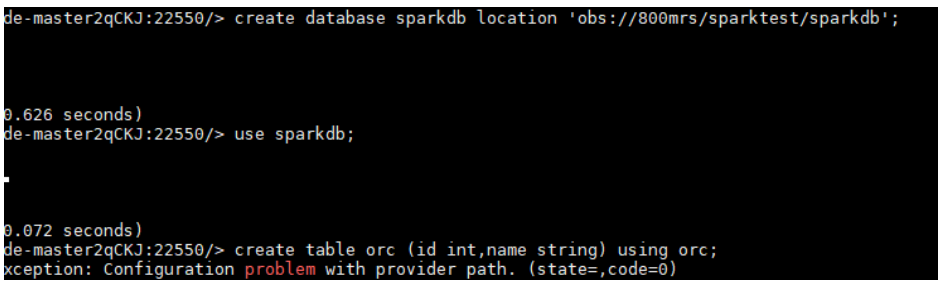
由于当前JobHistory开启了本地磁盘缓存功能，访问应用时，会将应用的Event log的解析结果缓存到本地磁盘中，第二次访问时，能大大加快响应速度。因此，出现此种情况时，仅需稍作等待，重新访问原来的链接即可，此时不会再出现需要长时间等待的现象。

12.23.8.14 对接 OBS 场景中，spark-beeline 登录后指定 loaction 到 OBS 建表失败

问题

对接OBS ECS/BMS集群，spark-beeline登录后，指定location到OBS建表报错失败。

图 12-69 错误信息



回答

HDFS上ssl.jceks文件权限不足，导致建表失败。

```
Caused by: org.apache.hadoop.security.AccessControlException: Permission denied: user=root, access=READ, inode="/user/spark2x/jars/8.0.2/ssl.jceks":spark2x:hadoop:-rw-------
    at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:410)
    at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:264)
    at com.huawei.hadoop.adapter.hdfs.plugin.HWAccessControlEnforce.checkPermission(HWAccessControlEnforce.java:54)
    at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:194)
    at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1957)
    at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1941)
    at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPathAccess(FSDirectory.java:1891)
    at org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getBlockLocations(FSDirStatAndListingOp.java:175)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getBlockLocations(FSNamesystem.java:1990)
    at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getBlockLocations(NameNodeRpcServer.java:762)
    at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getBlockLocations(ClientNamenodeProtocolServerSideTranslatorPB.java:445)
    at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:528)
    at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1036)
    at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:985)
    at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:913)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1737)
    at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2876)
```

解决方法

1. 使用**omm**用户登录Spark2x所在节点，执行如下命令：

   vi ${BIGDATA_HOME}/FusionInsight_Spark2x_8.1.0.1/install/FusionInsight-Spark2x-3.1.1/spark/sbin/fake_prestart.sh

2. 将“eval "${hdfsCmd}" -chmod 600 "${InnerHdfsDir}"/ssl.jceks >> "${PRESTART_LOG}" 2>&1”修改成“eval "${hdfsCmd}" -chmod 644 "${InnerHdfsDir}"/ssl.jceks >> "${PRESTART_LOG}" 2>&1”。
3. 重启SparkResource实例。

12.23.8.15 Spark shuffle 异常处理

问题

在部分场景Spark shuffle阶段会有如下异常

```
2021-06-18 02:53:08,364 | INFO  | [shuffle-server-6-1] | DIGEST41:Unmatched MACs | javax.security.sasl.unwrap(DigestMD5Base.java:1481)
2021-06-18 02:53:08,368 | WARN  | [shuffle-server-6-1] | Exception in connection from /[redacted] | org.apache.spark.network.server.TransportChannelHandler.exceptionCaught(TransportChannelHandler.java:87)
io.netty.handler.codec.DecoderException: javax.security.sasl.SaslException: DIGEST-MD5: Out of order sequencing of messages from server. Got: 16 Expected: 14
        at io.netty.handler.codec.MessageToMessageDecoder.channelRead(MessageToMessageDecoder.java:98)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:379)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:365)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelRead(AbstractChannelHandlerContext.java:357)
        at org.apache.spark.network.util.TransportFrameDecoder.channelRead(TransportFrameDecoder.java:102)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:379)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:365)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelRead(AbstractChannelHandlerContext.java:357)
        at io.netty.channel.DefaultChannelPipeline$HeadContext.channelRead(DefaultChannelPipeline.java:1410)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:379)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerContext.java:365)
        at io.netty.channel.DefaultChannelPipeline.fireChannelRead(DefaultChannelPipeline.java:919)
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:163)
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:714)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysOptimized(NioEventLoop.java:650)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:576)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:493)
        at io.netty.util.concurrent.SingleThreadEventExecutor$4.run(SingleThreadEventExecutor.java:989)
        at io.netty.util.internal.ThreadExecutorMap$2.run(ThreadExecutorMap.java:74)
        at io.netty.util.concurrent.FastThreadLocalRunnable.run(FastThreadLocalRunnable.java:30)
        at java.lang.Thread.run(Thread.java:748)
Caused by: javax.security.sasl.SaslException: DIGEST-MD5: Out of order sequencing of messages from server. Got: 16 Expected: 14
        at com.sun.security.sasl.digest.DigestMD5Base$DigestPrivacy.unwrap(DigestMD5Base.java:1489)
        at com.sun.security.sasl.digest.DigestMD5Base.unwrap(DigestMD5Base.java:213)
        at org.apache.spark.network.sasl.SparkSaslServer.unwrap(SparkSaslServer.java:149)
        at org.apache.spark.network.sasl.SaslEncryption$DecryptionHandler.decode(SaslEncryption.java:126)
        at org.apache.spark.network.sasl.SaslEncryption$DecryptionHandler.decode(SaslEncryption.java:101)
        at io.netty.handler.codec.MessageToMessageDecoder.channelRead(MessageToMessageDecoder.java:88)
        ... 20 more
```

解决方法

JDBC应该：

登录FusionInsight Manager管理界面，修改JDBCServer的参数“spark.authenticate.enableSaslEncryption”值为“false”，并重启对应的实例。

客户端作业：

客户端应用在提交应用的时候，修改spark-defaults.conf配置文件的“spark.authenticate.enableSaslEncryption”值为“false”。

12.24 使用 Sqoop

12.24.1 从零开始使用 Sqoop

Sqoop是一款开源的工具，主要用于在Hadoop(Hive)与传统的数据库(MySQL、PostgreSQL...)间进行数据的传递，可以将一个关系型数据库（例如：MySQL、

Oracle、PostgreSQL等）中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。

前提条件

- MRS 3.1.0及之后版本在创建集群时已勾选Sqoop组件。
- 安装客户端，具体请参考**安装客户端（3.x及之后版本）**。例如安装目录为“/opt/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

sqoop export (HDFS 到 MySQL)

步骤1 登录客户端所在节点。

步骤2 执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤3 使用sqoop命令操作sqoop客户端。

sqoop export --connect jdbc:mysql://10.100.231.134:3306/test --username root --password xxxxxx --table component13 -export-dir hdfs://hacluster/user/hive/warehouse/component_test3 --fields-terminated-by ',' -m 1

表 12-430 参数说明

参数	说明
-direct	快速模式，利用了数据库的导入工具，如MySQL的mysqlimport，可以比jdbc连接的方式更为高效的将数据导入到关系数据库中。
-export-dir <dir>	存放数据的HDFS的源目录。
-m或-num-mappers <n>	启动n个map来并行导入数据，默认是4个，该值请勿高于集群的最大Map数。
-table <table-name>	要导入的目的关系数据库表。
-update-key <col-name>	后面接条件列名，通过该参数可以将关系数据库中已经存在的数据进行更新操作，类似于关系数据库中的update操作。
-update-mode <mode>	更新模式，有两个值updateonly和默认的allowinsert，该参数只能在关系数据表里不存在要导入的记录时才能使用，比如要导入的hdfs中有一条id=1的记录，如果在表里已经有一条记录id=2，那么更新会失败。
-input-null-string <null-string>	可选参数，如果没有指定，则字符串null将被使用。
-input-null-non-string <null-string>	可选参数，如果没有指定，则字符串null将被使用。

参数	说明
-staging-table <staging-table-name>	创建一个与导入目标表同样数据结构的表，将所有数据先存放在该表中，然后由该表通过一次事务将结果写入到目标表中。 该参数是用来保证在数据导入关系数据库表的过程中的事务安全性，因为在导入的过程中可能会有多个事务，那么一个事务失败会影响到其它事务，比如导入的数据会出现错误或出现重复的记录等等情况，那么通过该参数可以避免这种情况。
-clear-staging-table	如果该staging-table非空，则通过该参数可以在运行导入前清除staging-table里的数据。

----结束

sqoop import（MySQL 到 Hive 表）

步骤1 登录客户端所在节点。

步骤2 执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤3 使用sqoop命令操作sqoop客户端。

sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password xxxxxx --table component --hive-import --hive-table component_test2 --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile

表 12-431 参数说明

参数	说明
-append	将数据追加到hdfs中已经存在的dataset中。使用该参数，sqoop将把数据先导入到一个临时目录中，然后重新给文件命名到一个正式的目录中，以避免和该目录中已存在的文件重名。
-as-avrodatafile	将数据导入到一个Avro数据文件中。
-as-sequencefile	将数据导入到一个sequence文件中。
-as-textfile	将数据导入到一个普通文本文件中，生成该文本文件后，可以在hive中通过sql语句查询出结果。

参数	说明
-boundary-query <statement>	边界查询，在导入前先通过SQL查询得到一个结果集，然后导入的数据就是该结果集内的数据，格式如：**-boundary-query 'select id,creationdate from person where id = 3'**，表示导入的数据为id=3的记录，或者**select min(<split-by>), max(<split-by>) from <table name>**。 注意：查询的字段中不能有数据类型为字符串的字段，否则会报错：java.sql.SQLException: Invalid value for getLong()。
-columns<col,col,col…>	指定要导入的字段值，格式如：-columns id,username
-direct	快速模式，利用了数据库的导入工具，如MySQL的mysqlimport，可以比jdbc连接的方式更为高效的将数据导入到关系数据库中。
-direct-split-size	在使用上面direct直接导入的基础上，对导入的流按字节数分块，特别是使用直连模式从PostgreSQL导入数据时，可以将一个到达设定大小的文件分为几个独立的文件。
-inline-lob-limit	设定大对象数据类型的最大值。
-m或-num-mappers	启动n个map来并行导入数据，默认是4个，该值请勿高于集群的最大Map数。
-query，-e<statement>	从查询结果中导入数据，该参数使用时必须指定-target-dir、-hive-table，在查询语句中一定要有where条件且在where条件中需要包含$CONDITIONS。 示例：-query 'select * from person where $CONDITIONS ' -target-dir /user/hive/warehouse/person -hive-table person
-split-by<column-name>	表的列名，用来切分工作单元，一般后面跟主键ID。
-table <table-name>	关系数据库表名，数据从该表中获取。
-target-dir <dir>	指定hdfs路径。
-warehouse-dir <dir>	与**-target-dir**不能同时使用，指定数据导入的存放目录，适用于导入hdfs，不适合导入hive目录。
-where	从关系数据库导入数据时的查询条件，示例：-where 'id = 2'
-z,-compress	压缩参数，默认数据不压缩，通过该参数可以使用gzip压缩算法对数据进行压缩，适用于SequenceFile，text文本文件，和Avro文件。
–compression-codec	Hadoop压缩编码，默认为gzip。
–null-string <null-string>	替换null字符串，如果没有指定，则字符串null将被使用。

参数	说明
–null-non-string<null-string>	替换非String的null字符串，如果没有指定，则字符串null将被使用。
-check-column (col)	增量导入参数，用来作为判断的列名，如id。
-incremental (mode) append 或lastmodified	增量导入参数。 append：追加，比如对大于last-value指定的值之后的记录进行追加导入。 lastmodified：最后的修改时间，追加last-value指定的日期之后的记录。
-last-value (value)	增量导入参数，指定自从上次导入后列的最大值（大于该指定的值），也可以自己设定某一值。

----结束

Sqoop 使用样例

- sqoop import（MySQL到HDFS）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--query 'SELECT * FROM component where $CONDITIONS and component_id ="MRS 1.0_002"' --target-dir /tmp/component_test --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile**

- sqoop export （obs到MySQL）

  sqoop export --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--table component14 -export-dir obs://obs-file-bucket/xx/part-m-00000 --fields-terminated-by ',' -m 1**

- sqoop import（MySQL到obs）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--table component --target-dir obs://obs-file-bucket/xx --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile**

- sqoop import（MySQL到Hive外obs表）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--table component --hive-import --hive-table component_test01 --fields-terminated-by "," -m 1 --as-textfile**

12.24.2 Sqoop1.4.7 适配 MRS 3.x 集群

Sqoop是专为Apache Hadoop和结构化数据库（如关系型数据库）设计的高效传输大量数据的工具。客户需要在MRS中使用sqoop进行数据迁移，MRS旧版本中未自带Sqoop，客户可参考此文档自行安装使用。MRS 3.1.0及之后版本已支持创建集群时勾选Sqoop组件，请创建集群时勾选即可。

前提条件

已安装MRS客户端的节点，且已安装 jdk环境。

```
[root@node-masterlfKEj bin]# echo $JAVA_HOME
/opt/Bigdata/client/JDK/jdk1.8.0_242
```

Sqoop1.4.7 适配步骤

步骤1 下载开源sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz包（下载地址**http://archive.apache.org/dist/sqoop/1.4.7/**）。

步骤2 将下载好的sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz包放入已安装MRS客户端的节点的“/opt/Bigdata/client”目录并解压。

tar zxvf sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz

步骤3 从MySQL官网下载MySQL jdbc驱动程序“mysql-connector-java-xxx.jar”，具体MySQL jdbc驱动程序选择参见下表。

表 12-432 版本信息

jdbc驱动程序版本	MySQL版本
Connector/J 5.1	MySQL 4.1、MySQL 5.0、MySQL 5.1、MySQL 6.0 alpha
Connector/J 5.0	MySQL 4.1、MySQL 5.0 servers、distributed transaction (XA)
Connector/J 3.1	MySQL 4.1、MySQL 5.0 servers、MySQL 5.0 except distributed transaction (XA)
Connector/J 3.0	MySQL 3.x、MySQL 4.1

步骤4 将MySQL 驱动包放入Sqoop的lib目录下（/opt/Bigdata/client/sqoop-1.4.7.bin__hadoop-2.6.0/lib）并修改jar包的属组和权限，参考**图12-70**的omm:wheel 和755的属组和权限。

图 12-70 MySQL 驱动包的属组和权限

```
-rwxr-xr-x. 1 omm wheel 1007502 Apr 28  2020 mysql-connector-java-5.1.47.jar
```

步骤5 使用MRS客户端中Hive的lib目录下（/opt/Bigdata/client/Hive/Beeline/lib）的jackson开头的jar包替换Sqoop的lib下的相应jar包。

图 12-71 jackson 开头的 jar

```
-rwxr-xr-x. 1 omm wheel   46989 Apr 28  2020 jackson-annotations-2.6.3.jar
-rwxr-xr-x. 1 omm wheel  258876 Apr 28  2020 jackson-core-2.6.5.jar
-rwxr-xr-x. 1 omm wheel  232248 Apr 28  2020 jackson-core-asl-1.9.13.jar
-rwxr-xr-x. 1 omm wheel 1171380 Apr 28  2020 jackson-databind-2.6.5.jar
-rwxr-xr-x. 1 omm wheel   18336 Apr 28  2020 jackson-jaxrs-1.9.13.jar
-rwxr-xr-x. 1 omm wheel  780664 Apr 28  2020 jackson-mapper-asl-1.9.13.jar
-rwxr-xr-x. 1 omm wheel   27084 Apr 28  2020 jackson-xc-1.9.13.jar
```

步骤6 将MRS Hive客户端中（/opt/Bigdata/client/Hive/Beeline/lib）的jline的包，拷贝到Sqoop的lib下。

步骤7 执行**vim $JAVA_HOME/jre/lib/security/java.policy**增加如下配置：

permission javax.management.MBeanTrustPermission "register";

步骤8 执行如下命令，进入Sqoop的conf目录并增加配置：

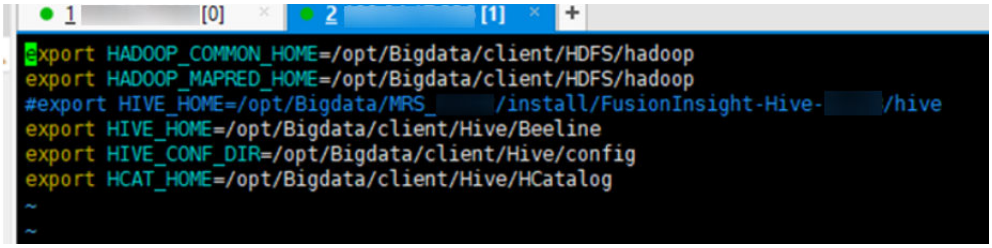
cd /opt/Bigdata/client/sqoop-1.4.7.bin__hadoop-2.6.0/conf

cp sqoop-env-template.sh sqoop-env.sh

步骤9 执行**vim sqoop-env.sh** 设置Sqoop的环境变量，Hadoop、Hive的目录根据实际目录修改。

```
export HADOOP_COMMON_HOME=/opt/Bigdata/client/HDFS/hadoop
export HADOOP_MAPRED_HOME=/opt/Bigdata/client/HDFS/hadoop
export HIVE_HOME=/opt/Bigdata/MRS_1.9.X/install/FusionInsight-Hive-3.1.0/hive(请按照实际路径填写)
export HIVE_CONF_DIR=/opt/Bigdata/client/Hive/config
export HCAT_HOME=/opt/Bigdata/client/Hive/HCatalog
```

图 12-72 设置 Sqoop 的环境变量



步骤10 编写Sqoop脚本 例如：

```
/opt/Bigdata/FusionInsight_Current/1_19_SqoopClient/install/FusionInsight-Sqoop-1.4.7/bin/sqoop import
--connect jdbc:mysql://192.168.0.183:3306/test
--driver com.mysql.jdbc.Driver
--username 'root'
--password 'xxx'
--query "SELECT id, name  FROM tbtest WHERE \$CONDITIONS"
--hcatalog-database default
--hcatalog-table test
--num-mappers 1
```

----结束

12.24.3 Sqoop 常用命令及参数介绍

Sqoop 常用命令介绍

表 12-433 Sqoop 常用命令介绍

命令	说明
import	数据导入到集群
export	集群数据导出
codegen	获取数据库中某张表数据生成Java并打包jar
create-hive-table	创建Hive表
eval	执行sql并查看结果

命令	说明
import-all-tables	导入某个数据库下的所有表到HDFS中
job	生成一个sqoop任务
list-databases	列举数据库名
list-tables	列举表名
merge	将HDFS不同目录下的数据合在一起并存放到指定目录
metastore	启动元数据库，记录sqoop job的元数据
help	打印帮助信息
version	打印版本信息

公用参数介绍

表 12-434 公用参数介绍

分类	参数	说明
连接数据库	--connect	连接关系型数据库的url
	--connection-manager	指定连接管理类
	--driver jdbc	连接驱动包
	--help	帮助信息
	--password	连接数据库密码
	--username	连接数据库的用户名
	--verbose	在控制台打印详细信息
import参数	--fields-terminated-by	设定字段分隔符，和Hive表或hdfs文件保持一致
	--lines-terminated-by	设定行分隔符，和hive表或hdfs文件保持一致
	--mysql-delimiters	MySQL默认分隔符设置
export参数	--input-fields-terminated-by	字段分隔符
	--input-lines-terminated-by	行分隔符

分类	参数	说明
hive 参数	--hive-delims-replacement	用自定义的字符替换数据中的\r\n等字符
	--hive-drop-import-delims	在导入数据到hive时，去掉\r\n等字符
	--map-column-hive	生成hive表时可以更改字段的数据类型
	--hive-partition-key	创建分区
	--hive-partition-value	导入数据库指定分区
	--hive-home	指定hive安装目录
	--hive-import	表示操作是从关系型数据库导入到hive中
	--hive-overwrite	覆盖hive已有数据
	--create-hive-table	创建Hive表，默认false, 如果目标表不存在，则会创建目标表
	--hive-table	指定hive表
	--table	关系型数据库表名
	--columns	指定需要导入的关系型数据表字段
	--query	指定查询语句，将查询结果导入
hcatalog参数	--hcatalog-database	指定hive库，使用hcatalog方式导入hive库
	--hcatalog-table	指定hive表，使用hcatalog方式导入hive表
其他参数	-m或--num-mappers	后跟数字，表示sqoop任务的分片数
	--split-by	按照某一字段进行分片，配合-m
	--target-dir	指定hdfs临时目录
	--null-string string	类型为null时替换字符串
	--null-non-string	非string类型为null时替换字符串
	--check-column	增量判断的字段
	--incremental append或lastmodified	增量导入参数 append：追加，比如对大于last-value指定的值之后的记录进行追加导入。 lastmodified：最后的修改时间，追加last-value指定的日期之后的记录。

分类	参数	说明
	--last-value	指定一个值，用于标记增量导入
	--input-null-string	替换null字符串，如果没有指定，则字符串null将被使用。
	--input-null-non-string	替换非String的null字符串，如果没有指定，则字符串null将被使用。

12.24.4 Sqoop 常见问题

12.24.4.1 报错找不到 QueryProvider 类

问题

报错找不到QueryProvider类。

```
2021-04-06 15:57:10,756 INFO manager.SqlManager: Using default fetchSize of 1000
2021-04-06 15:57:10,756 INFO tool.CodeGenTool: Beginning code generation
Apr 06, 2021 3:57:10 PM java.util.logging.LogManager$RootLogger log
SEVERE: Error loading factory org.apache.calcite.jdbc.CalciteJdbc41Factory
java.lang.NoClassDefFoundError: org/apache/calcite/linq4j/QueryProvider
        at java.lang.ClassLoader.defineClass1(Native Method)
        at java.lang.ClassLoader.defineClass(ClassLoader.java:757)
        at java.security.SecureClassLoader.defineClass(SecureClassLoader.java:142)
        at java.net.URLClassLoader.defineClass(URLClassLoader.java:468)
        at java.net.URLClassLoader.access$100(URLClassLoader.java:74)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:369)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:363)
        at java.security.AccessController.doPrivileged(Native Method)
        at java.net.URLClassLoader.findClass(URLClassLoader.java:362)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:419)
        at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:352)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:352)
        at java.lang.ClassLoader.defineClass1(Native Method)
        at java.lang.ClassLoader.defineClass(ClassLoader.java:757)
        at java.security.SecureClassLoader.defineClass(SecureClassLoader.java:142)
        at java.net.URLClassLoader.defineClass(URLClassLoader.java:468)
        at java.net.URLClassLoader.access$100(URLClassLoader.java:74)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:369)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:363)
        at java.security.AccessController.doPrivileged(Native Method)
        at java.net.URLClassLoader.findClass(URLClassLoader.java:362)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:419)
        at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:352)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:352)
        at java.lang.ClassLoader.defineClass1(Native Method)
        at java.lang.ClassLoader.defineClass(ClassLoader.java:757)
```

回答

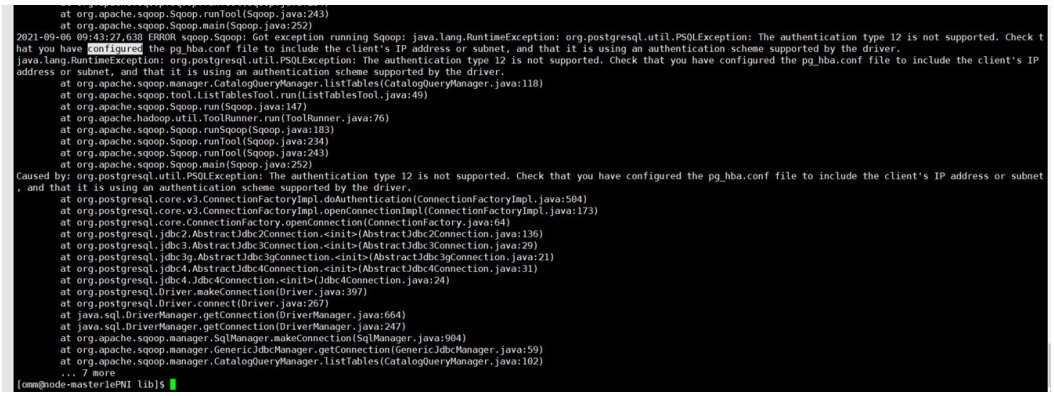
搜索mrs客户端目录，将以下两个jar包放入sqoop的lib目录下。

```
-rwxr-xr-x. 1 omm wheel 4813045 Apr  6 15:56 calcite-core-1.19.0.jar
-rwxr-xr-x. 1 omm wheel  459944 Apr  6 16:01 calcite-linq4j-1.19.0.jar
```

12.24.4.2 连接 postgresql 或者 gaussdb 时报错

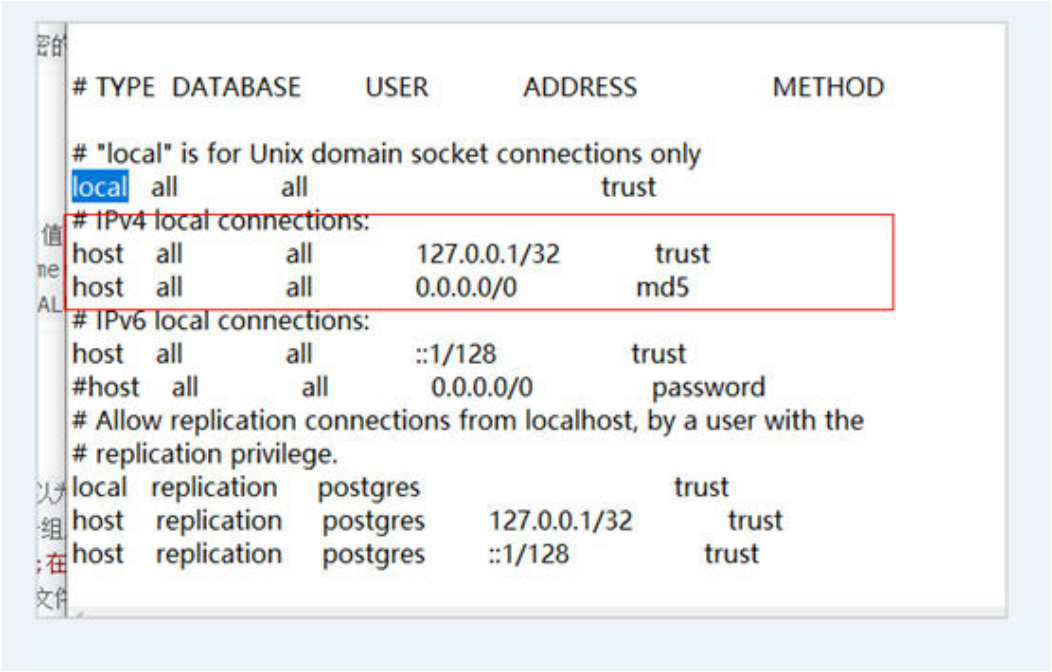
问题

连接postgresql或者gaussdb时报错。



回答

调整数据库的pg_hba.conf文件，将address改成sqoop所在节点的ip。



，

12.24.4.3 使用 hive-table 方式同步数据到 obs 上的 hive 表报错

问题

使用hive-table方式同步数据到obs上的hive表报错。

```
2021-09-03 16:28:11,611 ERROR tools.DistCp: XAttrs not supported on at least one file system:
org.apache.hadoop.tools.CopyListing$XAttrsNotSupportedException: XAttrs not supported for file system:
obs://fdd-fs
        at org.apache.hadoop.tools.util.DistCpUtils.checkFileSystemXAttrSupport(DistCpUtils.java:555)
        at org.apache.hadoop.tools.DistCp.configureOutputFormat(DistCp.java:341)
        at org.apache.hadoop.tools.DistCp.createJob(DistCp.java:308)
        at org.apache.hadoop.tools.DistCp.createAndSubmitJob(DistCp.java:218)
        at org.apache.hadoop.tools.DistCp.execute(DistCp.java:197)
        at org.apache.hadoop.tools.DistCp.run(DistCp.java:155)
```

回答

修改数据同步方式，将**-hive-table**改成**-hcatalog-table**。

12.24.4.4 使用 hive-table 方式同步数据到 orc 表或者 parquet 表失败

问题

使用hive-table方式同步数据到orc表或者parquet表失败，报错中会有kite-sdk的包名。

回答

修改数据同步方式，将**-hive-table**改成**-hcatalog-table**。

12.24.4.5 使用 hive-table 方式同步数据报错

问题

使用hive-table方式同步数据报错。

```
        ... 41 more
14:41:42.891 [bfe1438c-07bb-43fd-91d9-910a348f6e91 main] ERROR org.apache.hadoop.hive.metastore.ObjectStore - Version information not found in metastore. The process will exit.
14:41:42.892 [bfe1438c-07bb-43fd-91d9-910a348f6e91 main] ERROR org.apache.hadoop.hive.metastore.RetryingHMSHandler - ExitSecurityException
        at org.apache.sqoop.util.SubprocessSecurityManager.checkExit(SubprocessSecurityManager.java:83)
        at java.lang.Runtime.exit(Runtime.java:107)
        at java.lang.System.exit(System.java:973)
        at org.apache.hadoop.hive.metastore.ObjectStore.checkSchema(ObjectStore.java:9555)
        at org.apache.hadoop.hive.metastore.ObjectStore.verifySchema(ObjectStore.java:9531)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.hive.metastore.RawStoreProxy.invoke(RawStoreProxy.java:97)
        at com.sun.proxy.$Proxy37.verifySchema(Unknown Source)
        at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.getMSForConf(HiveMetaStore.java:903)
        at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.getMS(HiveMetaStore.java:895)
        at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.createDefaultDB(HiveMetaStore.java:978)
        at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.init(HiveMetaStore.java:585)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.hive.metastore.RetryingHMSHandler.invokeInternal(RetryingHMSHandler.java:148)
        at org.apache.hadoop.hive.metastore.RetryingHMSHandler.invoke(RetryingHMSHandler.java:109)
        at org.apache.hadoop.hive.metastore.RetryingHMSHandler.<init>(RetryingHMSHandler.java:81)
        at org.apache.hadoop.hive.metastore.RetryingHMSHandler.getProxy(RetryingHMSHandler.java:94)
        at org.apache.hadoop.hive.metastore.HiveMetaStore.newRetryingHMSHandler(HiveMetaStore.java:9683)
        at org.apache.hadoop.hive.metastore.HiveMetaStoreClient.<init>(HiveMetaStoreClient.java:185)
        at org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient.<init>(SessionHiveMetaStoreClient.java:96)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.hadoop.hive.metastore.utils.JavaUtils.newInstance(JavaUtils.java:84)
        at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.<init>(RetryingMetaStoreClient.java:97)
```

回答

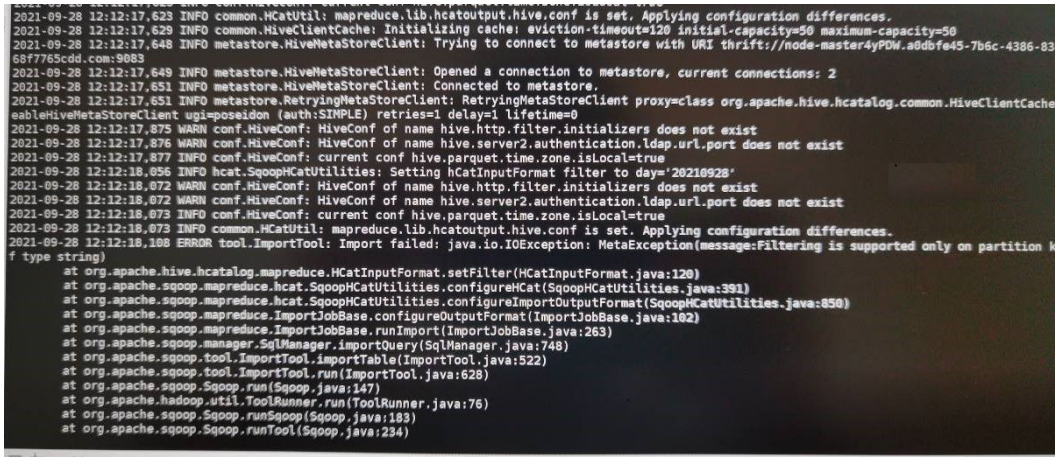
修改hive-site.xml，加入如下值。

```
<property>
<name>hive.metastore.schema.verification</name>
<value>false</value>
</property>
```

12.24.4.6 使用 hcatalog 方式同步 hive parquet 表报错

问题

同步hive parquet表，其分区字段为非string类型，无法正常使用hive import导入，只能考虑使用hcatalog方式，但是hcatalog方式报错如下：



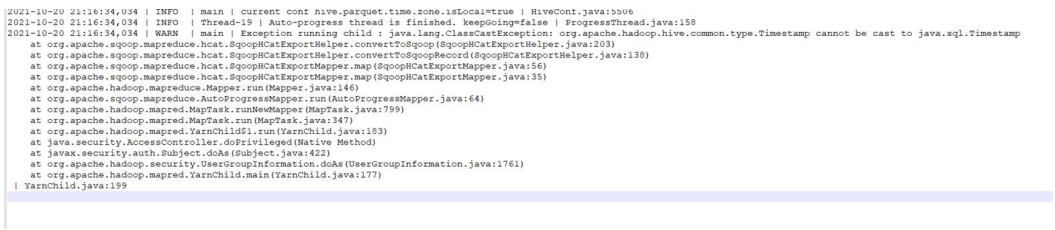
回答

1. 修改sqoop源码SqoopHCatUtilities中的代码，将限制代码去掉。
2. 修改hive客户端中的hive-site.xml文件，修改hive.metastore.integral.jdo.pushdown参数为true。

12.24.4.7 使用 Hcatalog 方式同步 Hive 和 MySQL 之间的数据，timestamp 和 data 类型字段会报错

问题

使用Hcatalog方式同步Hive和MySQL之间的数据，timestamp和data类型字段会报错：



回答

- 调整Sqoop源码包中的代码，将timestamp强制转换类型和Hive保持一致。
- 将Hive中的字段类型修改为String。

12.25 使用 Storm

12.25.1 从零开始使用 Storm

用户可以在MRS集群的客户端中提交和删除Storm拓扑等基本功能。

前提条件

已安装MRS集群客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient”。

cd /opt/hadoopclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 启用Kerberos认证的集群，执行以下命令认证用户身份。未启用Kerberos认证的集群无需执行。

kinit *Storm用户*

步骤5 执行以下命令，提交Storm拓扑：

storm jar *拓扑包路径 拓扑Main方法的类名称 拓扑名称*

界面提示以下信息表示提交成功：

```
Finished submitting topology: topo1
```

步骤6 执行以下命令，查看Storm中的拓扑。启用Kerberos认证的集群，只有属于“stormadmin”或“storm”的用户可以查看所有拓扑。

storm list

步骤7 执行以下命令，删除Storm中的拓扑。

storm kill *拓扑名称*

----结束

12.25.2 使用 Storm 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Storm客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/hadoopclient”。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）

操作步骤

步骤1 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若安装了Storm多实例，在使用Storm命令提交拓扑时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，Storm-2实例：

source Storm-2/component_env

步骤5 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤6 执行命令进行客户端操作。

例如执行以下命令：

- **cql**
- **storm**

说明

同一个storm客户端不能同时连接安全和非安全的ZooKeeper。

----结束

12.25.3 使用客户端提交 Storm 拓扑

操作场景

用户可以根据业务需要，在集群的客户端中提交Storm拓扑，持续处理用户的流数据。启用Kerberos认证的集群，需要提交拓扑的用户属于“stormadmin”或“storm”组。

前提条件

已刷新客户端。

操作步骤

步骤1 根据业务情况，准备好客户端，登录安装客户端的节点。

请根据客户端所在位置，参考**使用MRS客户端**章节，登录安装客户端的节点。

步骤2 执行以下命令，设置拓扑的jar包权限。

例如修改“/opt/storm/topology.jar”的权限：

chmod 600 /opt/storm/topology.jar

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 若安装了Storm多实例，在使用Storm命令提交拓扑时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，Storm-2实例：

source Storm-2/component_env

步骤6 启用Kerberos认证的集群，执行以下命令认证用户身份。未启用Kerberos认证的集群无需执行。

kinit *Storm用户*

步骤7 MRS 3.x之前版本：执行以下命令，提交Storm拓扑。

storm jar *拓扑包路径 拓扑Main方法的类名称 拓扑名称*

界面提示以下信息表示提交成功：

```
Finished submitting topology: topo1
```

说明

- 如果需要拓扑支持采样消息，则还需要增加参数“topology.debug”和“topology.eventlogger.executors”。
- 拓扑如何处理数据是拓扑自身行为。样例拓扑随机生成字符并分隔字符串，需要查看处理情况时，请启用采样功能并参见**查看Storm拓扑日志**。

步骤8 MRS 3.x及后续版本：执行以下命令，提交拓扑任务。

storm jar *topology-jar-path class 入参列表*

- topology-jar-path：表示拓扑的jar包所在路径。
- class：表示拓扑使用的main方法所在类名称。
- 入参列表：表示拓扑使用的main方法入参。

显示以下信息表示拓扑提交成功：

```
Finished submitting topology: topology1
```

说明

- 登录认证用户必须与所加载环境变量（component_env）一一对应，否则使用storm命令提交拓扑任务出错。
- 加载客户端环境变量且对应用户登录成功后，该用户可以在任意storm客户端下执行storm命令来提交拓扑任务，但提交拓扑命令执行完成后，提交成功的拓扑仍然在用户所对应的Storm集群中，不会出现在其他Storm集群中。
- 如果修改了集群域名，需要在提交拓扑前重新设置域名信息，进入cql语句执行命令。

步骤9 执行以下命令，查看Storm中的拓扑。启用Kerberos认证的集群，只有属于“stormadmin”或“storm”的用户可以查看所有拓扑。

storm list

----结束

12.25.4 访问 Storm 的 WebUI

操作场景

用户可以通过Storm的WebUI，在图形化界面使用Storm。

Storm的WebUI支持查看以下信息：

- Storm集群汇总信息
- Nimbus汇总信息
- 拓扑汇总信息
- Supervisor汇总信息
- Nimbus配置信息

前提条件

- 获取用户“admin”帐号密码。“admin”密码在创建集群时由用户指定。
- 使用其他用户访问Storm WebUI，需要用户属于“storm”或“stormadmin”用户组。

操作步骤

步骤1 进入组件管理页面：

- MRS 3.x之前版本，单击集群名称，登录集群详情页面，选择“组件管理”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务”。

步骤2 登录Storm WebUI：

- MRS 3.x之前版本：选择“Storm”，在“Storm 概述”的“Storm Web UI”，单击任意一个UI链接，打开Storm的WebUI。

  说明

  第一次访问Storm WebUI，需要在浏览器中添加站点信任以继续打开页面。

- MRS 3.x及后续版本：选择“Storm > 概览”，在“基本信息”的“Storm Web UI”，单击任意一个UI链接，打开Storm的WebUI。

----结束

相关任务

- 单击拓扑名称，可查看指定拓扑的详细信息、拓扑状态、Spouts信息、Bolts信息和拓扑配置。
- 在“Topology actions”区域，用户可以对拓扑执行激活、去激活、重部署、删除操作、调试、停止调试和修改日志级别，即“Activate”、“Deactivate”、“Rebalance”、“Kill”、“Debug”、“Stop Debug”、“Change Log Level”。重部署和删除操作需要设置操作执行的等待时间，单位为秒。

- 在“Topology Visualization”区域，用户可以执行拓扑可视化操作，即单击“Show Visualization”。拓扑可视化后，WebUI将显示拓扑结构图。

12.25.5 管理 Storm 拓扑

操作场景

用户可以使用Storm的WebUI管理拓扑。“storm”用户组的用户只能管理由自己提交的拓扑任务，“stormadmin”用户组的用户可以管理所有拓扑任务。

操作步骤

步骤1 访问Storm的WebUI，请参考**访问Storm的WebUI**。

步骤2 在“Topology summary”区域，单击指定的拓扑名称。

步骤3 通过“Topology actions”管理Storm拓扑。

- 激活拓扑
  单击“Activate”，转化当前拓扑为激活状态。
- 去激活拓扑
  单击“Deactivate”，转化当前拓扑为去激活状态。
- 重部署拓扑
  单击“Rebalance”，将当前拓扑重新部署执行，需要输入执行重部署的等待时间，单位为秒。一般在集群中节点数发生变化时进行，以更好利用集群资源。
- 删除拓扑
  单击“Kill”，将当前拓扑删除，需要输入执行操作的等待时间，单位为秒。
- 采样、停止采样拓扑消息
  单击“Debug”，在弹出窗口输入流数据采样消息的数值，单位为百分比，表示从开始采样到停止采样这段时间内所有数据的采集比例。例如输入“10”，则采集比例为10%。
  如果需要停止采样，则单击“Stop Debug”。

  说明

  只有在提交拓扑时启用采样功能，才支持此功能。查看采样处理数据，请参见**查看Storm拓扑日志**。
- 修改拓扑日志级别
  单击“Change Log Level”，可以为Storm日志指定新的日志信息级别。

步骤4 显示拓扑结构图。

在“Topology Visualization”区域单击“Show Visualization”，执行拓扑可视化操作。

----结束

12.25.6 查看 Storm 拓扑日志

操作场景

用户需要查看Storm拓扑在worker进程中的执行情况时，需要查看worker中关于拓扑的日志。如果需要查询拓扑在运行时数据处理的日志，提交拓扑并启用“Debug”功

能后可以查看日志。仅启用Kerberos认证的流集群支持该场景，且用户需要是拓扑的提交者，或者加入“stormadmin”。

前提条件

- 在工作环境完成网络配置。
- 需要查看处理数据的拓扑，提交时已启用采样功能。

查看 worker 进程日志

步骤1 访问Storm的WebUI，请参考**访问Storm的WebUI**。

步骤2 在“Topology Summary”区域单击指定的拓扑名称，打开拓扑的详细信息。

步骤3 单击要查看日志的“Spouts”或“Bolts”任务，在“Executors (All time)”区域单击“Port”列的端口值，查看详细日志内容。

----结束

查看拓扑处理数据日志

步骤1 访问Storm的WebUI，请参考**访问Storm的WebUI**。

步骤2 在“Topology Summary”区域单击指定的拓扑名称，打开拓扑的详细信息。

步骤3 单击“Debug”，输入采样数据的百分比数值，并单击“OK”开始采样。

步骤4 单击拓扑的“Spouts”或“Bolts”任务，在“Component summary”单击“events”打开处理数据日志。

----结束

12.25.7 Storm 常用参数

本章节内容适用于MRS 3.x及后续版本。

参数入口

参数入口，请参考**修改集群服务配置参数**。

参数说明

表 12-435 参数说明

配置参数	说明	默认值
supervisor.slots.ports	supervisor上能够运行workers的端口列表。每个worker占用一个端口，且每个端口只运行一个worker。通过这项配置可以设置每台机器上运行的worker数量。端口的取值范围是1024到65535，不同端口使用逗号分隔。	6700,6701,6702,6703

配置参数	说明	默认值
WORKER_GC_OPTS	supervisor启动worker时使用的jvm选项。需要根据业务中对内存等的使用来进行设置，例如是简单业务处理，建议1G，既“-Xmx1G”；如果有窗口缓存，根据窗口大小计算：每条记录大小*周期*2。	-Xms1G -Xmx1G -XX:+UseG1GC -XX:+PrintGCDetails -Xloggc:artifacts/gc.log -XX:+PrintGCDateStamps -XX:+PrintGCTimeStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -XX:+HeapDumpOnOutOfMemoryError -XX:HeapDumpPath=artifacts/heapdump
default.schedule.mode	默认调度器的调度模式。目前支持两个值，具体值与含义如下： ● “AVERAGE”：使用按空闲Slot数目为优先级的调度机制 ● “RATE”：使用按空闲Slot比率为优先级的调度机制	AVERAGE
nimbus.thrift.threads	设置主用Nimbus对外提供服务时的最大连接线程数。当Storm集群规模较大，Supervisor实例数量较多时，需要增加线程数。	512

12.25.8 配置 Storm 业务用户密码策略

操作场景

本章节内容适用于MRS 3.x及后续版本。

使用Storm业务用户提交一个拓扑以后，该任务需要使用提交拓扑的用户身份持续运行。在拓扑运行的过程中，worker进程可能需要正常重启以保持拓扑工作。若业务用户的密码被修改，或密码使用天数超过了默认密码策略指定的最大有效期，则会影响拓扑正常运行。系统管理员需要根据企业安全要求，为Storm业务用户配置独立的密码策略。

说明

如果不为Storm业务用户配置独立的密码策略，在修改业务用户密码以后，可以删除旧的拓扑并重新提交，使拓扑继续运行。

对系统的影响

- 为Storm业务用户配置独立的密码策略后，此用户将不受Manager界面上的“密码策略”配置影响。
- 为Storm业务用户配置独立的密码策略后，如果配置了跨集群互信，请根据此密码策略，在Manager为Storm业务用户重置密码。

前提条件

系统管理员已明确业务需求，并创建好“人机”用户，例如“testpol”。

操作步骤

步骤1 以“omm”用户登录集群内任意节点。

步骤2 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤3 执行以下命令，导出环境变量。

EXECUTABLE_HOME="${CONTROLLER_HOME}/kerberos_user_specific_binay/kerberos"

LD_LIBRARY_PATH=${EXECUTABLE_HOME}/lib:$LD_LIBRARY_PATH

PATH=${EXECUTABLE_HOME}/bin:$PATH

步骤4 执行以下命令，并输入Kerberos管理员密码，进入Kerberos管理控制台。

kadmin -p kadmin/admin

说明

第一次使用“kadmin/admin”用户需要修改“kadmin/admin”密码。

界面显示如下信息，则表示已成功进入Kerberos管理控制台。

```
kadmin:
```

步骤5 执行以下命令，查看创建好的“Human-Machine”用户的具体信息。

getprinc *用户名*

例如，查看“testpol”用户的详细信息：

getprinc testpol

界面显示如下信息，说明指定用户使用了默认的密码策略：

```
Principal: testpol@<系统域名>
......
Policy: default
```

步骤6 执行以下命令，为Storm业务用户创建独立的密码策略，例如“streampol”：

addpol -maxlife 0day -minlife 0sec -history 1 -maxfailure 5 -failurecountinterval 5min -lockoutduration 5min -minlength 8 -minclasses 4 streampol

其中“-maxlife”表示密码最大有效期，“0day”表示永不过期。

步骤7 执行以下命令，查看新创建的策略“streampol”。

getpol streampol

界面显示如下信息，说明新策略已指定密码不过期：

```
Policy: streampol
 Maximum password life: 0 days 00:00:00
 ......
```

步骤8 执行以下命令，将新的策略“streampol”应用到Storm用户“testpol”。

modprinc -policy streampol testpol

其中“streampol”是策略名称，“testpol”是用户名。

界面显示如下信息，说明指定用户的属性已修改：

```
Principal "testpol@<系统域名>" modified.
```

步骤9 执行以下命令，查看Storm用户“testpol”用户的当前信息。

getprinc testpol

界面显示如下信息，说明指定用户使用了新的密码策略：

```
Principal: testpol@<系统域名>
 ......
 Policy: streampol
```

----结束

12.25.9 迁移 Storm 业务至 Flink

12.25.9.1 概述

本章节内容适用于MRS 3.x及后续版本。

Flink从0.10.0版本开始提供了一套API可以将使用Storm API编写的业务平滑迁移到Flink平台上，只需要极少的改动即可完成。通过这项转换可以覆盖大部分的业务场景。

Flink支持两种方式的业务迁移：

1. 完整迁移Storm业务：转换并运行完整的由Storm API开发的Storm拓扑。
2. 嵌入式迁移Storm业务：在Flink的DataStream中嵌入Storm的代码，如使用Storm API编写的Spout/Bolt。

Flink提供了flink-storm包用来完成上述转换。

12.25.9.2 完整迁移 Storm 业务

操作场景

该任务指导用户通过Storm业务完整迁移的方式转换并运行完整的由Storm API开发的Storm拓扑。

操作步骤

步骤1 打开Storm业务工程，修改工程的pom文件，增加“flink-storm” 、“flink-core”和“flink-streaming-java_2.11”的引用。如下：

```
<dependency>
   <groupId>org.apache.flink</groupId>
   <artifactId>flink-storm_2.11</artifactId>
   <version>1.4.0</version>
   <exclusions>
      <exclusion>
         <groupId>*</groupId>
         <artifactId>*</artifactId>
      </exclusion>
   </exclusions>
</dependency>
<dependency>
   <groupId>org.apache.flink</groupId>
   <artifactId>flink-core</artifactId>
   <version>1.4.0</version>
   <exclusions>
      <exclusion>
         <groupId>*</groupId>
         <artifactId>*</artifactId>
      </exclusion>
   </exclusions>
</dependency>
```

```
<dependency>
   <groupId>org.apache.flink</groupId>
   <artifactId>flink-streaming-java_2.11</artifactId>
   <version>1.4.0</version>
   <exclusions>
      <exclusion>
         <groupId>*</groupId>
         <artifactId>*</artifactId>
      </exclusion>
   </exclusions>
</dependency>
```

说明

如果是非maven工程，则手动收集如上jar包，添加到工程的classpath中。

步骤2 修改拓扑提交部分代码，下面以WordCount为例：

1. Storm拓扑的构造部分保持不变，无需修改，包括使用Storm API开发的Spout和Bolt都无需修改。

```
TopologyBuilder builder = new TopologyBuilder();
builder.setSpout("spout", new RandomSentenceSpout(), 5);
builder.setBolt("split", new SplitSentenceBolt(), 8).shuffleGrouping("spout");
builder.setBolt("count", new WordCountBolt(), 12).fieldsGrouping("split", new Fields("word"));
```

2. 拓扑的提交部分需要修改，Storm的提交示例如下：

```
Config conf = new Config();
conf.setNumWorkers(3);

StormSubmitter.submitTopology("word-count", conf, builder.createTopology());
```

需要进行如下修改：

```
Config conf = new Config();
 conf.setNumWorkers(3);

 //将Storm的Config转化为Flink的StormConfig
 StormConfig stormConfig = new StormConfig(conf);

 //使用Storm的TopologBuilder构造FlinkTopology
 FlinkTopology topology = FlinkTopology.createTopology(builder);

 //获取StreamExecutionEnvironment
 StreamExecutionEnvironment env = topology.getExecutionEnvironment();

 //将StormConfig设置到Job的环境变量中，用于构造Bolt和Spout
 //如果Bolt和Spout初始化时不需要config，则不用设置
 env.getConfig().setGlobalJobParameters(stormConfig);
 //执行拓扑提交
 topology.execute();
```

3. 重新打包之后使用flink命令行进行提交：

   flink run -class {MainClass} WordCount.jar

----结束

12.25.9.3 嵌入式迁移 Storm 业务

操作场景

该任务指导用户通过嵌入式迁移的方式在Flink的DataStream中嵌入Storm的代码，如使用Storm API编写的Spout/Bolt。

操作步骤

步骤1 在Flink中，对Storm拓扑中的Spout和Bolt进行嵌入式转换，将之转换为Flink的Operator，代码示例如下：

```
//set up the execution environment
final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();

//get input data
final DataStream<String> text = getTextDataStream(env);

final DataStream<Tuple2<String, Integer>> counts = text

  //split up the lines in pairs (2-tuples) containing: (word,1)
  //this is done by a bolt that is wrapped accordingly
  .transform("CountBolt",
    TypeExtractor.getForObject(new Tuple2<String, Integer>("", 0)),
    new BoltWrapper<String, Tuple2<String, Integer>>(new CountBolt()))
  //group by the tuple field "0" and sum up tuple field "1"
  .keyBy(0).sum(1);
// execute program
env.execute("Streaming WordCount with bolt tokenizer");
```

步骤2 修改完成后使用Flink命令进行提交。

flink run -class {MainClass} WordCount.jar

----结束

12.25.9.4 迁移 Storm 对接的外部安全组件业务

迁移 Storm 对接 HDFS 和 HBase 组件的业务

如果Storm的业务使用的storm-hdfs或者storm-hbase插件包进行的对接，那么在按照**完整迁移Storm业务**进行迁移时，需要指定特定安全参数，如下：

```
//初始化Storm的Config
Config conf = new Config();

//初始化安全插件列表
List<String> auto_tgts = new ArrayList<String>();
//添加AutoTGT插件
auto_tgts.add("org.apache.storm.security.auth.kerberos.AutoTGT");
//添加AutoHDFS插件
//如果对接HBase，则如下更改为：auto_tgts.add("org.apache.storm.hbase.security.AutoHBase");
auto_tgts.add("org.apache.storm.hdfs.common.security.AutoHDFS");

//设置安全参数
conf.put(Config.TOPOLOGY_AUTO_CREDENTIALS, auto_tgts);
//设置worker个数
conf.setNumWorkers(3);

//将Storm的Config转化为Flink的StormConfig
StormConfig stormConfig = new StormConfig(conf);

//使用Storm的TopologBuilder构造FlinkTopology
FlinkTopology topology = FlinkTopology.createTopology(builder);

//获取StreamExecutionEnvironment
StreamExecutionEnvironment env = topology.getExecutionEnvironment();

//将StormConfig设置到Job的环境变量中，用于构造Bolt和Spout
//如果Bolt和Spout初始化时不需要config，则不用设置
env.getConfig().setGlobalJobParameters(stormConfig);

//执行拓扑提交
topology.execute();
```

增加如上的安全插件配置后，可以避免HDFSBolt和HBaseBolt在初始化过程中的无谓登录，因为Flink已经实现准备好了安全上下文，无需再登录。

迁移 Storm 对接其他安全组件的业务

如果Storm的业务使用的storm-kakfa-client和storm-solr等插件包进行的对接时，需要注意，之前所配置的安全插件需要去掉，如下：

```
List<String> auto_tgts = new ArrayList<String>();
//keytab方式
auto_tgts.add("org.apache.storm.security.auth.kerberos.AutoTGTFromKeytab");

//将客户端配置的plugin列表写入config指定项中
//安全模式必配
//普通模式不用配置，请注释掉该行
conf.put(Config.TOPOLOGY_AUTO_CREDENTIALS, auto_tgts);
```

如上所配置的AutoTGTFromKeytab插件在进行业务迁移时，必须删除，否则会引起相应Bolt或Spout初始化时登录异常。

12.25.10 Storm 日志介绍

本章节内容适用于MRS 3.x及后续版本。

日志描述

日志路径：Storm相关日志的默认存储路径为“/var/log/Bigdata/storm/*角色名*”（运行日志），“/var/log/Bigdata/audit/storm/*角色名*”（审计日志）。

- Nimbus：“/var/log/Bigdata/storm/nimbus”（运行日志），“/var/log/Bigdata/audit/storm/nimbus”（审计日志）
- Supervisor：“/var/log/Bigdata/storm/supervisor”（运行日志），“/var/log/Bigdata/audit/storm/supervisor”（审计日志）
- UI：“/var/log/Bigdata/storm/ui”（运行日志），“/var/log/Bigdata/audit/storm/ui”（审计日志）
- Logviewer：“/var/log/Bigdata/storm/logviewer”（运行日志），“/var/log/Bigdata/audit/storm/logviewer”（审计日志）

日志归档规则：Storm的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过10MB的时候会自动压缩，压缩后的日志文件名规则为：“<原有日志名>.log.[编号].gz”。默认最多保留最近的20个压缩文件，压缩文件保留个数和压缩文件阈值可以配置。

审计日志压缩后的日志文件名规则为：“audit.log.[yyyy-MM-dd].[编号].zip”。该文件永远都不会删除。

表 12-436 Storm 日志列表

日志类型	日志文件名	描述
运行日志	nimbus/access.log	Nimbus用户访问日志。
	nimbus/nimbus-<PID>-gc.log	Nimbus进程的GC日志。
	nimbus/checkavailable.log	Nimbus可用性检查日志。
	nimbus/checkService.log	Nimbus可服务性检查日志。
	nimbus/metrics.log	Nimbus监控统计的日志。
	nimbus/nimbus.log	Nimbus进程运行日志。
	nimbus/postinstall.log	Nimbus安装后的工作日志。
	nimbus/prestart.log	Nimbus启动前的工作日志。
	nimbus/start.log	Nimbus启动的工作日志。
	nimbus/stop.log	Nimbus停止的工作日志。
	supervisor/access.log	Supervisor用户访问日志。
	supervisor/metrics.log	Supervisor监控统计的日志。

日志类型	日志文件名	描述
	supervisor/postinstall.log	Supervisor安装后的工作日志。
	supervisor/prestart.log	Supervisor启动前的工作日志。
	supervisor/start.log	Supervisor启动的工作日志。
	supervisor/stop.log	Supervisor停止的工作日志。
	supervisor/supervisor.log	Supervisor进程运行日志。
	supervisor/supervisor-<PID>-gc.log	Supervisor进程的GC日志。
	ui/access.log	UI用户访问日志。
	ui/metric.log	UI监控统计的日志。
	ui/ui-<PID>-gc.log	UI进程的GC日志。
	ui/postinstall.log	UI安装后的工作日志。
	ui/prestart.log	UI启动前的工作日志。
	ui/start.log	UI启动的工作日志。
	ui/stop.log	UI停止的工作日志。
	ui/ui.log	UI进程运行日志。
	logviewer/access.log	Logviewer用户访问日志。
	logviewer/metric.log	Logviewer监控统计的日志。
	logviewer/logviewer-<PID>-gc.log	Logviewer进程的GC日志。
	logviewer/logviewer.log	logviewer运行日志。
	logviewer/postinstall.log	logviewer安装后的工作日志。
	logviewer/prestart.log	logviewer启动前的工作日志。
	logviewer/start.log	logviewer启动的工作日志。
	logviewer/stop.log	logviewer停止的工作日志。

日志类型	日志文件名	描述
	supervisor/[topologyId]-worker-[端口号].log	Worker进程运行日志，一个端口占用一个日志文件，系统默认包含29100,29101,29102,29103,29304五个端口。
	supervisor/metadata/[topologyid]-worker-[端口号].yaml	worker日志元数据文件，logviewer在清理日志的时候会以该文件来作为清理依据。该文件会被logviewer日志清理线程根据一定条件自动删除。
	nimbus/cleanup.log	Nimbus卸载的清理日志。
	logviewer/cleanup.log	logviewer卸载的清理日志。
	ui/cleanup.log	UI卸载的清理日志。
	supervisor/cleanup.log	Supervisor卸载的清理日志。
	leader_switch.log	Storm主备倒换运行日志。
审计日志	nimbus/audit.log	Nimbus审计日志。
	ui/audit.log	UI审计日志。
	supervisor/audit.log	Supervisor审计日志。
	logviewer/audit	Logviewer审计日志。

日志级别

Storm提供了如**表12-437**所示的日志级别。

运行日志和审计日志的级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-437 日志级别

级别	描述
ERROR	ERROR表示系统运行的错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。

级别	描述
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 请参考**修改集群服务配置参数**，进入Storm的“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 保存配置，在弹出窗口中单击“确定”使配置生效。

----结束

日志格式

Storm的日志格式如下所示：

表 12-438 日志格式

日志类型	格式	示例
运行日志	%d{yyyy-MM-dd HH:mm:ss,SSS} | %-5p | [%t] | %m | %logger (%F:%L) %n	2015-03-11 23:04:00,241 | INFO | [RMI TCP Connection(2646)-10.0.0.2] | The baseSleepTimeMs [1000] the maxSleepTimeMs [1000] the maxRetries [1] | backtype.storm.utils.StormBoundedExponentialBackoffRetry (StormBoundedExponentialBackoffRetry.java:46)
	<yyyy-MM-dd HH:mm:ss,SSS><HostName><RoleName><logLevel><Message>	2017-03-28 02:57:52 493 10-5-146-1 storm- INFO Nimbus start normally
审计日志	<用户名><用户IP><时间><操作><操作对象><操作结果>	UserName=storm/hadoop, UserIP=10.10.0.2, Time=Tue Mar 10 01:15:35 CST 2015, Operation=Kill, Resource=test, Result=Success

12.25.11 性能调优

12.25.11.1 Storm 性能调优

操作场景

通过调整Storm参数设置，可以提升特定业务场景下Storm的性能。

本章节适用于MRS 3.x及后续版本。

修改服务配置参数，请参考**修改集群服务配置参数**。

拓扑调优

当需要提升Storm数据量处理性能时，可以通过拓扑调优的操作提高效率。建议在可靠性要求不高的场景下进行优化。

表 12-439 调优参数

配置参数	默认值	调优场景
topology.acker.executors	null	Acker的执行器数量。当业务应用对可靠性要求较低，允许不处理部分数据，可设置参数值为“null”或“0”，以关闭Acker的执行器，减少流控制，不统计消息时延，提高性能。
topology.max.spout.pending	null	Spout消息缓存数，仅在Acker不为0或者不为null的情况下生效。Spout将发送到下游Bolt的每条消息加入到pending队列，待下游Bolt处理完成并确认后，再从pending队列移除，当pending队列占满时Spout暂停消息发送。增加pending值可提高Spout的每秒消息吞吐量，提高性能，但延时同步增加。
topology.transfer.buffer.size	32	每个worker进程Distuptor消息队列大小，建议在4到32之间，增大消息队列可以提升吞吐量，但延时可能会增加。
RES_CPUSET_PERCENTAGE	80	设置各个节点上的Supervisor角色实例（包含其启动并管理的Worker进程）所使用的物理CPU百分比。根据Supervisor所在节点业务量需求，适当调整参数值，优化CPU使用率。

JVM 调优

当应用程序需要处理大量数据从而占用更多的内存时，存在worker内存大于2GB的情况，推荐使用G1垃圾回收算法。

表 12-440 调优参数

配置参数	缺省值	调优场景
WORKER_GC_OPTS	-Xms1G -Xmx1G -XX:+UseG1GC -XX:+PrintGCDetails -Xloggc:artifacts/gc.log -XX:+PrintGCDateStamps -XX:+PrintGCTimeStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -XX:+HeapDumpOnOutOfMemoryError -XX:HeapDumpPath=artifacts/heapdump	应用程序内存中需要保存大量数据，worker进程使用的内存大于2G，那么建议使用G1垃圾回收算法，可修改参数值为“-Xms2G -Xmx5G -XX:+UseG1GC”。

12.26 使用 Tez

12.26.1 使用前须知

本章节适用于MRS 3.x及后续版本。

12.26.2 Tez 常用参数

参数入口

在Manager系统中，选择“集群 > 服务 > Tez > 配置”，选择“全部配置”。在搜索框中输入参数名称。

参数说明

表 12-441 参数说明

配置参数	说明	缺省值
property.tez.log.dir	Tez日志目录。	/var/log/Bigdata/tez/tezui
property.tez.log.level	Tez的日志级别。	INFO

12.26.3 访问 TezUI

Tez提供Tez任务执行过程图形化展示功能，使用户可以通过界面的方式查看Tez任务执行细节。

前提条件

已安装Yarn服务的TimelineServer实例。

使用介绍

登录Manager系统，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，在Manager界面选择“集群 > 服务 > Tez”，在“基本信息”中单击“Tez WebUI”右侧的链接，打开Tez WebUI。可查看执行的Tez任务执行细节。

12.26.4 日志介绍

日志描述

日志路径：Tez相关日志的默认存储路径为“/var/log/Bigdata/tez/*角色名*”。

TezUI：“/var/log/Bigdata/tez/tezui”（运行日志），“/var/log/Bigdata/audit/tez/tezui”（审计日志）。

日志归档规则：Tez的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过20MB的时候（此日志文件大小可进行配置），会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件，压缩文件保留个数和压缩文件阈值可以配置。

表 12-442 Tez 日志列表

日志类型	日志文件名	描述
运行日志	tezui.out	TezUI运行环境信息日志
	tezui.log	TezUI进程的运行日志
	tezui-omm-<日期>-gc.log.<编号>	TezUI进程的GC日志
	prestartDetail.log	TezUI启动前的工作日志

日志类型	日志文件名	描述
	check-serviceDetail.log	TezUI服务启动是否成功的检查日志
	postinstallDetail.log	TezUI安装后的工作日志
	startDetail.log	TezUI进程启动日志
	stopDetail.log	TezUI进程停止日志
审计日志	tezui-audit.log	TezUI审计日志

日志级别

TezUI提供了如表12-443所示的日志级别。

运行日志的级别优先级从高到低分别是ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-443 日志级别

级别	描述
ERROR	ERROR表示系统运行的错误信息。
WARN	WARN表示当前事件处理存在异常信息。
INFO	INFO表示记录系统及各事件正常运行状态信息。
DEBUG	DEBUG表示记录系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 登录Manager。

步骤2 选择“集群 > 服务 > Tez > 配置”。

步骤3 选择“全部配置”。

步骤4 左边菜单栏中选择“TezUI > 日志”。

步骤5 选择所需修改的日志级别。

步骤6 单击“保存”，在弹出窗口中单击“确定”保存配置。

步骤7 单击“实例”，勾选“TezUI”角色，选择“更多 > 重启实例”，输入用户密码后，在弹出窗口单击“确定”。

步骤8 等待实例重启完成，配置生效。

----结束

日志格式

Tez的日志格式如下所示：

表 12-444 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<LogLevel>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-07-31 11:44:21,378 | INFO | TezUI-health-check | Start health check | com.XXX.tez.HealthCheck.run(HealthCheck.java:30)
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<LogLevel>|<产生该日志的线程名字>|<User Name><User IP><Time><Operation><Resource><Result><Detail >|<日志事件的发生位置>	2018-12-24 12:16:25,319 | INFO | HiveServer2-Handler-Pool: Thread-185 | UserName=hive UserIP=10.153.2.204 Time=2018/12/24 12:16:25 Operation=CloseSession Result=SUCCESS Detail= | org.apache.hive.service.cli.thrift.ThriftCLIService.logAuditEvent(ThriftCLIService.java:434)

12.26.5 常见问题

12.26.5.1 TezUI 无法展示 Tez 任务执行细节

问题

登录Manager界面，跳转Tez WebUI界面，已经提交的Tez任务未展示，如何解决。

回答

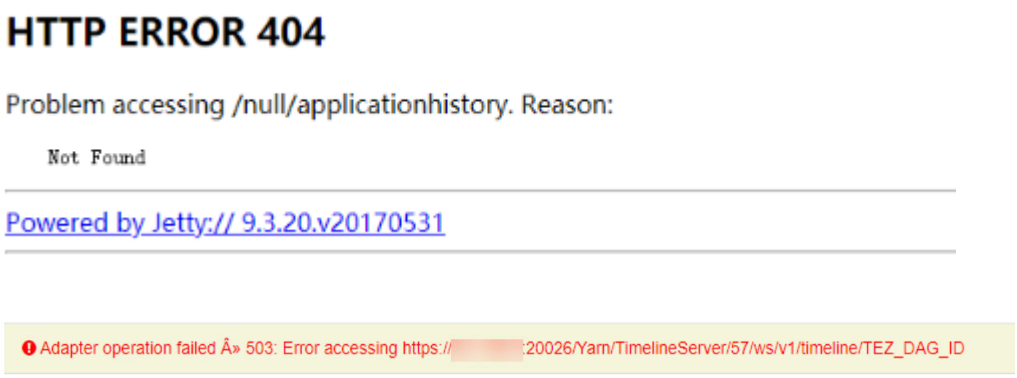
Tez WebUI展示的Tez任务数据，需要Yarn的TimelineServer支持，确认提交任务之前TimelineServer已经开启且正常运行。

在设置Hive执行引擎为Tez的同时，需要设置参数“yarn.timeline-service.enabled”为“true”，详情请参考**切换Hive执行引擎为Tez**。

12.26.5.2 进入 Tez 原生界面显示异常

问题

登录Manager界面，跳转Tez WebUI界面，显示404异常或503异常。



回答

Tez WebUI依赖Yarn的TimelineServer实例，需要预先安装TimelineServer，且处于良好状态。

12.26.5.3 TezUI 界面无法查看 yarn 日志

问题

登录Tez WebUI界面，单击Logs跳转yarn日志界面失败，无法加载数据。

回答

Tez WebUI跳转Yarn Logs界面时，目前是通过hostname进行访问，需要在windows机器，配置hostname到ip的映射。具体方法为：

修改windows机器C:\Windows\System32\drivers\etc\hosts文件，增加一行hostname到ip的映射, 例：10.244.224.45 10-044-224-45，保存后重新访问正常。

12.26.5.4 TezUI HiveQueries 界面表格数据为空

问题

登录Manager界面，跳转Tez WebUI界面，已经提交的任务，Hive Queries界面未展示数据，如何解决。

回答

Tez WebUI展示的Hive Queries任务数据，需要设置以下3个参数：

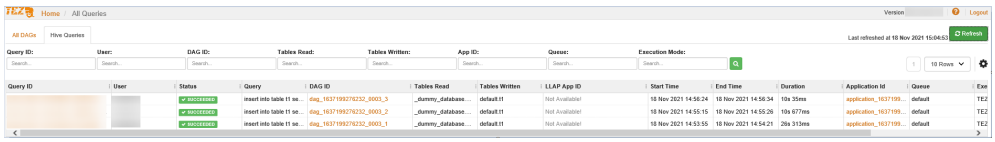
在FusionInsight Manager页面，选择“集群 > 服务 > Hive > 配置 > 全部配置 > HiveServer > 自定义”，在hive-site.xml中增加以下配置：

属性名	属性值
hive.exec.pre.hooks	org.apache.hadoop.hive.ql.hooks.ATSHook
hive.exec.post.hooks	org.apache.hadoop.hive.ql.hooks.ATSHook
hive.exec.failure.hooks	org.apache.hadoop.hive.ql.hooks.ATSHook

说明

TezUI数据展示依赖于Yarn组件的TimelineServer实例，如果TimelineServer实例故障或未启动，需设置hive自定义参数yarn-site.xml中**yarn.timeline-service.enabled**=false，否则hive任务会执行失败。

参数设置完成后，Hive Queries界面即可展示数据，但无法展示历史数据，展示效果如下:



12.27 使用 Yarn

12.27.1 Yarn 常用参数

队列资源分配

Yarn服务提供队列给用户使用，用户分配对应的系统资源给各队列使用。完成配置后，您可以单击“刷新队列”按钮或者重启Yarn服务使配置生效。

参数入口：

MRS 3.x之前的版本集群执行以下操作：

用户在MRS控制台上，选择“租户管理 > 资源分布策略”。

参数说明以default为例，其他队列的配置类似，单击“修改”编辑。

表 12-445 参数说明

配置参数	说明	默认值
资源容量	队列的资源容量（百分比）。当系统非常繁忙时，应保证每个队列的容量得到满足，而如果每个队列应用程序较少，可将剩余资源共享给其他队列。注意，所有队列的容量之和应小于100。	20
最大资源容量	队列的资源使用上限（百分比）。由于存在资源共享，因此一个队列使用的资源量可能超过其容量，而最多使用资源量可通过该参数限制。	100

MRS 3.x及后续版本集群执行以下操作：

用户可在Manager系统中，选择“租户资源 > 动态资源计划 > 队列配置”。

参数说明以修改Superior调度器的default租户为例，其他队列的配置类似，单击“修改”编辑。

表 12-446 队列配置参数

参数名	描述
AM最多占有资源（%）	表示当前队列内所有Application Master所占的最大资源百分比。
每个YARN容器最多分配核数	表示当前队列内单个YARN容器可分配的最多核数，默认为-1，表示取值范围内不限制。
每个YARN容器最大分配内存（MB）	表示当前队列内单个YARN容器可分配的最大内存，默认为-1，表示取值范围内不限制。
最多运行任务数	表示当前队列最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
每个用户最多运行任务数	表示每个用户在当前队列中最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
最多挂起任务数	表示当前队列最多同时可挂起任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可挂起任务。取值范围为-1～2147483647。

参数名	描述
资源分配规则	表示单个用户任务间的资源分配规则，包括FIFO和FAIR。 一个用户若在当前队列上提交了多个任务，FIFO规则代表一个任务完成后再执行其他任务，按顺序执行。FAIR规则代表各个任务同时获取到资源并平均分配资源。
默认资源标签	表示在指定资源标签（Label）的节点上执行任务。
Active状态	• ACTIVE表示当前队列可接受并执行任务。 • INACTIVE表示当前队列可接受但不执行任务，若提交任务，任务将处于挂起状态。
Open状态	• OPEN表示当前队列处于打开状态。 • CLOSED表示当前队列处于关闭状态，若提交任务，任务直接会被拒绝。

在 UI 显示 container 日志

默认情况下，系统会将container日志收集到HDFS中。如果您不需要将container日志收集到HDFS中，可以配置参数见**表12-447**。具体配置操作请参考**修改集群服务配置参数**。

表 12-447 参数说明

配置参数	说明	默认值
yarn.log-aggregation-enable	设置是否将container日志收集到HDFS中。 • 设置为true，表示日志会被收集到HDFS目录中。默认目录为“{yarn.nodemanager.remote-app-log-dir}/${user}/{thisParam}”，该路径可通过界面上的“yarn.nodemanager.remote-app-log-dir-suffix”参数进行配置。 • 设置为false，表示日志不会收集到HDFS中。 修改参数值后，需重启Yarn服务使其生效。 **说明** 在修改值为false并生效后，生效前的日志无法在UI中获取。您可以在“yarn.nodemanager.remote-app-log-dir-suffix”参数指定的路径中获取到生效前的日志。 如果需要在UI上查看之前产生的日志，建议将此参数设置为true。	true

在 WebUI 显示更多历史作业

默认情况下，Yarn WebUI界面支持任务列表分页功能，每个分页最多显示5000条历史作业，总共最多保留10000条历史作业。如果您需要在WebUI上查看更多的作业，可以配置参数如**表12-448**。具体配置操作请参考**修改集群服务配置参数**。

表 12-448 参数说明

配置参数	说明	默认值
yarn.resourcemanager.max-completed-applications	设置在WebUI总共显示的历史作业数量。	10000
yarn.resourcemanager.webapp.pagination.enable	是否开启Yarn WebUI的任务列表后台分页功能。	true
yarn.resourcemanager.webapp.pagination.threshold	开启Yarn WebUI的任务列表后台分页功能后，每个分页显示的最大作业数量。	5000

说明

- 显示更多的历史作业，会影响性能，增加打开Yarn WebUI的时间，建议开启后台分页功能，并根据实际硬件性能修改“yarn.resourcemanager.max-completed-applications”参数。
- 修改参数值后，需重启Yarn服务使其生效。

12.27.2 创建 Yarn 角色

操作场景

该任务指导系统管理员创建并设置Yarn的角色。Yarn角色可设置Yarn管理员权限以及Yarn队列资源管理。

说明

如果当前组件使用了Ranger进行权限控制，须基于Ranger配置相关策略进行权限管理。对于MRS 3.x及后续版本集群，具体操作可参考**添加Yarn的Ranger访问权限策略**。

前提条件

- 系统管理员已明确业务需求。
- 登录Manager。

操作步骤

MRS 3.x以前版本集群执行以下操作：

步骤1 选择“系统设置 > 角色管理 > 添加角色”。

步骤2 在“角色名称”和“描述”输入角色名字与描述。

步骤3 设置角色“权限”请参见**表12-449**。

Yarn权限：

- “Cluster Admin Operations”：Yarn管理员权限。
- “Scheduler Queue”：队列资源管理。

表 12-449 设置角色

任务场景	角色授权操作
设置Yarn管理员权限	在“权限”的表格中选择“Yarn”，勾选“Cluster Admin Operations”。 **说明** 设置Yarn管理员权限需要重启Yarn服务，才能使保存的角色配置生效。
设置用户在指定Yarn队列提交任务的权限	1. 在“权限”的表格中选择“Yarn > Scheduler Queue”。 2. 在指定队列的“权限”列，勾选“Submit”。
设置用户在指定Yarn队列管理任务的权限	1. 在“权限”的表格中选择“Yarn > Scheduler Queue”。 2. 在指定队列的“权限”列，勾选“Admin”。

如果Yarn角色包含了某个父队列的“提交”或“管理”权限，则角色默认子队列也继承此权限，将自动添加子队列的“提交”或“管理”权限。子队列继承的权限不在“配置资源权限”表格显示被选中。

如果设置Yarn角色时仅勾选到某个父队列的“提交”权限，使用拥有该角色权限的用户提交任务时，注意需要手动指定队列名称，否则当父队列下有多个子队列时，系统并不会自动判断，从而将任务提交到了“default”队列。

步骤4 单击“确定”完成。

----结束

MRS 3.x及以后版本集群执行以下操作：

步骤1 选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，然后“角色名称”和“描述”输入角色名字与描述。

步骤3 设置角色“配置资源权限”请参见**表12-450**。

Yarn权限：

- “集群管理操作权限”：Yarn管理员权限。
- “调度队列”：队列资源管理。

表 12-450 设置角色

任务场景	角色授权操作
设置Yarn管理员权限	在“配置资源权限”的表格中选择“*待操作集群的名称* > Yarn”，勾选“集群管理操作权限”。 **说明** 设置Yarn管理员权限需要重启Yarn服务，才能使保存的角色配置生效。
设置用户在指定Yarn队列提交任务的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Yarn > 调度队列 > root”。 2. 在指定队列的“权限”列，勾选“提交”。
设置用户在指定Yarn队列管理任务的权限	1. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Yarn > 调度队列 > root”。 2. 在指定队列的“权限”列，勾选“管理”。

如果Yarn角色包含了某个父队列的“提交”或“管理”权限，则角色默认子队列也继承此权限，将自动添加子队列的“提交”或“管理”权限。子队列继承的权限不在“配置资源权限”表格显示被选中。

如果设置Yarn角色时仅勾选到某个父队列的“提交”权限，使用拥有该角色权限的用户提交任务时，注意需要手动指定队列名称，否则当父队列下有多个子队列时，系统并不会自动判断，从而将任务提交到了“default”队列。

步骤4 单击“确定”完成。

----结束

12.27.3 使用 Yarn 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Yarn客户端。

前提条件

- 已安装客户端。

  例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。普通模式不需要下载keytab文件及修改密码操作。

使用 Yarn 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行Yarn命令。例如：

yarn application -list

----结束

客户端常见使用问题

1. 当执行Yarn客户端命令时，客户端程序异常退出，报“java.lang.OutOfMemoryError”的错误。

   这个问题是由于Yarn客户端运行时的所需的内存超过了Yarn客户端设置的内存上限（默认为128MB）。对于MRS 3.x后续版本集群，可以通过修改“<客户端安装路径>/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改Yarn客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   export CLIENT_GC_OPTS="-Xmx1G"
   ```

   对于MRS 3.x之前版本集群，可以通过修改“<客户端安装路径>/HDFS/component_env”中的“GC_OPTS_YARN”来修改Yarn客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   export GC_OPTS_YARN="-Xmx1G"
   ```

   在修改完后，使用如下命令刷新客户端配置，使之生效：

   source <*客户端安装路径*>/**bigdata_env**

2. 如何设置Yarn客户端运行时的日志级别?

   Yarn客户端运行时的日志是默认输出到Console控制台的，其级别默认是INFO级别。有的时候为了定位问题，需要开启DEBUG级别日志，可以通过导出一个环境变量来设置，命令如下：

   export YARN_ROOT_LOGGER=DEBUG,console

   在执行完上面命令后，再执行Yarn Shell命令时，即可打印出DEBUG级别日志。

   如果想恢复INFO级别日志，可执行如下命令：

   export YARN_ROOT_LOGGER=INFO,console

12.27.4 配置 NodeManager 角色实例使用的资源

操作场景

如果部署NodeManager的各个节点硬件资源（如CPU核数、内存总量）不一样，而NodeManager可用硬件资源设置为相同的值，可能造成性能浪费或状态异常，需要修改各个NodeManager角色实例的配置，使硬件资源得到充分利用。

对系统的影响

保存新的配置需要重启NodeManager角色实例，此时对应的角色实例不可用。

前提条件

- MRS 3.x之前的版本集群：已登录MRS控制台。
- MRS 3.x及后续版本集群：已登录Manager。

操作步骤

MRS 3.x之前的版本集群执行以下操作：

步骤1 选择“集群列表 > 现有集群”，单击集群名称。选择“组件管理 > Yarn > 实例”。

步骤2 单击“角色”列“NodeManager”角色实例名称，并切换到“实例配置”。单击“基础配置”下拉菜单，选择“全部配置”，在搜索框中输入以下参数。

步骤3 “yarn.nodemanager.resource.cpu-vcores”设置当前节点上NodeManager可使用的虚拟CPU核数，建议按节点实际逻辑核数的1.5到2倍配置。
“yarn.nodemanager.resource.memory-mb”设置当前节点上NodeManager可使用的物理内存大小，建议按节点实际物理内存大小的75%~90%配置。

说明

“yarn.scheduler.maximum-allocation-vcores”可配置单个Container最多CPU可用核数，“yarn.scheduler.maximum-allocation-mb”可配置单个Container最大内存可用值。不支持实例级别的修改，需要在Yarn服务的配置中修改参数值，并重启Yarn服务。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例”，单击“确定”。重启NodeManager角色实例。

界面提示“操作成功。”，单击“完成”，NodeManager角色实例成功启动。

----结束

MRS 3.x及后续版本集群也可执行以下操作：

步骤1 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”。

步骤2 单击部署NodeManager节点对应角色实例名称，并切换到“实例配置”，选择“全部配置”。

步骤3 “yarn.nodemanager.resource.cpu-vcores”设置当前节点上NodeManager可使用的虚拟CPU核数，建议按节点实际逻辑核数的1.5到2倍配置。
“yarn.nodemanager.resource.memory-mb”设置当前节点上NodeManager可使用的物理内存大小，建议按节点实际物理内存大小的75%配置。

说明

“yarn.scheduler.maximum-allocation-vcores”可配置单个Container最多CPU可用核数，“yarn.scheduler.maximum-allocation-mb”可配置单个Container最大内存可用值。不支持实例级别的修改，需要在Yarn服务的配置中修改参数值，并重启Yarn服务。

步骤4 单击“保存”，单击“确定”。重启NodeManager角色实例。

界面提示“操作成功”，单击“完成”，NodeManager角色实例成功启动。

----结束

12.27.5 更改 NodeManager 的存储目录

操作场景

Yarn NodeManager定义的存储目录不正确或Yarn的存储规划变化时，系统管理员需要在Manager中修改NodeManager的存储目录，以保证Yarn正常工作。NodeManager的存储目录包含本地存放目录“yarn.nodemanager.local-dirs”和日志目录“yarn.nodemanager.log-dirs”。适用于以下场景：

- 更改NodeManager角色的存储目录，所有NodeManager实例的存储目录将同步修改。
- 更改NodeManager单个实例的存储目录，只对单个实例生效，其他节点NodeManager实例存储目录不变。

对系统的影响

- 更改NodeManager角色的存储目录需要停止并重新启动集群，集群未启动前无法提供服务。
- 更改NodeManager单个实例的存储目录需要停止并重新启动实例，该节点NodeManager实例未启动前无法提供服务。
- 服务参数配置如果使用旧的存储目录，需要更新为新目录。
- 更改NodeManager的存储目录以后，需要重新下载并安装客户端。

前提条件

- 在各个数据节点准备并安装好新磁盘，并格式化磁盘。
- 规划好新的目录路径，用于保存旧目录中的数据。
- 准备好系统管理员用户**admin**。

操作步骤

MRS 3.x之前的版本集群执行以下操作：

步骤1 检查环境。

1. 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。选择“组件管理”，查看Yarn的“健康状态”是否为“良好”。
   - 是，执行**步骤1.3**。
   - 否，Yarn状态不健康，执行**步骤1.2**。
2. 请先修复Yarn异常，任务结束。
3. 确定修改NodeManager的存储目录场景。
   - 更改NodeManager角色的存储目录，执行**步骤2**。
   - 更改NodeManager单个实例的存储目录，执行**步骤3**。

步骤2 更改NodeManager角色的存储目录。

1. 选择“集群列表 > 现有集群”，单击集群名称。选择“组件管理 > Yarn > 停止”，停止Yarn服务。
2. 登录弹性云服务器，以**root**用户登录到安装Yarn服务的各个节点中，执行如下操作。

a. 创建目标目录。

例如目标目录为“${BIGDATA_DATA_HOME}/data2”：

执行**mkdir ${BIGDATA_DATA_HOME}/data2**

b. 挂载目标目录到新磁盘。

例如挂载“${BIGDATA_DATA_HOME}/data2”到新磁盘。

c. 修改新目录的权限。

例如新目录路径为“${BIGDATA_DATA_HOME}/data2”：

执行**chmod 750 ${BIGDATA_DATA_HOME}/data2 -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/data2 -R**

3. 在MRS控制台界面，选择“集群列表 > 现有集群”，单击集群名称。选择“组件管理 > Yarn > 实例”，选择对应主机的NodeManager实例，单击“实例配置”，“选择”“全部配置”。

将配置项“yarn.nodemanager.local-dirs”或“yarn.nodemanager.log-dirs”修改为新的目标目录。

例如：如果修改“yarn.nodemanager.local-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/localdir”。如果修改“yarn.nodemanager.log-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/containerlogs”。

4. 单击“保存配置”，勾选“重新启动受影响的服务或实例”，单击“确定”。重启Yarn服务。

界面提示“操作成功”，单击“完成”，Yarn成功启动，任务结束。

步骤3 更改NodeManager单个实例的存储目录。

1. 选择“集群列表 > 现有集群”，单击集群名称。选择“组件管理 > Yarn > 实例”，勾选需要修改存储目录的NodeManager单个实例，选择“更多 > 停止实例”。
2. 登录弹性云服务器，以**root**用户登录到这个NodeManager节点，执行如下操作。

a. 创建目标目录。

例如目标目录为“${BIGDATA_DATA_HOME}/data2”：

执行**mkdir ${BIGDATA_DATA_HOME}/data2**。

b. 挂载目标目录到新磁盘。

例如挂载“${BIGDATA_DATA_HOME}/data2”到新磁盘。

c. 修改新目录的权限。

例如新目录路径为“${BIGDATA_DATA_HOME}/data2”：

执行**chmod 750 ${BIGDATA_DATA_HOME}/data2 -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/data2 -R**。

3. 在MRS控制台，单击指定的NodeManager实例并切换到“实例配置”。

将配置项“yarn.nodemanager.local-dirs”或“yarn.nodemanager.log-dirs”修改为新的目标目录。

例如：如果修改“yarn.nodemanager.local-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/localdir”。如果修改“yarn.nodemanager.log-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/containerlogs”。

4. 单击“保存配置”，勾选“重新启动受影响的服务或实例”。单击“确定”。重启NodeManager实例。

界面提示“操作成功”，单击“完成”，NodeManager实例启动成功。

----结束

MRS 3.x及后续版本集群也可执行以下操作：

步骤1 检查环境。

1. 登录Manager，选择“集群 > *待操作集群的名称* > 服务”查看Yarn的状态“运行状态”是否为“良好”。
   - 是，执行**1.c**。
   - 否，Yarn状态不健康，执行**1.b**。
2. 修复Yarn异常，任务结束。
3. 确定修改NodeManager的存储目录场景。
   - 更改NodeManager角色的存储目录，执行**2**。
   - 更改NodeManager单个实例的存储目录，执行**3**。

步骤2 更改NodeManager角色的存储目录。

1. 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 停止服务”，停止Yarn服务。
2. 以**root**用户登录到安装Yarn服务的各个节点中，执行如下操作。
   a. 创建目标目录。
      例如目标目录为“${BIGDATA_DATA_HOME}/data2”：
      执行**mkdir ${BIGDATA_DATA_HOME}/data2**
   b. 挂载目标目录到新磁盘。
      例如挂载“${BIGDATA_DATA_HOME}/data2”到新磁盘。
   c. 修改新目录的权限。
      例如新目录路径为“${BIGDATA_DATA_HOME}/data2”：
      执行**chmod 750 ${BIGDATA_DATA_HOME}/data2 -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/data2 -R**
3. 在Manager管理界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”，选择对应主机的NodeManager实例，单击“实例配置”，选择“全部配置”。
   将配置项“yarn.nodemanager.local-dirs”或“yarn.nodemanager.log-dirs”修改为新的目标目录。
   例如：如果修改“yarn.nodemanager.local-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/localdir”。如果修改“yarn.nodemanager.log-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/containerlogs”。
4. 单击“保存”，单击“确定”。重启Yarn服务。
   界面提示“操作成功”，单击“完成”，Yarn成功启动，任务结束。

步骤3 更改NodeManager单个实例的存储目录。

1. 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”，勾选需要修改存储目录的NodeManager单个实例，选择“更多 > 停止实例”。
2. 以**root**用户登录到这个NodeManager节点，执行如下操作。
   a. 创建目标目录。
      例如目标目录为“${BIGDATA_DATA_HOME}/data2”：
      执行**mkdir ${BIGDATA_DATA_HOME}/data2**。
   b. 挂载目标目录到新磁盘。
      例如挂载“${BIGDATA_DATA_HOME}/data2”到新磁盘。

c. 修改新目录的权限。

例如新目录路径为“${BIGDATA_DATA_HOME}/data2”：

执行**chmod 750 ${BIGDATA_DATA_HOME}/data2 -R**和**chown omm:wheel ${BIGDATA_DATA_HOME}/data2 -R**。

3. 在Manager管理界面，单击指定的NodeManager实例并切换到“实例配置”。

将配置项“yarn.nodemanager.local-dirs”或“yarn.nodemanager.log-dirs”修改为新的目标目录。

例如：如果修改“yarn.nodemanager.local-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/localdir”。如果修改“yarn.nodemanager.log-dirs”参数，则将其值修改为“/srv/BigData/data2/nm/containerlogs”。

4. 单击“保存”，单击“确定”。重启NodeManager实例。

界面提示“操作成功”，单击“完成”，NodeManager实例启动成功。

----结束

12.27.6 配置 YARN 严格权限控制

配置场景

在安全模式的多租户场景下，一个集群可以支持多个用户使用以及支持多个用户任务提交、运行，用户之间是不可见，需要有一个权限控制机制，使用户的任务信息不被其他用户获取。

例如，用户A提交的应用正在运行，此时用户B登录系统并查看应用列表，用户B不应该访问到A用户的应用信息。

配置描述

- 查看Yarn服务配置参数

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入**表12-451**中参数名称。

表 12-451 参数描述

参数	描述	默认值
yarn.acl.enable	Yarn权限控制启用开关。	true
yarn.webapp.filter-entity-list-by-user	严格视图启用开关，开启后，登录用户只能查看该用户有权限查看的内容。当要开启该功能时，同时需要设置参数“yarn.acl.enable”为true。 **说明** 此参数适用于MRS 3.x及后续版本集群。	true

- 查看Mapreduce服务配置参数

参考**修改集群服务配置参数**进入Mapreduce服务参数“全部配置”界面，在搜索框中输入**表12-452**中参数名称。

表 12-452 参数描述

参数	描述	默认值
mapreduce.cluster.acls.enabled	MR JobHistoryServer权限控制启用开关。该参数为客户端参数，当JobHistoryServer服务端开启权限控制之后该参数生效。	true
yarn.webapp.filter-entity-list-by-user	MR JobHistoryServer严格视图启用开关，开启后，登录用户只能查看该用户有权限查看的内容。该参数为JobHistoryServer的服务端参数，表示JHS开启了权限控制，但是否要对某一个特定的Application进行控制，是由客户端参数：“mapreduce.cluster.acls.enabled”决定。 **说明** 此参数适用于MRS 3.x及后续版本集群。	true

须知

以上配置会影响restful API和shell命令结果，即以上配置开启后，restful API调用和shell命令运行所返回的内容只包含调用用户有权查看的信息。

当yarn.acl.enable或mapreduce.cluster.acls.enabled设置为false时，即关闭Yarn或Mapreduce的权限校验功能。此时任何用户都可以在Yarn或MapReduce上提交任务和查看任务信息，存在安全风险，请谨慎使用。

12.27.7 配置 Container 日志聚合功能

配置场景

YARN提供了Container日志聚合功能，可以将各节点Container产生的日志收集到HDFS，释放本地磁盘空间。日志收集的方式有两种：

- 应用完成后将Container日志一次性收集到HDFS。
- 应用运行过程中周期性收集Container输出的日志片段到HDFS。

配置描述

参数入口：

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入**表12-453**中参数名称。

其中“yarn.nodemanager.remote-app-log-dir-suffix”参数还需要在YARN的客户端进行配置，且在ResourceManager、NodeManager和JobHistory节点的配置与在YARN的客户端的配置必须一致。

周期性收集日志功能目前仅支持MapReduce应用，且MapReduce应用必须进行相应的日志文件滚动输出配置，需要在MapReduce客户端节点的“mapred-site.xml”配置文件中进行如**表12-455**所示的配置。

表 12-453 参数说明

参数	描述	默认值
yarn.log-aggregation-enable	设置是否打开Container日志聚合功能。 ● 设置为“true”，表示打开该功能，日志会被收集到HDFS目录中。 ● 设置为“false”，表示关闭该功能，表示日志不会收集到HDFS中。 修改参数值后，需重启YARN服务使其生效。 **说明** ● 在修改值为“false”并生效后，生效前的日志无法在WebUI中获取。 ● 如果需要在UI上查看之前产生的日志，建议将此参数设置为“true”。	true
yarn.nodemanager.log-aggregation.roll-monitoring-interval-seconds	NodeManager周期性日志收集的时间间隔。 ● 设置为-1或0时，表示周期性收集日志功能关闭，日志在应用运行完成后一次性收集。 ● 收集周期最小可设定为3600秒。当设置为大于0秒且小于3600秒时，收集周期将使用3600秒。 定义NodeManager唤醒并上传日志的间隔周期。设置为-1或0表示禁用滚动监控，应用任务结束后日志汇聚。取值范围大于等于-1。	-1

参数	描述	默认值
yarn.nodemanager.disk-health-checker.log-dirs.max-disk-utilization-per-disk-percentage	配置Container日志目录可以占用每块磁盘上YARN的磁盘配额的最大百分比。当日志目录占用空间超过此设定值时，将触发周期性日志收集服务启动一次周期外的日志收集活动，以释放本地磁盘空间。每个磁盘上可提供给Container logs的最大可使用率。当Container logs使用超过这个限制，会触发滚动汇聚。 ● 对于MRS 3.x之前的版本集群，磁盘配额最大百分比的有效取值范围为0～100，如果配置小于等于0，会被强制重置为25；如果配置大于100，则被强制重置为25。 ● 对于MRS 3.x及后续版本集群，磁盘配额最大百分比的有效取值范围为-1～100，如果配置小于-1，会被强制重置为25；如果配置大于100，则被强制重置为25。而配置为-1时则关闭Container日志目录的磁盘容量检测功能。 **说明** ● Container日志目录实际可用磁盘百分比=YARN磁盘可用百分比（“yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage”）* 日志目录可用百分比（“yarn.nodemanager.disk-health-checker.log-dirs.max-disk-utilization-per-disk-percentage”）。 ● 只有启用了周期性收集日志功能的应用才会在日志目录磁盘配额超过设定阈值时被触发启动日志收集。	25
yarn.nodemanager.remote-app-log-dir-suffix	设置HDFS用于存放Container日志的文件夹名称。该配置加上“yarn.nodemanager.remote-app-log-dir”，构成了Container日志的完整存放目录。目录为：“{yarn.nodemanager.remote-app-log-dir}/${user}/{yarn.nodemanager.remote-app-log-dir-suffix}”。 **说明** *{user}*为运行任务时的用户名。	logs
yarn.nodemanager.log-aggregator.on-fail.remain-log-in-sec	设置Container日志归集失败后日志在本地保留的时间。单位：秒。 ● 设置为0时，本地日志将马上删除。 ● 设置为正数时，表示本地日志将保留这段时间。	604800

参考**修改集群服务配置参数**进入Mapreduce服务参数“全部配置”界面，在搜索框中输入**表12-454**中参数名称。

表 12-454 参数说明

参数	描述	默认值
yarn.log-aggregation.retain-seconds	汇聚日志的保存时间。单位：秒。 ● 设置为-1时，表示HDFS上面的Container聚合日志将永久保留。 ● 设置为0或正数时，表示HDFS上面的Container聚合日志将保留这段时间，超时将被删除。 **说明** 当时间设置太短时，有可能会增加NameNode的负担，建议根据实际情况设置一个合理的时间值。	1296000
yarn.log-aggregation.retain-check-interval-seconds	设置扫描HDFS保存的Container聚合日志的间隔时间。单位：秒。 ● 设置为-1或0时，间隔时间将为“yarn.log-aggregation.retain-seconds”该配置时间的十分之一。 **说明** 当该配置设置为-1或0时，“yarn.log-aggregation.retain-seconds”不能设置为0。 ● 设置为正数时，将周期性的间隔这段时间以后对HDFS上的container聚合日志进行扫描。 **说明** 当时间设置太短时，有可能会增加NameNode的负担，建议根据实际情况设置一个合理的时间。	86400

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入**表12-455**中参数名称。

表 12-455 MapReduce 应用日志文件滚动输出配置

参数	描述	默认值
mapreduce.task.userlog.limit.kb	MR应用程序单个task日志文件大小限制。当日志文件达到该限制时，会新建一个日志文件进行输出。设置为“0”表示不限制日志文件大小。	51200

参数	描述	默认值
yarn.app.mapreduce.task.container.log.backups	MR应用程序task日志保留的最大个数。 设置为“0”表示不滚动输出。 使用CRLA（ContainerRollingLogAppender）时任务日志备份文件的数量。默认使用CLA（ContainerLogAppender）且container日志不回滚。 当mapreduce.task.userlog.limit.kb和yarn.app.mapreduce.task.container.log.backups都大于0时，任务启用CRLA。取值范围0~999。	10
yarn.app.mapreduce.am.container.log.limit.kb	MR应用程序单个AM日志文件大小限制。单位：KB，当日志文件达到该限制时，会新建一个日志文件进行输出。设置为“0”表示不限制单个AM日志文件大小。	51200
yarn.app.mapreduce.am.container.log.backups	MR应用程序AM日志保留的最大个数。设置为“0”表示不滚动输出。使用CRLA（ContainerRollingLogAppender）时ApplicationMaster日志备份文件的数量。默认使用CLA（ContainerLogAppender）且容器日志不回滚。 当yarn.app.mapreduce.am.container.log.limit.kb和yarn.app.mapreduce.am.container.log.backups都大于0时，ApplicationMaster启用CRLA。取值范围0~999。	20
yarn.app.mapreduce.shuffle.log.backups	MR应用程序shuffle日志保留的最大个数。设置为“0”表示不滚动输出。 当yarn.app.mapreduce.shuffle.log.limit.kb和yarn.app.mapreduce.shuffle.log.backups都大于0时，syslog.shuffle将采用CRLA。取值范围0~999。	10
yarn.app.mapreduce.shuffle.log.limit.kb	MR应用程序单个shuffle日志文件大小限制，单位KB。当日志文件达到该限制时，会新建一个日志文件进行输出。设置为“0”不限制单个shuffle日志文件大小。取值范围大于等于0。	51200

12.27.8 启用 CGroups 功能

本章节适用于MRS 3.x及后续版本集群。

配置场景

CGroups是一个Linux内核特性。它可以将任务集及其子集聚合或分离成具备特定行为的分层组。在YARN中，CGroups特性对容器（container）使用的资源（例如CPU使用率）进行限制。本特性大大降低了限制容器CPU使用的难度。

说明

当前CGroups仅用于限制CPU使用率。

配置描述

有关如何配置CPU隔离与安全的CGroups功能的详细信息，请参见Hadoop官网：**http://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/NodeManagerCgroups.html**

由于CGroups为Linux内核特性，是通过**LinuxContainerExecutor**进行开放。请参考官网资料对**LinuxContainerExecutor**进行安全配置。您可通过官网资料了解系统用户和用户组配置对应的文件系统权限。详情请参见：**http://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-common/SecureMode.html#LinuxContainerExecutor**

说明

- 请勿修改对应文件系统中各路径所属的用户、用户组及对应的权限，否则可能导致本功能异常。
- 当参数“yarn.nodemanager.resource.percentage-physical-cpu-limit”配置过小，导致可使用的核不足1个时，例如4核节点，将此参数设置为20%，不足1个核，那么将会使用系统全部的核。Linux的一些版本不支持Quota模式，例如Cent OS。在这种情况下，可以使用CPUset模式。

配置cpuset模式，即YARN只能使用配置的CPU，需要添加以下配置。

表 12-456 cpuset 配置

参数	描述	默认值
yarn.nodemanager.linux-container-executor.cgroups.cpu-set-usage	设置为“true”时，应用以cpuset模式运行。	false

配置strictcpuset模式，即container只能使用配置的CPU，需要添加以下配置。

表 12-457 CPU 硬隔离参数配置

参数	描述	默认值
yarn.nodemanager.linux-container-executor.cgroups.cpu-set-usage	设置为“true”时，应用以cpuset模式运行。	false

参数	描述	默认值
yarn.nodemanager.linux-container-executor.cgroups.cpuset.strict.enabled	设置为true时，container只能使用配置的CPU。	false

要从cpuset模式切换到Quota模式，必须遵循以下条件：

- 配置“yarn.nodemanager.linux-container-executor.cgroups.cpu-set-usage”=“false”。
- 删除container文件夹（如果存在）。
- 删除cpuset.cpus文件中设置的所有CPU。

操作步骤

步骤1 登录Manager系统。选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，选择“全部配置”。

步骤2 在左侧导航栏选择“NodeManager > 自定义”，找到yarn-site.xml文件。

步骤3 添加**表12-456**和**表12-457**中的参数为自定义参数。

根据配置文件与参数作用，在“yarn-site.xml”所在行“名称”列输入参数名，在“值”列输入此参数的参数值。

单击“+”增加自定义参数。

步骤4 单击“保存”，在弹出的“保存配置”窗口中确认修改参数，单击“确定”。界面提示“操作成功”，单击“完成”，配置保存成功。

保存完成后请重新启动配置过期的Yarn服务以使配置生效。

----结束

12.27.9 配置 AM 失败重试次数

配置场景

在资源不足导致ApplicationMaster启动失败的情况下，调整如下参数值，提高容错性，保证客户端应用的正常运行。

配置描述

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入**表12-458**中参数名称。

表 12-458 参数说明

参数	描述	默认值
yarn.resourcemanager.am.max-attempts	ApplicationMaster重试次数，增加重试次数，可以防止资源不足导致的AM启动失败问题。适用于所有ApplicationMaster的全局设置。每个ApplicationMaster都可以使用API设置一个单独的最大尝试次数，但这个次数不能大于全局的最大次数。如果大于了，那ResourceManager将会覆写这个单独的最大尝试次数。以允许至少一次重试。取值范围大于等于1。	5

12.27.10 配置 AM 自动调整分配内存

本章节适用于MRS 3.x及后续版本集群。

配置场景

启动该配置的过程中，ApplicationMaster在创建container时，分配的内存会根据任务总数的浮动自动调整，资源利用更加灵活，提高了客户端应用运行的容错性。

配置描述

参数入口：

在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，选择“全部配置”，在搜索框中输入参数名称“mapreduce.job.am.memory.policy”。

配置说明：

配置项的默认值为空，此时不会启动自动调整的策略，ApplicationMaster的内存仍受“yarn.app.mapreduce.am.resource.mb”配置项的影响。

配置参数的值由5个数值组成，中间使用“：”与“，”分隔，格式为：**baseTaskCount:taskStep:memoryStep,minMemory:maxMemory**，在键入时会严格校验格式。

表 12-459 配置数值说明

数值名称	描述	设定要求
baseTaskCount	任务总量基数，只有当应用的task总数（map端与reduce端之和）不小于该值时配置才会起作用	不能为空且大于零
taskStep	任务增量步进，与memoryStep共同决定内存调整量	不能为空且大于零
memoryStep	内存增量步进，在"yarn.app.mapreduce.am.resource.mb"配置的基础上对内存向上调整	不能为空且大于零，单位：MB

数值名称	描述	设定要求
minMemory	内存自动调整下限，若调整后的内存不大于该值，仍保持"yarn.app.mapreduce.am.resource.mb"的配置	不能为空且大于零，且不大于maxMemory的设定值 单位：MB
maxMemory	内存自动调整上限，若调整后的内存超过该值，则使用该值作为最终调整值	不能为空且大于零，且不小于minMemory的设定值 单位：MB

配置示例

配置情况：

- yarn.app.mapreduce.am.resource.mb=1536
- mapreduce.job.am.memory.policy=100:10:50,1200:2000
- 某应用task总数=120

计算过程：

调整后内存=1536+[（120-100）/10]*50=1636，满足1200<1636且2000>1636，最终ApplicationMaster内存会设定为1636MB。

若memStep修改为250，调整后内存=1536+[（120-100）/10]*250=2136，超过maxMemory=2000的限制，最终ApplicationMaster内存会设定为2000MB。

说明

对于计算后的调整值低于设定的“minMemory”值的情形，虽然此时配置不会生效但后台仍然会打印出这个调整值，用于为用户提供“minMemory”参数调整的依据，保证配置可以生效。

12.27.11 配置访问通道协议

配置场景

服务端配置了web访问为https通道，如果客户端没有配置，默认使用http访问，客户端和服务端的配置不同，就会导致访问结果显示乱码。在客户端和服务端配置相同的“yarn.http.policy”参数，可以防止客户端访问结果显示乱码。

操作步骤

步骤1 在Manager系统中，选择“集群 > 服务 > Yarn > 配置”，选择“全部配置”，在搜索框中输入参数名称“yarn.http.policy”。

- 安全模式下配置为“HTTPS_ONLY”。
- 普通模式下配置为“HTTP_ONLY”。

步骤2 以客户端安装用户，登录安装客户端的节点。

步骤3 执行以下命令，进入客户端安装路径。

cd /opt/client

步骤4 执行以下命令编辑“yarn-site.xml”文件。

vi Yarn/config/yarn-site.xml

修改“yarn.http.policy”的参数值。

安全模式下，“yarn.http.policy”配置成“HTTPS_ONLY”。

普通模式下，“yarn.http.policy”配置成“HTTP_ONLY”。

步骤5 执行**:wq**命令保存。

步骤6 重启客户端使配置生效。

----结束

12.27.12 检测内存使用情况

配置场景

针对所提交应用的内存使用无法预估的情况，可以通过修改服务端的配置项控制是否对内存使用进行检测。

若不检测内存使用，Container会占用内存直到内存溢出；若检测内存使用，当内存使用超过配置的内存大小时，相应的Container会被kill掉。

配置描述

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。

表 12-460 参数说明

参数	描述	默认值
yarn.nodemanager.vmem-check-enabled	是否进行虚拟内存检测的开关。如果任务使用的内存量超出分配值，则直接将任务强制终止。 ● 设置为true时，进行虚拟内存检测； ● 设置为false时，不进行虚拟内存检测。	MRS 3.x之前的版本集群:false MRS 3.x及后续版本集群:true
yarn.nodemanager.pmem-check-enabled	是否进行物理内存检测的开关。如果任务使用的内存量超出分配值，则直接将任务强制终止。 ● 设置为true时，进行物理内存检测； ● 设置为false时，不进行物理内存检测。	true

12.27.13 配置自定义调度器的 WebUI

配置场景

如果用户在ResourceManager中配置了自定义的调度器，可以通过以下配置项为其配置相应的Web展示页面及其他Web应用。

配置描述

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。

表 12-461 配置自定义调度器的 WebUI

参数	描述	默认值
hadoop.http.rmwebapp.scheduler.page.classes	在RM WebUI中为自定义调度器加载相应的web页面。仅当“yarn.resourcemanager.scheduler.class”配置为自定义调度器时此配置项生效。	-
yarn.http.rmwebapp.external.classes	在RM的Web服务中加载用户自定义的web应用。	-

12.27.14 配置 YARN Restart 特性

配置场景

YARN Restart特性包含两部分内容：ResourceManager Restart和NodeManager Restart。

- 当启用ResourceManager Restart时，升主后的ResourceManager就可以通过加载之前的主ResourceManager的状态信息，并通过接收所有NodeManager上container的状态信息，重构运行状态继续执行。这样应用程序通过定期执行检查点操作保存当前状态信息，就可以避免工作内容的丢失。
- 当启用NodeManager Restart时，NodeManager在本地保存当前节点上运行的container信息，重启NodeManager服务后通过恢复此前保存的状态信息，就不会丢失在此节点上运行的container进度。

配置描述

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。

ResourceManager Restart特性配置如下。

表 12-462 ResourceManager Restart 参数配置

参数	描述	默认值
yarn.resourcemanager.recovery.enabled	设置是否让ResourceManager在启动后恢复状态。如果设置为true，那yarn.resourcemanager.store.class也必须设置。	true
yarn.resourcemanager.store.class	指定用于保存应用程序和任务状态以及证书内容的state-store类。	MRS 3.x之前的版本集群：org.apache.hadoop.yarn.server.resourcemanager.recovery.ZKRMStateStore MRS 3.x及后续版本集群： org.apache.hadoop.yarn.server.resourcemanager.recovery.AsyncZKRMStateStore
yarn.resourcemanager.zk-state-store.parent-path	ZKRMStateStore在ZooKeeper上的保存目录。	/rmstore
yarn.resourcemanager.work-preserving-recovery.enabled	启用ResourceManager Work preserving功能。该配置仅用于YARN特性验证。	true
yarn.resourcemanager.state-store.async.load	对已完成的application采用ResourceManager异步恢复方式。	MRS 3.x之前的版本集群：false MRS 3.x及后续版本集群：true
yarn.resourcemanager.zk-state-store.num-fetch-threads	启用异步恢复功能，增加工作线程的数量可以加快恢复ZK中保存的任务信息的速度，取值范围大于0。	MRS 3.x之前的版本集群：1 MRS 3.x及后续版本集群：20

NodeManager Restart特性配置如下。

表 12-463 NodeManager Restart 参数配置

参数	描述	默认值
yarn.nodemanager.recovery.enabled	当Nodemanager重启时是否启用日志失败收集功能，是否恢复未完成的Application。	true

参数	描述	默认值
yarn.nodemanager.recovery.dir	NodeManager用于保存container状态的本地目录。适用于MRS 3.x及后续版本集群。	${SRV_HOME}/tmp/yarn-nm-recovery
yarn.nodemanager.recovery.supervised	NodeManager是否在监控下运行。开启此特性后NodeManager在退出后不会清理containers，NodeManager会假设自己会立即重启和恢复containers。	true

12.27.15 配置 AM 作业保留

本章节适用于MRS 3.x及后续版本集群。

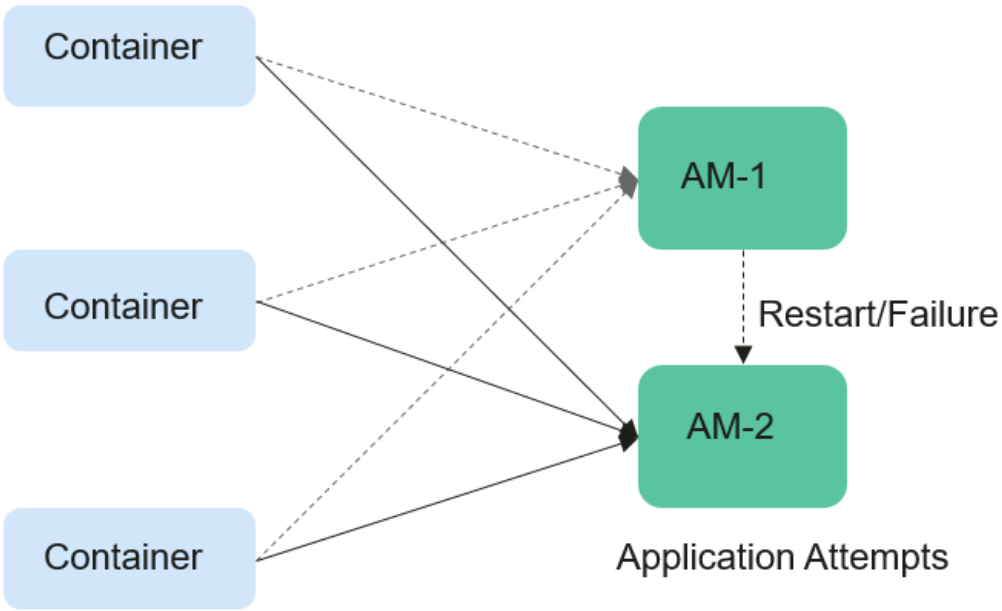
配置场景

在YARN中，ApplicationMaster(AM)与Container类似，都运行在NodeManager(NM)上（本文中忽略未管理的AM）。AM可能由于多种原因崩溃、退出或关闭。如果AM停止运行，ResourceManager(RM)会关闭ApplicationAttempt中管理的所有Container，其中包括当前在NM上运行的所有Container。RM会在另一计算节点上启动新的ApplicationAttempt。

对于不同类型的应用，希望以不同方式处理AM重启的事件。MapReduce类应用的目标是不丢失任务，但允许丢失当前运行的Container。但是对于长周期的YARN服务而言，用户可能并不希望由于AM的故障而导致整个服务停止运行。

YARN支持在新的ApplicationAttempt启动时，保留之前Container的状态，因此运行中的作业可以继续无故障的运行。

图 12-73 AM 作业保留



配置描述

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。

根据**表12-464**，对如下参数进行设置。

表 12-464 AM 作业保留相关参数

参数	说明	默认值
yarn.app.mapreduce.am.work-preserve	是否开启AM作业保留特性。	false
yarn.app.mapreduce.am.umbilical.max.retries	AM作业保留特性中，运行的容器尝试恢复的最大次数。	5
yarn.app.mapreduce.am.umbilical.retry.interval	AM作业保留特性中，运行的容器尝试恢复的时间间隔。单位：毫秒。	10000
yarn.resourcemanager.am.max-attempts	ApplicationMaster的重试次数。增加重试次数可以避免当资源不足时造成AM启动失败。 适用于所有ApplicationMaster的全局设置。每个ApplicationMaster都可以使用API设置一个单独的最大尝试次数，但这个次数不能大于全局的最大次数。如果大于了，那ResourceManager将会覆写这个单独的最大尝试次数。取值范围大于等于1。	2

12.27.16 配置本地化日志级别

本章节适用于MRS 3.x及后续版本集群。

配置场景

container本地化默认的日志级别是INFO。用户可以通过配置“yarn.nodemanager.container-localizer.java.opts”来改变日志级别。

配置描述

在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，选择“全部配置”，在NodeManager的配置文件“yarn-site.xml”中配置下面的参数来更改日志级别。

表 12-465 参数描述

参数	描述	默认值
yarn.nodemanager.container-localizer.java.opts	附加的jvm参数是提供给本地化container进程使用的。	-Xmx256m -Djava.security.krb5.conf=${KRB5_CONFIG}

默认值-Xmx256m -Djava.security.krb5.conf=${KRB5_CONFIG}和默认日志级别是INFO。为了更改container本地化的日志级别，添加下面的内容。

```
-Dhadoop.root.logger=<LOG_LEVEL>,localizationCLA
```

示例：

为了更改本地化日志级别为DEBUG，参数值应该为

```
-Xmx256m -Dhadoop.root.logger=DEBUG,localizationCLA
```

说明

允许的日志级别是：FATAL，ERROR，WARN，INFO，DEBUG，TRACE和ALL。

12.27.17 配置运行任务的用户

本章节适用于MRS 3.x及后续版本集群。

配置场景

目前YARN支持启动NodeManager的用户运行所有用户提交的任务，也支持以提交任务的用户运行任务。

配置描述

在Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，选择“全部配置”。在搜索框中输入参数名称。

表 12-466 参数描述

参数	描述	默认值
yarn.nodemanager.linux-container-executor.user	运行任务的用户。	默认为空。 **说明** 默认为空，实际以提交任务的用户来运行任务。
yarn.nodemanager.container-executor.class	启动任务的executor。	org.apache.hadoop.yarn.server.nodemanager.EnhancedLinuxContainerExecutor

说明

- “yarn.nodemanager.linux-container-executor.user”配置运行container的用户。默认空表示运行container的用户就是提交任务的用户。该参数仅在“yarn.nodemanager.container-executor.class”配置为“org.apache.hadoop.yarn.server.nodemanager.EnhancedLinuxContainerExecutor”时有效。
- 非安全模式下，当“yarn.nodemanager.linux-container-executor.user”设置为omm时，也需设置“yarn.nodemanager.linux-container-executor.nonsecure-mode.local-user”为omm。
- 建议“yarn.nodemanager.linux-container-executor.user”和“yarn.nodemanager.container-executor.class”这两个参数都采用默认值，这样安全性更高。

12.27.18 Yarn 日志介绍

日志描述

Yarn相关日志的默认存储路径如下：

- ResouceManager：“/var/log/Bigdata/yarn/rm”（运行日志），“/var/log/Bigdata/audit/yarn/rm”（审计日志）
- NodeManager：“/var/log/Bigdata/yarn/nm”（运行日志），“/var/log/Bigdata/audit/yarn/nm”（审计日志）

日志归档规则：Yarn的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过50MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的100个压缩文件，压缩文件保留个数可以在Manager界面中配置。

日志归档规则：

表 12-467 Yarn 日志列表

日志类型	日志文件名	描述
运行日志	hadoop-<SSH_USER>-<process_name>-<hostname>.log	Yarn组件日志，记录Yarn组件运行时候所产生的大部分日志。
	hadoop-<SSH_USER>-<process_name>-<hostname>.out	Yarn运行环境信息日志。
	<process_name>-<SSH_USER>-<DATE>-<PID>-gc.log	垃圾回收日志。
	yarn-haCheck.log	ResourceManager主备状态检测日志。
	yarn-service-check.log	Yarn服务健康状态检查日志。
	yarn-start-stop.log	Yarn服务启停操作日志。
	yarn-prestart.log	Yarn服务启动前集群操作的记录日志。

日志类型	日志文件名	描述
	yarn-postinstall.log	Yarn服务安装后启动前的工作日志。
	hadoop-commission.log	Yarn入服日志。
	yarn-cleanup.log	Yarn服务卸载时候的清理日志。
	yarn-refreshqueue.log	Yarn刷新队列日志。
	upgradeDetail.log	升级日志记录。
	stderr/stdin/syslog	Yarn服务上运行的应用所对应的container日志。
	yarn-application-check.log	Yarn服务上运行的应用检查日志。
	yarn-appsummary.log	Yarn服务上运行的应用的运行结果日志。
	yarn-switch-resourcemanager.log	Yarn主备倒换运行日志。
	ranger-yarn-plugin-enable.log	Yarn启用Ranger鉴权的日志
	yarn-nodemanager-period-check.log	Yarn nodemanager的周期检查日志
	yarn-resourcemanager-period-check.log	Yarn resourcemanager的周期检查日志
	hadoop.log	Hadoop的客户端日志
	env.log	实例启停前的环境信息日志。
审计日志	yarn-audit-<process_name>.log ranger-plugin-audit.log	Yarn操作审计日志。
	SecurityAuth.audit	Yarn安全审计日志。

日志级别

Yarn中提供了如表12-468所示的日志级别。其中日志级别优先级从高到低分别是OFF、FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-468 日志级别

级别	描述
FATAL	FATAL表示当前事件处理存在严重错误信息。

级别	描述
ERROR	ERROR表示当前事件处理存在错误信息。
WARN	WARN表示当前事件处理存在异常告警信息。
INFO	INFO表示记录系统及各事件正常运行状态信息
DEBUG	DEBUG表示记录系统及系统的调试信息

如果您需要修改日志级别，请执行如下操作：

步骤1 参考**修改集群服务配置参数**，进入Yarn服务“全部配置”页面。

步骤2 在左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 单击“保存配置”，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后立即生效，不需要重启服务。

----结束

日志格式

Yarn的日志格式如下所示：

表 12-469 日志格式

日志类型	格式	示例
运行日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2014-09-26 14:18:59,109 | INFO | main | Client environment:java.compiler=<NA> | org.apache.zookeeper.Environment.logEnv(Environment.java:100)
审计日志	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2014-09-26 14:24:43,605 | INFO | main-EventThread | USER=omm OPERATION=refreshAdminAcls TARGET=AdminService RESULT=SUCCESS | org.apache.hadoop.yarn.server.resourcemanager.RMAuditLogger$LogLevel$6.printLog(RMAuditLogger.java:91)

12.27.19 Yarn 性能调优

12.27.19.1 抢占任务

操作场景

抢占任务可精简队列中的job运行并提高资源利用率，由ResourceManager的capacity scheduler实现，其简易流程如下：

1. 假设存在两个队列A和B。其中队列A的capacity为25%，队列B的capacity为75%。
2. 初始状态下，任务1发送给队列A，此任务需要75%的集群资源。之后任务2发送到了队列B，此任务需要50%的集群资源。
3. 任务1将会使用队列A提供的25%的集群资源，并从队列B获取的50%的集群资源。队列B保留25%的集群资源。
4. 启用抢占任务特性，则任务1使用的资源将会被抢占。队列B会从队列A中获取25%的集群资源以满足任务2的执行。
5. 当任务2完成后，集群中存在足够的资源时，任务1将重新开始执行。

操作步骤

参数入口：

参考**修改集群服务配置参数**进入Yarn服务参数“全部配置”界面，在搜索框中输入参数名称。

表 12-470 Preemption 配置

参数	描述	默认值
yarn.resourcemanager.scheduler.monitor.enable	根据“yarn.resourcemanager.scheduler.monitor.policies”中的策略，启用新的scheduler监控。设置为“true”表示启用监控，并根据scheduler的信息，启动抢占的功能。设置为“false”表示不启用。	false
yarn.resourcemanager.scheduler.monitor.policies	设置与scheduler配合的“SchedulingEditPolicy”的类的清单。	org.apache.hadoop.yarn.server.resourcemanager.monitor.capacity.ProportionalCapacityPreemptionPolicy

参数	描述	默认值
yarn.resourcemanager.monitor.capacity.preemption.observe_only	● 设置为“true”，则执行策略，但是不对集群资源进程抢占操作。 ● 设置为“false”，则执行策略，且根据策略启用集群资源抢占的功能。	false
yarn.resourcemanager.monitor.capacity.preemption.monitoring_interval	根据策略监控的时间间隔，单位为毫秒。如果将该参数设置为更大的值，容量检测将不那么频繁地运行。	3000
yarn.resourcemanager.monitor.capacity.preemption.max_wait_before_kill	应用发送抢占需求到停止container（释放资源）的时间间隔，单位为毫秒。取值范围大于等于0。 默认情况下，若ApplicationMaster15秒内没有终止container，ResourceManager等待15秒后会强制终止。	15000
yarn.resourcemanager.monitor.capacity.preemption.total_preemption_per_round	在一个周期内能够抢占资源的最大的比例。可使用这个值来限制从集群回收容器的速度。计算出了期望的总抢占值之后，策略会伸缩回这个限制。	0.1
yarn.resourcemanager.monitor.capacity.preemption.max_ignored_over_capacity	集群中资源总量乘以此配置项的值加上某个队列（例如队列A）原有的资源量为资源抢占盲区。当队列A中的任务实际使用的资源超过该抢占盲区时，超过部分的资源将会被抢占。取值范围：0~1。 **说明** 设置的值越小越有利于资源抢占。	0
yarn.resourcemanager.monitor.capacity.preemption.natural_termination_factor	设置抢占目标，Container只会抢占所配置比例的资源。 示例，如果设置为0.5，则在5*“yarn.resourcemanager.monitor.capacity.preemption.max_wait_before_kill”的时间内，任务会回收所抢占资源的近95%。即接连抢占5次，每次抢占待抢占资源的0.5，呈几何收敛，每次的时间间隔为“yarn.resourcemanager.monitor.capacity.preemption.max_wait_before_kill”。取值范围：0~1。	1

12.27.19.2 任务优先级

操作场景

集群的资源竞争场景如下：

1. 提交两个低优先级的应用Job 1和Job 2。
2. 正在运行中的Job 1和Job 2有部分task处于running状态，但由于集群或队列资源容量有限，仍有部分task未得到资源而处于pending状态。
3. 提交一个较高优先级的应用Job 3，此时会出现如下资源分配情况：当Job 1和Job 2中running状态的task运行结束并释放资源后，Job 3中处于pending状态的task将优先得到这部分新释放的资源。
4. Job 3完成后，资源释放给Job 1、Job 2继续执行。

用户可以在YARN中配置任务的优先级。任务优先级是通过ResourceManager的调度器实现的。

操作步骤

设置参数“mapreduce.job.priority”，使用命令行接口或API接口设置任务优先级。

- 命令行接口。

  提交任务时，添加“**-Dmapreduce.job.priority=***<priority>*”参数。

  <priority>可以设置为：
  - VERY_HIGH
  - HIGH
  - NORMAL
  - LOW
  - VERY_LOW
- API接口。

  用户也可以使用API配置对象的优先级。

  设置优先级，可通过**Configuration.set("mapreduce.job.priority", <priority>)**或**Job.setPriority(JobPriority priority)**设置。

12.27.19.3 节点配置调优

操作场景

合理配置大数据集群的调度器后，还可通过调节每个节点的可用内存、CPU资源及本地磁盘的配置进行性能调优。

具体包括以下配置项：

- 可用内存
- CPU虚拟核数
- 物理CPU使用百分比
- 内存和CPU资源的协调
- 本地磁盘

操作步骤

若您需要对参数配置进行调整，具体操作请参考**修改集群服务配置参数**。

- **可用内存**

除了分配给操作系统、其他服务的内存外，剩余的资源应尽量分配给YARN。通过如下配置参数进行调整。

例如，如果一个container默认使用512M，则内存使用的计算公式为：512M*container数。

默认情况下，Map或Reduce container会使用1个虚拟CPU内核和1024MB内存，ApplicationMaster使用1536MB内存。

参数	描述	默认值
yarn.nodemanager.resource.memory-mb	设置可分配给容器的物理内存数量。单位：MB，取值范围大于0。 建议配置成节点物理内存总量的75%~90%。若该节点有其他业务的常驻进程，请降低此参数值给该进程预留足够运行资源。	MRS 3.x及之后：16384 MRS 3.x之前：8192

- **CPU虚拟核数**

  建议将此配置设定在逻辑核数的1.5～2倍之间。如果上层计算应用对CPU的计算能力要求不高，可以配置为2倍的逻辑CPU。

参数	描述	默认值
yarn.nodemanager.resource.cpu-vcores	表示该节点上YARN可使用的虚拟CPU个数，默认是8。 目前推荐将该值设值为逻辑CPU核数的1.5～2倍之间。	8

- **物理CPU使用百分比**

  建议预留适量的CPU给操作系统和其他进程（数据库、HBase等）外，剩余的CPU核都分配给YARN。可以通过如下配置参数进行调整。

参数	描述	默认值
yarn.nodemanager.resource.percentage-physical-cpu-limit	表示该节点上YARN可使用的物理CPU百分比。默认是90，即不进行CPU控制，YARN可以使用节点全部CPU。该参数只支持查看，可通过调整YARN的RES_CPUSET_PERCENTAGE参数来修改本参数值。注意，目前推荐将该值设为可供YARN集群使用的CPU百分数。 例如：当前节点除了YARN服务外的其他服务（如HBase、HDFS、Hive等）及系统进程使用CPU为20%左右，则可以供YARN调度的CPU为1-20%=80%，即配置此参数为80。	90

- **本地磁盘**

由于本地磁盘会提供给MapReduce写job执行的中间结果，数据量大。因此配置的原则是磁盘尽量多，且磁盘空间尽量大，单个达到百GB以上规模更好。简单的做法是配置和data node相同的磁盘，只在最下一级目录上不同即可。

说明

多个磁盘之间使用逗号隔开。

参数	描述	默认值
yarn.nodemanager.log-dirs	日志存放地址（可配置多个目录）。 容器日志的存储位置。默认值为%{@auto.detect.datapart.nm.logs}。如果有数据分区，基于该数据分区生成一个类似/srv/BigData/hadoop/data1/nm/containerlogs,/srv/BigData/hadoop/data2/nm/containerlogs的路径清单。如果没有数据分区，生成默认路径/srv/BigData/yarn/data1/nm/containerlogs。除了使用表达式以外，还可以输入完整的路径清单，比如/srv/BigData/yarn/data1/nm/containerlogs或/srv/BigData/yarn/data1/nm/containerlogs,/srv/BigData/yarn/data2/nm/containerlogs。这样数据就会存储在所有设置的目录中，一般会是在不同的设备中。为保证磁盘IO负载均衡，需要提供几个路径且每个路径都对应一个单独的磁盘。应用程序的本地化后的日志目录存在于相对路径/application_%{appid}中。单独容器的日志目录，即container_{$contid}，是该路径下的子目录。每个容器目录都含容器生成的stderr、stdin及syslog文件。要新增目录，比如新增/srv/BigData/yarn/data2/nm/containerlogs目录，应首先删除/srv/BigData/yarn/data2/nm/containerlogs下的文件。之后，为/srv/BigData/yarn/data2/nm/containerlogs赋予跟/srv/BigData/yarn/data1/nm/containerlogs一样的读写权限，再将/srv/BigData/yarn/data1/nm/containerlogs修改为/srv/BigData/yarn/data1/nm/containerlogs,/srv/BigData/yarn/data2/nm/containerlogs。可以新增目录，但不要修改或删除现有目录。否则，NodeManager的数据将丢失，且服务将不可用。 【默认值】%{@auto.detect.datapart.nm.logs} 【注意】请谨慎修改该项。如果配置不当，将造成服务不可用。当角色级别的该配置项修改后，所有实例级别的该配置项都将被修改。如	%{@auto.detect.datapart.nm.logs}

参数	描述	默认值
	果实例级别的配置项修改后，其他实例的该配置项的值保持不变。	

参数	描述	默认值
yarn.nodemanager.local-dirs	本地化后的文件的存储位置。默认值为%{@auto.detect.datapart.nm.localdir}。如果有数据分区，基于该数据分区生成一个类似/srv/BigData/hadoop/data1/nm/localdir,/srv/BigData/hadoop/data2/nm/localdir的路径清单。如果没有数据分区，生成默认路径/srv/BigData/yarn/data1/nm/localdir。除了使用表达式以外，还可以输入完整的路径清单，比如/srv/BigData/yarn/data1/nm/localdir或/srv/BigData/yarn/data1/nm/localdir,/srv/BigData/yarn/data2/nm/localdir。这样数据就会存储在所有设置的目录中，一般会是在不同的设备中。为保证磁盘IO负载均衡，需要提供几个路径且每个路径都对应一个单独的磁盘。应用程序的本地化后的文件目录存在于相对路径/usercache/%{user}/appcache/application_%{appid}中。单独容器的工作目录，即container_%{contid}，是该路径下的子目录。要新增目录，比如新增/srv/BigData/yarn/data2/nm/localdir目录，应首先删除/srv/BigData/yarn/data2/nm/localdir下的文件。之后，为/srv/BigData/hadoop/data2/nm/localdir赋予跟/srv/BigData/hadoop/data1/nm/localdir一样的读写权限，再将/srv/BigData/yarn/data1/nm/localdir修改为/srv/BigData/yarn/data1/nm/localdir,/srv/BigData/yarn/data2/nm/localdir。可以新增目录，但不要修改或删除现有目录。否则，NodeManager的数据将丢失，且服务将不可用。 【默认值】%{@auto.detect.datapart.nm.localdir} 【注意】请谨慎修改该项。如果配置不当，将造成服务不可用。当角色级别的该配置项修改后，所有实例级别的该配置项都将被修改。如果实例级别的配置项修改后，其他实例的该配置项的值保持不变。	%{@auto.detect.datapart.nm.localdir}

12.27.20 Yarn 常见问题

12.27.20.1 任务完成后 Container 挂载的文件目录未清除

问题

使用了CGroups功能的场景下，任务完成后Container挂载的文件目录未清除。

回答

即使任务失败，Container挂载的目录也应该被清除。

上述问题是由于删除动作超时导致的。完成某些任务所使用的时间已远超过删除时间。

为避免出现这种场景，您可以参考**修改集群服务配置参数**，进入Yarn“全部配置”页面。在搜索框搜索“yarn.nodemanager.linux-container-executor.cgroups.delete-timeout-ms”配置项来修改删除时间的时长。参数值的单位为毫秒。

12.27.20.2 作业执行失败时会抛出 HDFS_DELEGATION_TOKEN 到期的异常

问题

安全模式下，为什么作业执行失败时会抛出HDFS_DELEGATION_TOKEN到期的异常?

回答

HDFS_DELEGATION_TOKEN到期的异常是由于token没有更新或者超出了最大生命周期。

在token的最大生命周期内确保下面的参数值大于作业的运行时间。

“dfs.namenode.delegation.token.max-lifetime”=“604800000”（默认是一星期）

参考**修改集群服务配置参数**，进入HDFS“全部配置”页面，在搜索框搜索该参数。

说明

建议在token的最大生命周期内参数值为多倍小时数。

12.27.20.3 重启 YARN，本地日志不被删除

问题

在以下两种情况下重启YARN，本地日志不会被定时删除，将被永久保留。

- 在任务运行过程中，重启YARN，本地日志不被删除。
- 在任务完成，日志归集失败后定时清除日志前，重启YARN，本地日志不被删除。

回答

NodeManager有个重启恢复机制（详情请参见**https://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/**

NodeManager.html#NodeManager_Restart），参考**修改集群服务配置参数**，进入Yarn“全部配置”页面。需将NodeManager的“yarn.nodemanager.recovery.enabled”配置项为“true”后才生效，默认为“true”，这样在YARN重启的异常场景时会定时删除多余的本地日志，避免问题的出现。

12.27.20.4 为什么执行任务时 AppAttempts 重试次数超过 2 次还没有运行失败

问题

系统默认的AppAttempts运行失败的次数为2，为什么在执行任务时，AppAttempts重试次数超过2次还没有运行失败?

回答

在执行任务过程中，若ContainerExitStatus的返回值为ABORTED、PREEMPTED、DISKS_FAILED、KILLED_BY_RESOURCEMANAGER这四种状态之一时，系统不会将其计入failed attempts中，因此出现上面的问题，只有当真正失败尝试2次之后才会运行失败。

12.27.20.5 为什么在 ResourceManager 重启后，应用程序会移回原来的队列

问题

将应用程序从一个队列移到另一个队列时，为什么在RM（ResourceManager）重启后，应用程序会被移回原来的队列?

回答

这是RM的使用限制，应用程序运行过程中移动到别的队列，此时RM重启，RM并不会在状态存储中存储新队列的信息。

假设用户提交一个MR任务到叶子队列test11上。当任务运行时，删除叶子队列test11，这时提交队列自动变为lost_and_found队列（找不到队列的任务会被放入lost_and_found队列中），任务暂停运行。要启动该任务，用户将任务移动到叶子队列test21上。在将任务移动到叶子队列test21后，任务继续运行，此时RM重启，重启后显示提交队列为lost_and_found队列，而不是test21队列。

发生上述情况的原因是，任务未完成时，RM状态存储中存储的还是应用程序移动前的队列状态。唯一的解决办法就是等RM重启后，再次移动应用程序，将新的队列状态信息写入状态存储中。

12.27.20.6 为什么 YARN 资源池的所有节点都被加入黑名单，而 YARN 却没有释放黑名单，导致任务一直处于运行状态

问题

为什么YARN资源池的所有节点都被加入黑名单，而YARN却没有释放黑名单，导致任务一直处于运行状态?

回答

在YARN中，当一个APP的节点被AM（ApplicationMaster）加入黑名单的数量达到一定比例（默认值为节点总数的33%）时，该AM会自动释放黑名单，从而不会出现由于所有可用节点都被加入黑名单而任务无法获取节点资源的现象。

在资源池场景下，假设该集群上有8个节点，通过NodeLabel特性将集群划分为两个资源池，pool A和pool B，其中pool B包含两个节点。用户提交了一个任务App1到pool B，由于HDFS空间不足，App1运行失败，导致pool B的两个节点都被App1的AM加入了黑名单，根据上述原则，2个节点小于8个节点的33%，所以YARN不会释放黑名单，使得App1一直无法得到资源而保持运行状态，后续即使被加入黑名单的节点恢复，App1也无法得到资源。

由于上述原则不适用于资源池场景，所以目前可通过调整客户端参数“yarn.resourcemanager.am-scheduling.node-blacklisting-disable-threshold”为：(nodes number of pool / total nodes)* 33%解决该问题。

12.27.20.7 ResourceManager 持续主备倒换

问题

RM（ResourceManager）在多个任务（比如2000个任务）正常并发运行时出现持续的主备倒换，导致YARN服务不可用。

回答

产生上述问题的原因是，full GC（GabageCollection）时间过长，超出了RM与ZK（ZooKeeper）之间定期交互时长的阈值，导致RM与ZK失联，从而造成RM主备倒换。

在多任务情况下，RM需要保存多个任务的鉴权信息，并通过心跳传递给各个NM（NodeManager），即心跳Response。心跳Response的生命周期短，默认值为1s，一般可以在JVM minor GC时被回收，但在多任务的情况下，集群规模较大，比如5000节点，多个节点的心跳Response会占用大量内存，导致JVM在minor GC时无法完全回收，无法回收的内存持续累积，最终触发JVM的full GC。JVM的GC都是阻塞式的，即在GC过程中不执行任何作业，所以若full GC的时间过长，超出了RM与ZK之间定期交互时长的阈值，就会出现主备倒换。

登录FusionInsight Manager，选择“集群 > 服务 > Yarn > 配置 > 全部配置”，在左侧选择“Yarn > 自定义”，在“yarn.yarn-site.customized.configs”中添加“yarn.resourcemanager.zk-timeout-ms”参数来增大RM与ZK之间定期交互时长的阈值（参数值的范围为小于等于90000毫秒），可以解决RM持续主备倒换的问题。

12.27.20.8 当一个 NodeManager 处于 unhealthy 的状态 10 分钟时，新应用程序失败

问题

当一个NM（NodeManager）处于unhealthy的状态10分钟时，新应用程序失败。

回答

当nodeSelectPolicy为SEQUENCE，且第一个连接到RM的NM不可用时，RM会在“yarn.nm.liveness-monitor.expiry-interval-ms”属性中指定的周期内，一直尝试为同一个NM分配任务。

可以通过两种方式来避免上述问题：

- 使用其他的nodeSelectPolicy，如RANDOM。
- 参考**修改集群服务配置参数**，进入Yarn“全部配置”页面。在搜索框搜索以下参数，通过“yarn-site.xml”文件更改以下属性：

  “yarn.resourcemanager.am-scheduling.node-blacklisting-enabled” = “true”；

  “yarn.resourcemanager.am-scheduling.node-blacklisting-disable-threshold” = “0.5”。

12.27.20.9 Superior 通过 REST 接口查看已结束或不存在的 applicationID，返回的页面提示 Error Occurred

问题

Superior通过REST接口查看已结束或不存在的applicationID，返回的页面提示Error Occurred。

回答

用户提交查看applicationID的请求，访问REST接口“https://<SS_REST_SERVER>/ws/v1/sscheduler/applications/{application_id}”。

由于Superior Scheduler只存储正在运行的applicationID，所以当查看的是已结束或不存在的applicationID，服务器会响应给浏览器“404”的状态码。但是由于chrome浏览器访问该REST接口时，优先以“application/xml”的格式响应，该行为会导致服务器端处理出现异常，所以返回的页面会提示“Error Occurred”。而IE浏览器访问该REST接口时，优先以“application/json”的格式响应，服务器会正确响应给浏览器“404”的状态码。

12.27.20.10 Superior 调度模式下，单个 NodeManager 故障可能导致 MapReduce 任务失败

问题

在Superior调度模式下，如果出现单个NodeManager故障，可能会导致Mapreduce任务失败。

回答

正常情况下，当一个application的单个task的attempt连续在一个节点上失败3次，那么该application的AppMaster就会将该节点加入黑名单，之后AppMaster就会通知调度器不要继续调度task到该节点，从而避免任务失败。

但是默认情况下，当集群中有33%的节点都被加入黑名单时，调度器会忽略黑名单节点。因此，该黑名单特性在小集群场景下容易失效。比如，集群只有3个节点，当1个

节点出现故障，黑名单机制失效，不管task的attempt在同一个节点失败多少次，调度器仍然会将task继续调度到该节点，从而导致application因为task失败达到最大attempt次数（MapReduce默认4次）而失败。

规避手段：

“yarn.resourcemanager.am-scheduling.node-blacklisting-disable-threshold”参数以百分比的形式配置忽略黑名单节点的阈值。建议根据集群规模，适当增大该参数的值，如3个节点的集群， 建议增大到50%。

说明

Superior调度器的框架设计是基于时间的异步调度，当NodeManager故障后，ResourceManager无法快速的感知到NodeManager已经出了问题(默认10mins)，因此在此期间，Superior调度器仍然会向该节点调度task，从而导致任务失败。

12.27.20.11 当应用程序从 lost_and_found 队列移动到其他队列时，应用程序不能继续执行

问题

当删除一个有部分应用程序正在运行的队列，这些应用程序会被移动到“lost_and_found”队列上。当这些应用程序移回运行正常的队列时，某些任务会被挂起，不能正常运行。

回答

如果应用程序没有设置标签表达式，那么该应用程序上新增的container/resource将使用其所在队列默认的标签表达式。如果队列没有默认的标签表达式，则将其标签表达设置为“default label”。

当应用程序（app1）提交到队列（Q1）上时，应用程序上新增的container/resource使用队列默认的标签表达式（“label1”）。若app1正在运行时Q1被删除，则app1被移动到“lost_and_found”队列上。由于“lost_and_found”队列没有标签表达式，其标签表达式设置为“default label”，此时app1上新增的container/resource也将其标签表达式设置为“default label”。当app1被移回正常运行的队列（例如，Q2）时，如果Q2支持调用app1中的所有标签表达式（包含“label1”和“default label”），则app1能正常运行直到结束；如果Q2仅支持调用app1中的部分标签表达式（例如，仅支持调用“default label”），那么app1在运行时，拥有“label1”标签表达式的部分任务的资源请求将无法获得资源，从而被挂起，不能正常运行。

因此当把应用程序从“lost_and_found”队列移动到其他运行正常的队列上时，需要保证目标队列能够调用该应用程序的所有标签表达式。

建议不要删除正在运行应用程序的队列。

12.27.20.12 如何限制存储在 ZKstore 中的应用程序诊断消息的大小

问题

如何限制存储在ZKstore中的应用程序诊断消息的大小?

回答

在某些情况下，已经观察到诊断消息可能无限增长。由于诊断消息存储在状态存储中，不建议允许诊断消息无限增长。因此，需要有一个属性参数用于设置诊断消息的最大大小。

若您需要设置“yarn.app.attempt.diagnostics.limit.kc”参数值，具体操作参考**修改集群服务配置参数**，进入Yarn“全部配置”页面，在搜素框搜索以下参数。

表 12-471 参数描述

参数	描述	默认值
yarn.app.attempt.diagnostics.limit.kc	定义每次应用连接的诊断消息的数据大小，以千字节为单位（字符数*1024）。当使用ZooKeeper来存储应用程序的行为状态时，需要限制诊断消息的大小，以防止YARN拖垮ZooKeeper。如果将“yarn.resourcemanager.state-store.max-completed-applications”设置为一个较大的数值，则需要减小该属性参数的值以限制存储的总数据大小。	64

12.27.20.13 为什么将非 ViewFS 文件系统配置为 ViewFS 时 MapReduce 作业运行失败

问题

为什么将非ViewFS文件系统配置为ViewFS时MR作业运行失败?

回答

通过集群将非ViewFS文件系统配置为ViewFS时，ViewFS中的文件夹的用户权限与默认NameService中的非ViewFS不同。因为目录权限不匹配，所以已提交的MR作业运行失败。

在集群中配置ViewFS的用户，需要检查并校验目录权限。在提交作业之前，应按照默认的NameService文件夹权限更改ViewFS文件夹权限。

下表列出了ViewFS中配置的目录的默认权限结构。如果配置的目录权限与下表不匹配，则必须相应地更改目录权限。

表 12-472 ViewFS 中配置的目录的默认权限结构

参数	描述	默认值	默认值及其父目录的默认权限
yarn.nodemanager.remote-app-log-dir	在默认文件系统上（通常是HDFS），指定NM应将日志聚合到哪个目录。	logs	777

参数	描述	默认值	默认值及其父目录的默认权限
yarn.nodemanager.remote-app-log-archive-dir	将日志归档的目录。	-	777
yarn.app.mapreduce.am.staging-dir	提交作业时使用的staging目录。	/tmp/hadoop-yarn/staging	777
mapreduce.jobhistory.intermediate-done-dir	MapReduce作业记录历史文件的目录。	${yarn.app.mapreduce.am.staging-dir}/history/done_intermediate	777
mapreduce.jobhistory.done-dir	由MR JobHistory Server管理的历史文件的目录。	${yarn.app.mapreduce.am.staging-dir}/history/done	777

12.27.20.14 开启 Native Task 特性后，Reduce 任务在部分操作系统运行失败

问题

开启Native Task特性后，Reduce任务在部分操作系统运行失败。

回答

运行包含Reduce的Mapreduce任务时，通过-Dmapreduce.job.map.output.collector.class=org.apache.hadoop.mapred.nativetask.NativeMapOutputCollectorDelegator命令开启Native Task特性，任务在部分操作系统运行失败，日志中提示错误“version 'GLIBCXX_3.4.20' not found”。该问题原因是操作系统的GLIBCXX版本较低，导致该特性依赖的libnativetask.so.1.0.0库无法加载，进而导致任务失败。

规避手段：

设置配置项mapreduce.job.map.output.collector.class的值为org.apache.hadoop.mapred.MapTask$MapOutputBuffer。

12.28 使用 ZooKeeper

12.28.1 从零开始使用 Zookeeper

Zookeeper是一个开源的，高可靠的，分布式一致性协调服务。Zookeeper设计目标是用来解决那些复杂，易出错的分布式系统难以保证数据一致性的。不必开发专门的协同应用，十分适合高可用服务保持数据一致性。

背景信息

在使用客户端前，除主管理节点以外的客户端，需要下载并更新客户端配置文件。

操作步骤

MRS 2.x及以前版本集群执行以下操作：

步骤1 下载客户端配置文件。

1. 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击待操作集群的名称。该集群为“用户指南 > 配置集群 > 创建自定义集群”中创建的集群。
2. 选择“组件管理”。
3. 单击“服务管理”，然后单击“下载客户端”。

   在“客户端类型”选择“仅配置文件”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤2 登录MRS Manager的主管理节点。

1. 在MRS控制台，选择“集群列表 > 现有集群”，单击集群名称，在“节点管理”页签中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。

   MRS Manager的主备管理节点默认安装在集群Master节点上。在主备模式下，由于Master1和Master2之间会切换，Master1节点不一定是MRS Manager的主管理节点，需要在Master1节点中执行命令，确认MRS Manager的主管理节点。命令请参考**步骤2.4**。
2. 以**root**用户使用密码方式登录Master1节点。操作方法，请参见“用户指南 > 连接集群 > 登录集群 > 登录集群节点”章节。
3. 切换至**omm**用户。

   sudo su - root

   su - omm
4. 执行以下命令确认MRS Manager的主管理节点。

   sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh

   回显信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“mgtomsdat-sh-3-01-1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“mgtomsdat-sh-3-01-2”为备管理节点）。

   ```
   Ha mode
   double
   NodeName            HostName                HAVersion       StartTime           HAActive
   HAAllResOK          HARunPhase
   192-168-0-30        mgtomsdat-sh-3-01-1     V100R001C01     2014-11-18 23:43:02
   active              normal                  Actived
   192-168-0-24        mgtomsdat-sh-3-01-2     V100R001C01     2014-11-21 07:14:02
   standby             normal                  Deactived
   ```
5. 使用**root**用户登录MRS Manager的主管理节点，例如“192-168-0-30”节点，并执行以下命令切换到**omm**用户。

   sudo su - omm

步骤3 执行以下命令切换到客户端安装目录。例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，更新主管理节点的客户端配置。

sh refreshConfig.sh /opt/client *客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/client/tmp/MRS-client/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

说明

步骤**步骤1**~**步骤4**的操作也可以参考“用户指南 > 连接集群 > 使用MRS客户端 > 更新客户端”页面的方法二操作。

步骤5 在Master节点使用客户端。

1. 在已更新客户端的主管理节点，例如“192-168-0-30”节点，执行以下命令切换到客户端目录。

   cd /opt/client

2. 执行以下命令配置环境变量。

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit MRS集群用户

   例如，kinit zookeeperuser。

4. 直接执行Zookeeper组件的客户端命令。

   zkCli.sh -server <zookeeper安装节点ip>:<port>

   例如：***zkCli.sh -server node-master1DGhZ:2181***

步骤6 运行Zookeeper客户端命令。

1. 创建ZNode。
   ```
   create /test
   ```
2. 查看ZNode信息。
   ```
   ls /
   ```
3. 向ZNode中写入数据。
   ```
   set /test "zookeeper test"
   ```
4. 查看写入ZNode中的数据。
   ```
   get /test
   ```
5. 删除创建的ZNode。
   ```
   delete /test
   ```

----结束

MRS 3.x及以后版本集群执行以下操作：

步骤1 下载客户端配置文件。

1. 登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。
2. 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 下载客户端”。
3. 下载集群客户端。

“选择客户端类型”选择“仅配置文件”，选择平台类型，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client/”。

步骤2 登录Manager的主管理节点。

1. 以**root**用户登录任意部署Manager的节点。
2. 执行以下命令确认主备管理节点。

   sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh

   界面打印信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“node-master1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“node-master2”为备管理节点）。

   ```
   HAMode
   double
   NodeName            HostName        HAVersion       StartTime             HAActive
   HAAllResOK          HARunPhase
   192-168-0-30        node-master1    V100R001C01     2020-05-01 23:43:02   active
   normal              Actived
   192-168-0-24        node-master2    V100R001C01     2020-05-01 07:14:02   standby
   normal              Deactived
   ```

3. 以**root**用户登录主管理节点，并执行以下命令切换到**omm**用户。

   sudo su - omm

步骤3 执行以下命令切换到客户端安装目录。例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，更新主管理节点的客户端配置。

sh refreshConfig.sh /opt/client *客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/client /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

步骤5 在Master节点使用客户端。

1. 在已更新客户端的主管理节点，例如“192-168-0-30”节点，执行以下命令切换到客户端目录。

   cd /opt/client

2. 执行以下命令配置环境变量。

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，具体请参见配置拥有对应权限的角色，参考为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit MRS 集群用户

   例如，**kinit zookeeperuser**。

4. 直接执行Zookeeper组件的客户端命令。

   zkCli.sh -server <zookeeper安装节点ip>:<port>

   例如：**zkCli.sh -server node-master1DGhZ:2181**

步骤6 运行Zookeeper客户端命令。

1. 创建ZNode。
   ```
   create /test
   ```
2. 查看ZNode信息。
   ```
   ls /
   ```
3. 向ZNode中写入数据。
   ```
   set /test "zookeeper test"
   ```
4. 查看写入ZNode中的数据。
   ```
   get /test
   ```
5. 删除创建的ZNode。
   ```
   delete /test
   ```

----结束

12.28.2 ZooKeeper 常用参数

参数入口：

请参考**修改集群服务配置参数**，进入ZooKeeper“全部配置”页面。在搜索框中输入参数名称。

表 12-473 参数说明

配置参数	说明	默认值
skipACL	是否跳过ZooKeeper节点的权限检查。	no
maxClientCnxns	ZooKeeper的最大连接数，在连接数多的情况下，建议增加。	2000
LOG_LEVEL	日志级别，在调试的时候，可以改为DEBUG。	INFO
acl.compare.shortName	当Znode的ACL权限认证类型为SASL时，是否仅使用principal的用户名部分进行ACL权限认证。	true
synclimit	Follower与leader进行同步的时间间隔（单位为tick）。如果在指定的时间内leader没响应，连接将不能被建立。	15
tickTime	一次tick的时间（毫秒），它是ZooKeeper使用的基本时间单位，心跳、超时的时间都由它来规定。	4000

说明

ZooKeeper内部时间由参数ticktime和参数synclimit控制，如需调大ZooKeeper内部超时时间，需要调大客户端连接ZooKeeper的超时时间。

12.28.3 使用 ZooKeeper 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用ZooKeeper客户端。

前提条件

已安装客户端。例如安装目录为“/opt/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 执行以下命令进行用户认证。(普通模式跳过此步骤)

kinit *组件业务用户*

步骤5 执行以下命令登录客户端工具。

zkCli.sh -server *ZooKeeper角色实例所在节点业务IP: clientPort*

----结束

12.28.4 ZooKeeper 权限设置指南

操作场景

该操作指导用户对ZooKeeper的znode设置权限。

ZooKeeper通过访问控制列表（ACL）来对znode进行访问控制。ZooKeeper客户端为znode指定ACL，ZooKeeper服务器根据ACL列表判定某个请求znode的客户端是否有对应操作的权限。ACL设置涉及如下四个方面。

- 查看ZooKeeper中znode的ACL。
- 增加ZooKeeper中znode的ACL。
- 修改ZooKeeper中znode的ACL。
- 删除ZooKeeper中znode的ACL。

  ZooKeeper的ACL权限说明：

  ZooKeeper目前支持create，delete，read，write，admin五种权限，且ZooKeeper对权限的控制是znode级别的，而且不继承，即对父znode设置权限，

其子znode不继承父znode的权限。ZooKeeper中znode的默认权限为**world:anyone:cdrwa**，即任何用户都有所有权限。

说明

ACL有三部分：

第一部分是认证类型，如world指所有认证类型，sasl是kerberos认证类型；

第二部分是帐号，如anyone指的是任何人；

第三部分是权限，如cdrwa指的是拥有所有权限。

特别的，由于普通模式启动客户端不需要认证，sasl认证类型的ACL在普通模式下将不能使用。本文所有涉及sasl方式的鉴权操作均是在安全集群中进行。

表 12-474 Zookeeper 的五种 ACL

权限说明	权限简称	权限详情
创建权限	create(c)	可以在当前znode下创建子znode
删除权限	delete(d)	删除当前的znode
读权限	read(r)	获取当前znode的数据，可以列出当前znode所有的子znodes
写权限	write(w)	向当前znode写数据，写入子znode
管理权限	admin(a)	设置当前znode的权限

对系统的影响

须知

修改ZooKeeper的ACL是高危操作。修改ZooKeeper中znode的权限，可能会导致其他用户无权限访问该znode，导致系统功能异常。另外在3.5.6及以后版本，用户对于getAcl操作需要有读权限。

前提条件

- 已安装ZooKeeper客户端。例如安装目录为“/opt/client”。
- 已获取系统管理员用户和密码。

操作步骤

启动ZooKeeper客户端

步骤1 以**root**用户登录安装了ZooKeeper客户端的服务器。

步骤2 进入客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 执行以下命令认证用户身份，并输入用户密码（任意有权限的用户，这里以admin为例，普通模式不涉及）。

kinit admin

步骤5 在ZooKeeper客户端执行以下命令，进入ZooKeeper命令行。

sh zkCli.sh -server *ZooKeeper任意实例所在节点业务平面IP.clientPort*

默认的“clientPort”为“2181”

例如：**sh zkCli.sh -server 192.168.0.151:2181**

步骤6 登录ZooKeeper客户端后，使用**ls**命令，可以查看ZooKeeper中的znode列表。例如，可以查看根目录znode列表。

ls /

```
[zk: 192.168.0.151:2181(CONNECTED) 1] ls /
[hadoop-flag, hadoop-ha, test, test2, test3, test4, test5, test6, zookeeper]
```

查看ZooKeeper znode ACL信息

步骤7 启动ZooKeeper客户端。

步骤8 使用**getAcl**命令，可以查看znode。如下命令，可以查看到之前创建的名为test的znode的ACL权限。

getAcl */znode名称*

```
[zk: 192.168.0.151:2181(CONNECTED) 2] getAcl /test
'world,'anyone
: cdrwa
```

增加ZooKeeper znode ACL信息

步骤9 启动ZooKeeper客户端。

步骤10 查看旧的ACL信息，查看当前帐号是否有权限修改该znode的ACL信息的权限（a权限），如果没有权限，需要kinit登录有权限的用户，并重新启动ZooKeeper客户端。

getAcl */znode名称*

```
 [zk: 192.168.0.151:2181(CONNECTED) 3] getAcl /test
'world,'anyone
: cdrwa
```

步骤11 使用**setAcl**命令增加权限。设置新权限命令如下：

setAcl /test world:anyone:cdrwa,sasl:*用户名@<系统域名>:权限值*

例如对test的znode，需要增加admin用户的权限：

setAcl /test world:anyone:cdrwa,sasl:admin@HADOOP.COM:cdrwa

说明

增加权限时，需要保留已有权限。新增加权限和旧的权限用英文逗号隔开，新增加权限有三个部分：

第一部分是认证类型，如sasl指使用kerberos认证；

第二部分是帐号，如admin@HADOOP.COM指的是admin用户；

第三部分是权限，如cdrwa指的是拥有所有权限。

步骤12 setAcl后，可以使用**getAcl**命令查看增加权限是否成功：

getAcl */znode名称*

```
[zk: 192.168.0.151:2181(CONNECTED) 4] getAcl /test
'world,'anyone
: cdrwa
'sasl,'admin@<系统域名>
: cdrwa
```

修改ZooKeeper znode ACL信息

步骤13 启动ZooKeeper客户端。

步骤14 查看旧的ACL信息，查看当前帐号是否有权限修改该znode的ACL信息的权限（a权限），如果没有权限，需要kinit登录有权限的用户，并重新启动ZooKeeper客户端。

getAcl */znode名称*

```
[zk: 192.168.0.151:2181(CONNECTED) 5] getAcl /test
'world,'anyone
: cdrwa
'sasl,'admin@<系统域名>
: cdrwa
```

步骤15 使用**setAcl**命令修改权限。设置新权限命令如下：

s**etAcl /test sasl:***用户名@<系统域名>:权限值*

例如仅保留admin用户的所有权限，删除anyone用户的rw权限。

setAcl /test sasl:admin@HADOOP.COM:cdrwa

步骤16 setAcl后，可以使用**getAcl**命令查看修改权限是否成功：

getAcl */znode名称*

```
[zk: 192.168.0.151:2181(CONNECTED) 6] getAcl /test
'sasl,'admin@<系统域名>
: cdrwa
```

删除ZooKeeper znode ACL信息

步骤17 启动ZooKeeper客户端。

步骤18 查看旧的ACL信息，查看当前帐号是否有权限修改该znode的ACL信息的权限（a权限），如果没有权限，需要kinit登录有权限的用户，并重新启动ZooKeeper客户端。

getAcl */znode名称*

```
[zk: 192.168.0.151:2181(CONNECTED) 5] getAcl /test
'world,'anyone
: rw
'sasl,'admin@<系统域名>
: cdrwa
```

步骤19 使用**setAcl**命令增加权限。设置新权限命令如下：

s**etAcl /test sasl:***用户名@<系统域名>:权限值*

例如，仅保留admin用户是所有权限，取消anyone用户的rw权限。

setAcl /test sasl:admin@HADOOP.COM:cdrwa

步骤20 setAcl后，可以使用**getAcl**命令查看修改权限是否成功

getAcl */znode名称*

```
[zk: 192.168.0.151:2181(CONNECTED) 6] getAcl /test
'sasl,'admin@<系统域名>
: cdrwa
```

----结束

12.28.5 ZooKeeper 日志介绍

日志描述

日志存储路径：“/var/log/Bigdata/zookeeper/quorumpeer”（运行日志），“/var/log/Bigdata/audit/zookeeper/quorumpeer”（审计日志）

日志归档规则：ZooKeeper的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过30MB的时候，会自动压缩。最多保留20个压缩文件，压缩文件保留个数可以在Manager界面中配置。

表 12-475 ZooKeeper 日志列表

日志类型	日志文件名	描述
运行日志	zookeeper-<SSH_USER>-<process_name>-<hostname>.log	ZooKeeper系统日志，记录ZooKeeper系统运行时候所产生的大部分日志。
	check-serviceDetail.log	ZooKeeper服务启动是否成功的检查日志。
	zookeeper-<SSH_USER>-<DATA>-<PID>-gc.log	ZooKeeper垃圾回收日志。
	instanceHealthDetail.log	ZooKeeper实例健康状态检查日志
	zookeeper-omm-server-<hostname>.out	ZooKeeper运行异常退出日志。
	zk-err-<zkpid>.log	ZooKeeper致命错误日志。
	java_pid<zkpid>.hprof	ZooKeeper内存溢出日志。
	funcDetail.log	ZooKeeper实例启动日志。
	zookeeper-period-check.log	ZooKeeper实例健康检查日志。
	zookeeper-period-check-java.log	ZooKeeper配额监控周期检查日志。
审计日志	zk-audit-quorumpeer.log	ZooKeeper操作审计日志。

日志级别

ZooKeeper中提供了如表12-476所示的日志级别。日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-476 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO表示系统及各事件正常运行状态信息。
DEBUG	DEBUG表示系统及系统的调试信息。

如果您需要修改日志级别，请执行如下操作：

步骤1 参考**修改集群服务配置参数**章节，进入ZooKeeper服务“全部配置”页面。

步骤2 左边菜单栏中选择所需修改的角色所对应的日志菜单。

步骤3 选择所需修改的日志级别。

步骤4 单击“保存”，在弹出窗口中单击“确定”使配置生效。

说明

配置完成后立即生效，不需要重启服务。

----结束

日志格式

ZooKeeper的日志格式如下所示：

表 12-477 日志格式

日志类型	组件	格式	示例
运行日志	zookeeper quorumpeer	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-01-20 16:33:43,816	INFO	main	Defaulting to majority quorums	org.apache.zookeeper.server.quorum.QuorumPeerConfig.parseProperties(QuorumPeerConfig.java:335)

日志类型	组件	格式	示例
审计日志	zookeeper quorumpeer	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-01-20 16:33:54,313 | INFO | CommitProcessor:13 | session=0xd4b0679daea0000 ip=10.177.112.145 operation=create znode target=ZooKeeperServer znode=/zk-write-test-2 result=success | org.apache.zookeeper.ZKAuditLogger$LogLevel$5.printLog(ZKAuditLogger.java:70)

12.28.6 ZooKeeper 常见问题

12.28.6.1 创建大量 znode 后，ZooKeeper Sever 启动失败

问题

创建大量znode后，ZooKeeper集群处于故障状态不能自动恢复，尝试重启失败，ZooKeeper server日志显示如下内容：

follower：

```
2016-06-23 08:00:18,763 | WARN  | QuorumPeer[myid=26](plain=/10.16.9.138:2181)(secure=disabled) |
Exception when following the leader |
org.apache.zookeeper.server.quorum.Follower.followLeader(Follower.java:93)
java.net.SocketTimeoutException: Read timed out
    at java.net.SocketInputStream.socketRead0(Native Method)
    at java.net.SocketInputStream.socketRead(SocketInputStream.java:116)
    at java.net.SocketInputStream.read(SocketInputStream.java:170)
    at java.net.SocketInputStream.read(SocketInputStream.java:141)
    at java.io.BufferedInputStream.fill(BufferedInputStream.java:246)
    at java.io.BufferedInputStream.read(BufferedInputStream.java:265)
    at java.io.DataInputStream.readInt(DataInputStream.java:387)
    at org.apache.jute.BinaryInputArchive.readInt(BinaryInputArchive.java:63)
    at org.apache.zookeeper.server.quorum.QuorumPacket.deserialize(QuorumPacket.java:83)
    at org.apache.jute.BinaryInputArchive.readRecord(BinaryInputArchive.java:99)
    at org.apache.zookeeper.server.quorum.Learner.readPacket(Learner.java:156)
    at org.apache.zookeeper.server.quorum.Learner.registerWithLeader(Learner.java:276)
    at org.apache.zookeeper.server.quorum.Follower.followLeader(Follower.java:75)
    at org.apache.zookeeper.server.quorum.QuorumPeer.run(QuorumPeer.java:1094)
2016-06-23 08:00:18,764 | INFO  | QuorumPeer[myid=26](plain=/10.16.9.138:2181)(secure=disabled) |
shutdown called | org.apache.zookeeper.server.quorum.Follower.shutdown(Follower.java:198)
java.lang.Exception: shutdown Follower
    at org.apache.zookeeper.server.quorum.Follower.shutdown(Follower.java:198)
```

```
    at org.apache.zookeeper.server.quorum.QuorumPeer.stopFollower(QuorumPeer.java:1141)
    at org.apache.zookeeper.server.quorum.QuorumPeer.run(QuorumPeer.java:1098)
```

leader：

```
2016-06-23 07:30:57,481 | WARN  | QuorumPeer[myid=25](plain=/10.16.9.136:2181)(secure=disabled) |
Unexpected exception | org.apache.zookeeper.server.quorum.QuorumPeer.run(QuorumPeer.java:1108)
java.lang.InterruptedException: Timeout while waiting for epoch to be acked by quorum
    at org.apache.zookeeper.server.quorum.Leader.waitForEpochAck(Leader.java:1221)
    at org.apache.zookeeper.server.quorum.Leader.lead(Leader.java:487)
    at org.apache.zookeeper.server.quorum.QuorumPeer.run(QuorumPeer.java:1105)
2016-06-23 07:30:57,482 | INFO  | QuorumPeer[myid=25](plain=/10.16.9.136:2181)(secure=disabled) |
Shutdown called | org.apache.zookeeper.server.quorum.Leader.shutdown(Leader.java:623)
java.lang.Exception: shutdown Leader! reason: Forcing shutdown
    at org.apache.zookeeper.server.quorum.Leader.shutdown(Leader.java:623)
    at org.apache.zookeeper.server.quorum.QuorumPeer.stopLeader(QuorumPeer.java:1149)
    at org.apache.zookeeper.server.quorum.QuorumPeer.run(QuorumPeer.java:1110)
```

回答

创建大量节点后，follower与leader同步时数据量大，在集群数据同步限定时间内不能完成同步过程，导致超时，各个ZooKeeper server启动失败。

参考**修改集群服务配置参数**章节，进入ZooKeeper服务“全部配置”页面。不断尝试调大ZooKeeper配置文件“zoo.cfg”中的“syncLimit”和“initLimit”两参数值，直到ZooKeeperServer正常。

表 12-478 参数说明

参数	描述	默认值
syncLimit	follower与leader进行同步的时间间隔（时长为ticket时长的倍数）。如果在该时间范围内leader没响应，连接将不能被建立。	15
initLimit	follower连接到leader并与leader同步的时间（时长为ticket时长的倍数）。	15

如果将参数“initLimit”和“syncLimit”的参数值均配置为“300”之后，ZooKeeper server仍然无法恢复，则需确认没有其他应用程序正在kill ZooKeeper。例如，参数值为“300”，ticket时长为2000毫秒，即同步限定时间为300*2000ms=600s。

可能存在以下场景，在ZooKeeper中创建的数据过大，需要大量时间与leader同步，并保存到硬盘。在这个过程中，如果ZooKeeper需要运行很长时间，则需确保没有其他监控应用程序kill ZooKeeper而判断其服务停止。

12.28.6.2 为什么 ZooKeeper Server 出现 java.io.IOException: Len 的错误日志

问题

在父目录中创建大量的znode之后，当ZooKeeper客户端尝试在单个请求中获取该父目录中的所有子节点时，将返回失败。

客户端日志，如下所示：

```
2017-07-11 13:17:19,610 [myid:] - WARN  [New I/O worker #3:ClientCnxnSocketNetty
$ZKClientHandler@468] - Exception caught: [id: 0xb66cbb85, /10.18.97.97:49192 :>
```

```
10.18.97.97/10.18.97.97:2181] EXCEPTION: java.nio.channels.ClosedChannelException
java.nio.channels.ClosedChannelException
at org.jboss.netty.handler.ssl.SslHandler$6.run(SslHandler.java:1580)
at org.jboss.netty.channel.socket.ChannelRunnableWrapper.run(ChannelRunnableWrapper.java:40)
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.executeInIoThread(AbstractNioWorker.java:71)
at org.jboss.netty.channel.socket.nio.NioWorker.executeInIoThread(NioWorker.java:36)
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.executeInIoThread(AbstractNioWorker.java:57)
at org.jboss.netty.channel.socket.nio.NioWorker.executeInIoThread(NioWorker.java:36)
at org.jboss.netty.channel.socket.nio.AbstractNioChannelSink.execute(AbstractNioChannelSink.java:34)
at org.jboss.netty.handler.ssl.SslHandler.channelClosed(SslHandler.java:1566)
at org.jboss.netty.channel.Channels.fireChannelClosed(Channels.java:468
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.close(AbstractNioWorker.java:376)
at org.jboss.netty.channel.socket.nio.NioWorker.read(NioWorker.java:93)
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.process(AbstractNioWorker.java:109)
at org.jboss.netty.channel.socket.nio.AbstractNioSelector.run(AbstractNioSelector.java:312)
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.run(AbstractNioWorker.java:90)
at org.jboss.netty.channel.socket.nio.NioWorker.run(NioWorker.java:178)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
```

Leader节点的日志，如下所示：

```
2017-07-11 13:17:33,043 [myid:1] - WARN  [New I/O worker #7:NettyServerCnxn@445] - Closing
connection to /10.18.101.110:39856
java.io.IOException: Len error 45
at org.apache.zookeeper.server.NettyServerCnxn.receiveMessage(NettyServerCnxn.java:438)
at org.apache.zookeeper.server.NettyServerCnxnFactory
$CnxnChannelHandler.processMessage(NettyServerCnxnFactory.java:267)
at org.apache.zookeeper.server.NettyServerCnxnFactory
$CnxnChannelHandler.messageReceived(NettyServerCnxnFactory.java:187)
at org.jboss.netty.channel.SimpleChannelHandler.handleUpstream(SimpleChannelHandler.java:88)
at org.jboss.netty.channel.DefaultChannelPipeline.sendUpstream(DefaultChannelPipeline.java:564)
at org.jboss.netty.channel.DefaultChannelPipeline.sendUpstream(DefaultChannelPipeline.java:559)
at org.jboss.netty.channel.Channels.fireMessageReceived(Channels.java:268)
at org.jboss.netty.channel.Channels.fireMessageReceived(Channels.java:255)
at org.jboss.netty.channel.socket.nio.NioWorker.read(NioWorker.java:88)
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.process(AbstractNioWorker.java:109)
at org.jboss.netty.channel.socket.nio.AbstractNioSelector.run(AbstractNioSelector.java:312)
at org.jboss.netty.channel.socket.nio.AbstractNioWorker.run(AbstractNioWorker.java:90)
at org.jboss.netty.channel.socket.nio.NioWorker.run(NioWorker.java:178)
at org.jboss.netty.util.ThreadRenamingRunnable.run(ThreadRenamingRunnable.java:108)
at org.jboss.netty.util.internal.DeadLockProofWorker$1.run(DeadLockProofWorker.java:42)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
```

回答

在单个父目录中创建大量的znode后，当客户端尝试在单个请求中获取所有子节点时，服务端将无法返回，因为结果将超出可存储在znode上的数据的最大长度。

为了避免这个问题，应该根据客户端应用的实际情况将“jute.maxbuffer”参数配置为一个更高的值。

“jute.maxbuffer”只能设置为Java系统属性，且没有zookeeper前缀。如果要将“jute.maxbuffer”的值设为X，在ZooKeeper客户端或服务端启动时传入以下系统属性：-Djute.maxbuffer=X。

例如，将参数值设置为4MB：-Djute.maxbuffer=0x400000。

表 12-479 配置参数

参数	描述	默认值
jute.maxbuffer	指定可以存储在znode中的数据的最大长度。单位是Byte。默认值为0xfffff，即低于1MB。 **说明** 如果更改此选项，则必须在所有服务器和客户端上设置该系统属性，否则将出现问题。	0xfffff

12.28.6.3 为什么在 Zookeeper 服务器上启用安全的 netty 配置时，四个字母的命令不能与 linux 的 netcat 命令一起使用

问题

为什么在Zookeeper服务器上启用安全的netty配置时，四个字母的命令不能与linux的***netcat***命令一起使用?

例如：

echo stat /netcat host port

回答

Linux的***netcat***命令没有与Zookeeper服务器安全通信的选项，所以当启用安全的netty配置时，它不能支持Zookeeper四个字母的命令。

为了避免这个问题，用户可以使用下面的Java API来执行四个字母的命令。

```
org.apache.zookeeper.client.FourLetterWordMain
```

例如：

```
String[] args = new String[]{host, port, "stat"};
org.apache.zookeeper.client.FourLetterWordMain.main(args);
```

说明

netcat命令只能用于非安全的netty配置。

12.28.6.4 如何查看哪个 ZooKeeper 实例是 leader

问题

如何查看ZooKeeper实例的角色是leader还是follower?

回答

登录Manager，选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 实例”，单击相应的quorumpeer实例名称，进入对应实例的详情页面，即可查看到该实例的“服务器状态”。

12.28.6.5 使用 IBM JDK 时客户端无法连接 ZooKeeper

问题

使用IBM的JDK的情况下客户端连接ZooKeeper失败。

回答

可能原因为IBM的JDK和普通JDK的jaas.conf文件格式不一样。

在使用IBM JDK时，建议使用如下jaas.conf文件模板，其中“useKeytab”中的文件路径必须以“file://”开头，后面为绝对路径。

```
Client {
com.ibm.security.auth.module.Krb5LoginModule required
useKeytab="file://D:/install/HbaseClientSample/conf/user.keytab"
principal="hbaseuser1"
credsType="both";
};
```

12.28.6.6 ZooKeeper 客户端刷新 TGT 失败

问题

ZooKeeper客户端刷新TGT失败，无法连接ZooKeeper。报错内容如下：

```
Login: Could not renew TGT due to problem running shell command: '***/kinit -R'; exception was:org.apache.zookeeper.Shell$ExitCodeException: kinit: Ticket expired while renewing credentials
```

回答

ZooKeeper使用系统命令**kinit -R**对票据进行刷新，当前MRS版本已经取消了该命令的功能，如需运行长任务，建议使用keytab方式完成鉴权功能。

在“jaas.conf”配置文件中设置属性“useTicketCache=false”，设置“useKeyTab=true”，并指明keytab路径。

12.28.6.7 使用 deleteall 命令，删除大量 znode 时，偶现报错“Node does not exist”错误

问题

客户端连接非leader实例，使用deleteall命令删除大量znode时，报错Node does not exist，但是stat命令能够获取到node状态。

回答

由于网络问题或者数据量大导致leader和follower数据不同步。解决方法是客户端连接到leader实例进行删除操作。具体过程是首先根据**如何查看哪个ZooKeeper实例是leader**查看leader所在节点IP，使用连接客户端命令zkCli.sh -server leader节点IP:2181成功后进行deleteall命令删除操作，具体操作请参见**使用ZooKeeper客户端**。

12.29 附录

12.29.1 修改集群服务配置参数

- 用户可直接通过MRS管理控制台的集群管理页面修改各服务配置参数：
  a. 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。
  b. 选择“组件管理 > *服务名称* > 服务配置”。
     默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。
  c. 在导航树选择指定的参数分类，并在右侧修改参数值。
     不确定参数的具体位置时，支持在右上角输入参数名，系统将实时进行搜索并显示结果。
  d. 单击“保存配置”，并在确认对话框中单击“是”。
  e. 等待界面提示“操作成功”，单击“完成”，配置已修改。
     查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。也可在保存配置时直接勾选“重新启动受影响的服务或实例。”。
- MRS 3.x之前的版本，服务配置参数均支持登录MRS Manager进行修改:
  a. 登录MRS Manager。
  b. 单击“服务管理”。
  c. 单击服务视图中指定的服务名称。
  d. 单击“服务配置”。
     默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。
  e. 在导航树选择指定的参数分类，并在右侧修改参数值。
     不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。
  f. 单击“保存配置”，并在确认对话框中单击“是”。
  g. 等待界面提示“操作成功”，单击“完成”，配置已修改。
     查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。也可在保存配置时直接勾选“重新启动受影响的服务或实例。”
- MRS 3.x及后续版本，服务配置参数均支持登录FusionInsight Manager进行修改：
  a. 登录FusionInsight Manager。
  b. 选择“集群 > 服务”。
  c. 单击服务视图中指定的服务名称。
  d. 单击“配置”。
     默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。
  e. 在导航树选择指定的参数分类，并在右侧修改参数值。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

f. 单击“保存”，并在确认对话框中单击“确定”。

g. 等待界面提示“操作成功”，单击“完成”，配置已修改。

查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。

12.29.2 访问集群 Manager

12.29.2.1 访问 MRS Manager（MRS 3.x 之前版本）

操作场景

MRS 3.x之前版本集群使用MRS Manager对集群进行监控、配置和管理，用户可以在MRS控制台页面打开Manager管理页面。

访问 MRS Manager

步骤1 登录MRS管理控制台页面。

步骤2 单击“集群列表 > 现有集群”，在集群列表中单击指定的集群名称，进入集群信息页面。

步骤3 单击“前往 Manager”，打开“访问MRS Manager页面”。

- 若用户创建集群时已经绑定弹性公网IP：
  a. 选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
  b. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则，如需开放多个IP段为可信范围用于访问MRS Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

  说明

  - 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
  - 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问MRS Manager服务。

  c. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”
- 若用户创建集群时暂未绑定弹性公网IP：
  a. 在弹性公网IP下拉框中选择可用的弹性公网IP或单击“管理弹性公网IP”创建弹性公网IP。
  b. 选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
  c. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则，如需开放多个IP段为可信范围用于访问MRS Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问MRS Manager服务。

d. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

步骤4 单击“确定”，进入MRS Manager登录页面。

步骤5 输入默认用户名“admin”及创建集群时设置的密码，单击“登录”进入MRS Manager页面。

步骤6 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

说明

如需给其他用户开通访问MRS Manager的权限，请执行**步骤6**-**步骤9**，添加对应用户访问公网的IP地址为可信范围。

步骤7 单击弹性公网IP后边的“添加安全组规则”。

步骤8 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的授权对象是可信任的公网访问IP范围，禁止使用0.0.0.0/0,否则会有安全风险。”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤6**-**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤9 单击“确定”完成安全组规则添加。

----结束

为其他用户开通访问 MRS Manager 的权限

步骤1 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤2 单击弹性公网IP后边的“添加安全组规则”。

步骤3 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的授权对象是可信任的公网访问IP范围，禁止使用0.0.0.0/0,否则会有安全风险。”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤1**-**步骤4**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤4 单击“确定”完成安全组规则添加。

----结束

12.29.2.2 访问 FusionInsight Manager（MRS 3.x 及之后版本）

操作场景

MRS 3.x及之后版本的集群使用FusionInsight Manager对集群进行监控、配置和管理。用户在集群安装后可使用帐号登录FusionInsight Manager。

说明

如果不能正常登录组件的WebUI页面，请参考**通过ECS访问FusionInsight Manager**方式访问FusionInsight Manager。

通过弹性 IP 访问 FusionInsight Manager

步骤1 登录MRS管理控制台页面。

步骤2 单击“集群列表 > 现有集群”，在集群列表中单击指定的集群名称，进入集群信息页面。

步骤3 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中配置弹性IP信息。

1. 若创建MRS集群时暂未绑定弹性公网IP，在“弹性公网IP”下拉框中选择可用的弹性公网IP。若用户创建集群时已经绑定弹性公网IP，直接执行**步骤3.2**

   说明

   如果没有弹性公网IP，可先单击“管理弹性公网IP”弹性公网IP后，然后在弹性公网IP下拉框中选择的弹性公网IP。

2. 在“安全组”中选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
3. 添加安全组规则，默认填充的是用户访问弹性IP地址的规则，如需开放多个IP段为可信范围用于访问Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。
4. 勾选确认信息后，单击“确定”。

步骤4 单击“确定”，进入Manager登录页面。

步骤5 输入默认用户名“admin”及创建集群时设置的密码，单击“登录”进入Manager页面。

步骤6 在MRS管理控制台，在“现有集群” 列表，单击指定的集群名称，进入集群信息页面。

说明

如需给其他用户开通访问Manager的权限，请执行**步骤6**～**步骤9**，添加对应用户访问公网的IP地址为可信范围。

步骤7 单击弹性公网IP后边的“添加安全组规则”。

步骤8 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的*公网IP*/*端口号*是可信任的公网访问IP范围，我了解使用0.0.0.0/0会带来安全风险”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤6**-**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤9 单击“确定”完成安全组规则添加。

----结束

通过 ECS 访问 FusionInsight Manager

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在“现有集群”列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“集群管理页面”、“安全组”。

步骤3 在管理控制台首页服务列表中选择“弹性云服务器”，进入ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个标准镜像“Windows Server 2012 R2 Standard 64bit(40GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和Master节点的“默认安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为Master节点的默认安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的帐号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

例如Windows 2012操作系统可以使用Internet Explorer 11。

Manager访问地址为“集群管理页面”地址。访问时需要输入集群的用户名和密码，例如“admin”用户。

说明

- 如果使用其他集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。请咨询管理员。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

----结束

12.29.3 使用 MRS 客户端

12.29.3.1 安装客户端（3.x 及之后版本）

操作场景

该操作指导安装工程师安装MRS集群所有服务（不包含Flume）的客户端。Flume客户端安装请参见“组件操作指南 > 使用Flume > 安装Flume客户端”。

客户端可以安装集群内节点，也可以安装在集群外节点，本章节以安装目录“/opt/client”为例进行介绍，请以实际集群版本为准。

在集群外节点安装客户端前提条件

- 已准备一个Linux弹性云服务器，主机操作系统及版本建议参见**表12-480**。

  表 12-480 参考列表

CPU架构	操作系统	支持的版本号
x86计算	Euler	Euler OS 2.5
	SuSE	SUSE Linux Enterprise Server 12 SP4（SUSE 12.4）
	Red Hat	Red Hat-7.5-x86_64（Red Hat 7.5）
	CentOS	CentOS-7.6版本（CentOS 7.6）
鲲鹏计算(ARM)	Euler	Euler OS 2.8
	CentOS	CentOS-7.6版本（CentOS 7.6）

  同时为弹性云服务分配足够的磁盘空间，例如“40GB”。
- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。
- 弹性云服务器操作系统已安装NTP服务，且NTP服务运行正常。

  若未安装，在配置了**yum**源的情况下，可执行**yum install ntp -y**命令自行安装。
- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式）。

集群内节点安装客户端

1. 获取软件包。

   访问FusionInsight Manager（MRS 3.x及之后版本），在“集群”下拉列表中单击需要操作的集群名称。

   选择“更多 > 下载客户端”，弹出“下载集群客户端”信息提示框。

   说明

   在只安装单个服务的客户端的场景中，选择“集群 > 服务 > *服务名称* > 更多 > 下载客户端”，弹出“下载客户端”信息提示框。

2. “选择客户端类型”中选择“完整客户端”。

   “仅配置文件”下载的客户端配置文件，适用于应用开发任务中，完整客户端已下载并安装后，管理员通过Manager界面修改了服务端配置，开发人员需要更新客户端配置文件的场景。

   平台类型包括x86_64和aarch64两种：

   - x86_64：可以部署在X86平台的客户端软件包。
   - aarch64：可以部署在TaiShan服务器的客户端软件包。

   说明

   集群支持下载x86_64和aarch64两种类型客户端，但是客户端类型必须与待安装节点的架构匹配，否则客户端会安装失败。

3. 勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件。

   文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。单击“确定”，等待下载完成后，使用**omm**用户或**root**用户将获取的软件包复制到将要安装客户端的服务器文件目录。

   客户端软件包名称格式为：“FusionInsight_Cluster_*<集群ID>*_Services_Client.tar”。

   后续步骤及章节以FusionInsight_Cluster_1_Services_Client.tar进行举例。

   说明

   当用户无法获取**root**用户权限，需要用**omm**用户操作。

   如需安装客户端至集群内其他节点，则执行以下命令复制客户端到待安装客户端的节点：

   scp -p /*tmp/FusionInsight-Client*/**FusionInsight_Cluster_1_Services_Client.tar** *待安装客户端节点的IP地址:/opt/Bigdata/client*

4. 以**user_client**用户登录将要安装客户端的服务器。
5. 解压软件包。

   进入安装包所在目录，例如“/tmp/FusionInsight-Client”。执行如下命令解压安装包到本地目录。

   tar -xvf FusionInsight_Cluster_1_Services_Client.tar

6. 校验软件包。

   执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

   sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256

   ```
   FusionInsight_Cluster_1_Services_ClientConfig.tar: OK
   ```

7. 解压获取的安装文件。

   tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

8. 进入安装包所在目录，执行如下命令安装客户端到指定目录（绝对路径），例如安装到“/opt/client”目录。

   cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig

   执行**./install.sh /opt/client**命令，等待客户端安装完成（以下只显示部分屏显结果）。

   ```
   The component client is installed successfully
   ```

说明

- 如果已经安装的全部服务或某个服务的客户端使用了“/opt/client”目录，再安装其他服务的客户端时，需要使用不同的目录。
- 卸载客户端请删除客户端安装目录。
- 如果要求安装后的客户端仅能被该安装用户（如“user_client”）使用，请在安装时加“-o”参数，即执行**./install.sh /opt/client -o**命令安装客户端。
- 由于HBase使用的Ruby语法限制，如果安装的客户端中包含了HBase客户端，建议客户端安装目录路径只包含大写字母、小写字母、数字以及_-?.@+=字符。

使用客户端

1. 在已安装客户端的节点，执行**sudo su - omm**命令切换用户。执行以下命令切换到客户端目录：

   cd /opt/client

2. 执行以下命令配置环境变量：

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit admin**。

   说明

   启用Kerberos认证的MRS集群默认创建“admin”用户帐号，用于集群管理员维护集群。

4. 直接执行组件的客户端命令。

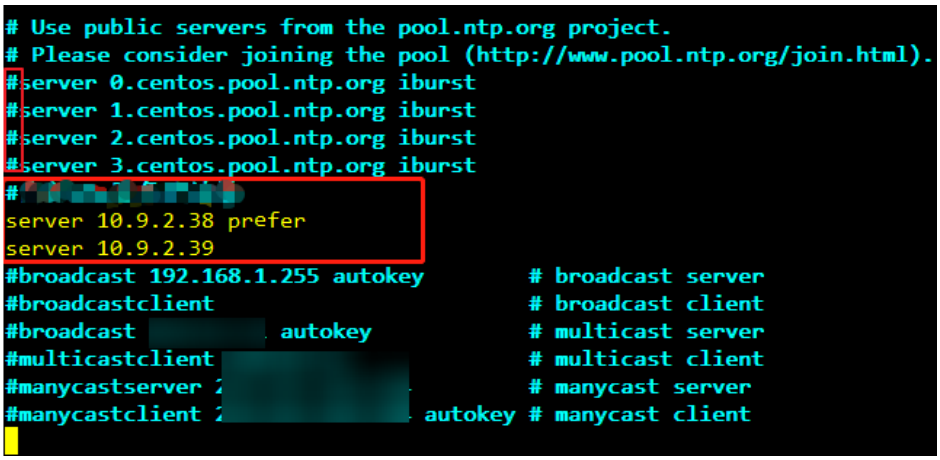
   例如：使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

集群外节点安装客户端

1. 根据**在集群外节点安装客户端前提条件**，创建一个满足要求的弹性云服务器。
2. 执行ntp时间同步，使集群外节点的时间与MRS集群时间同步。
   a. 执行**vi /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

   ```
   server master1_ip prefer
   server master2_ip
   ```

   图 12-74 增加 Master 节点的 IP



b. 执行**service ntpd stop**命令关闭NTP服务。
c. 执行**/usr/sbin/ntpdate** *主Master节点的IP地址*命令手动同步一次时间。
d. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。
e. 执行**ntpstat**命令查看时间同步结果。

3. 参考以下步骤，从FusionInsight Manager下载集群客户端软件包并复制到ECS节点后安装客户端。
   a. **访问FusionInsight Manager（MRS 3.x及之后版本）**，参考**集群内节点安装客户端**下载集群客户端到主管理节点的指定目录。
   b. 使用**root**用户登录主管理节点，执行以下命令复制客户端安装包到待安装客户端的节点：

      scp -p /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar *待安装客户端节点的IP地址*:**/tmp**
   c. 使用待安装客户端的用户登录待安装客户端节点。

      执行以下命令安装客户端，如果当前用户无客户端软件包以及客户端安装目录的操作权限，需使用**root**用户进行赋权：

      cd /tmp

      tar -xvf FusionInsight_Cluster_1_Services_Client.tar

      tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

      cd FusionInsight_Cluster_1_Services_ClientConfig

      ./install.sh /opt/client
   d. 执行以下命令，切换到客户端目录并配置环境变量：

      cd /opt/client

      source bigdata_env
   e. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

      kinit *MRS集群用户*

      例如，**kinit admin**。
   f. 直接执行组件的客户端命令。

      例如使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

12.29.3.2 安装客户端（3.x 之前版本）

操作场景

用户需要使用MRS客户端。MRS集群客户端可以安装在集群内的Master节点或者Core节点，也可以安装在集群外节点上。

MRS 3.x之前版本集群在集群创建后，在主Master节点默认安装有客户端，可以直接使用，安装目录为“/opt/client”。

MRS 3.x及之后版本客户端的安装请参考**安装客户端（3.x及之后版本）**。

说明

如果集群外的节点已安装客户端且只需要更新客户端，请使用安装客户端的用户例如“root”。

在集群外节点安装客户端前提条件

- 已准备一个弹性云服务器，主机操作系统及版本请参见**表12-481**。

  表 12-481 参考列表

操作系统	支持的版本号
Euler	• 可用：Euler OS 2.2 • 可用：Euler OS 2.3 • 可用：Euler OS 2.5

  例如，用户可以选择操作系统为**Euler**的弹性云服务器准备操作。

  同时为弹性云服务分配足够的磁盘空间，例如“40GB”。
- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。

  如果不同，请修改弹性云服务器安全组或配置弹性云服务器安全组的出入规则允许MRS集群所有安全组的访问。
- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式），请参见弹性云服务器《用户指南》中“实例> 登录Linux弹性云服务器 >SSH密码方式登录”。

在 Core 节点安装客户端

1. 登录MRS Manager页面，选择“服务管理 > 下载客户端”下载客户端安装包至主管理节点。

   说明

   如仅需更新客户端配置文件，请参考**更新客户端（3.x之前版本）**页面的方法二操作。
2. 使用IP地址搜索主管理节点并使用VNC登录主管理节点。
3. 在主管理节点，执行以下命令切换用户。

   sudo su - omm
4. 在MRS管理控制台，查看指定集群“节点管理”页面的“IP”地址。

   记录需使用客户端的Core节点IP地址。
5. 在主管理节点，执行以下命令，将客户端安装包从主管理节点文件拷贝到当前Core节点：

   scp -p /tmp/MRS-client/MRS_Services_Client.tar *Core节点的IP地址*:**/opt/client**
6. 使用“root”登录Core节点。

   Master节点支持Cloud-Init特性，Cloud-init预配置的用户名“root”，密码为创建集群时设置的密码。
7. 执行以下命令，安装客户端：

   cd /opt/client

   tar -xvf MRS_Services_Client.tar

   tar -xvf MRS_Services_ClientConfig.tar

   cd /opt/client/MRS_Services_ClientConfig

./install.sh *客户端安装目录*

例如，执行命令：

./install.sh /opt/client

8. 客户端的使用请参见**使用MRS客户端**。

使用 MRS 客户端

1. 在已安装客户端的节点，执行**sudo su - omm**命令切换用户。执行以下命令切换到客户端目录：

   cd /opt/client

2. 执行以下命令配置环境变量：

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit admin**。

   说明

   启用Kerberos认证的MRS集群默认创建“admin”用户帐号，用于集群管理员维护集群。

4. 直接执行组件的客户端命令。

   例如：使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

在集群外节点上安装客户端

步骤1 根据前提条件，创建一个满足要求的弹性云服务器。

步骤2 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 3.x之前版本）**，然后选择“服务管理”。

步骤3 单击“下载客户端”。

步骤4 在“客户端类型”选择“完整客户端”。

步骤5 在“下载路径”选择“远端主机”。

步骤6 将“主机IP”设置为ECS的IP地址，设置“主机端口”为“22”，并将“存放路径”设置为“/tmp”。

- 如果使用SSH登录ECS的默认端口“22”被修改，请将“主机端口”设置为新端口。
- “存放路径”最多可以包含256个字符。

步骤7 “登录用户”设置为“root”。

如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。

步骤8 在“登录方式”选择“密码”或“SSH私钥”。

- 密码：输入创建集群时设置的root用户密码。
- SSH私钥：选择并上传创建集群时使用的密钥文件。

步骤9 单击“确定”开始生成客户端文件。

若界面显示以下提示信息表示客户端包已经成功保存。单击“关闭”。客户端文件请到下载客户端时设置的远端主机的“存放路径”中获取。

```
下载客户端文件到远端主机成功。
```

若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从步骤2执行重新开始下载客户端。

```
连接到服务器失败，请检查网络连接或参数设置。
```

说明

生成客户端会占用大量的磁盘IO，不建议在集群处于安装中、启动中、打补丁中等非稳态场景下载客户端。

步骤10 使用VNC方式，登录弹性云服务器。参见弹性云服务器《用户指南》的**远程登录（VNC方式）**章节（“实例 > 登录Linux弹性云服务器 > 远程登录（VNC方式）”）。

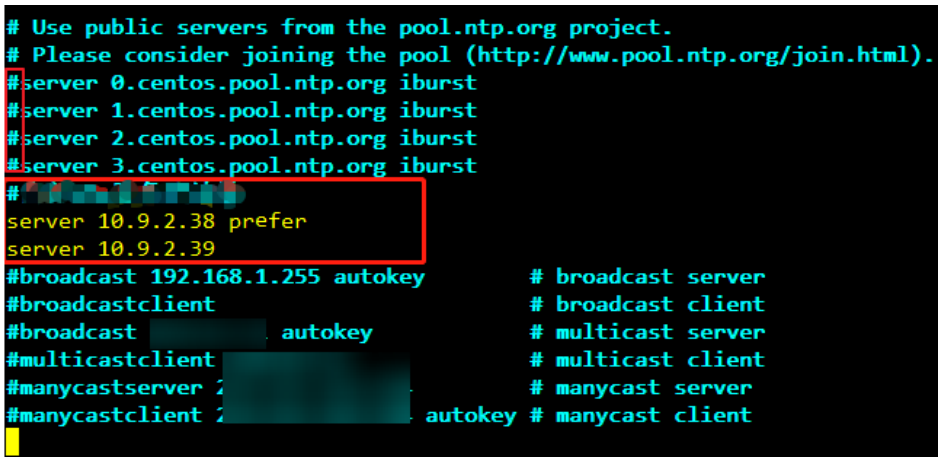
所有镜像均支持Cloud-init特性。Cloud-init预配置的用户名“root”，密码为创建集群时设置的密码。首次登录建议修改。

步骤11 执行ntp时间同步，使集群外节点的时间与MRS集群时间同步。

1. 检查安装NTP服务有没有安装，未安装请执行**yum install ntp -y**命令自行安装。
2. 执行**vim /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

   ```
   server master1_ip prefer
   server master2_ip
   ```

   图 12-75 增加 Master 节点的 IP

3. 执行**service ntpd stop**命令关闭NTP服务。
4. 执行**/usr/sbin/ntpdate *主Master节点的IP***命令手动同步一次时间。
5. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。
6. 执行**ntpstat**命令查看时间同步结果。

步骤12 在弹性云服务器，切换到**root**用户，并将步骤6中“存放路径”中的安装包复制到目录“/opt”，例如“存放路径”设置为“/tmp”时命令如下。

sudo su - root

cp /tmp/MRS_Services_Client.tar /opt

步骤13 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Services_Client.tar

步骤14 执行以下命令，校验文件包。

sha256sum -c MRS_Services_ClientConfig.tar.sha256

界面显示如下：

```
MRS_Services_ClientConfig.tar: OK
```

步骤15 执行以下命令，解压“MRS_Services_ClientConfig.tar”。

tar -xvf MRS_Services_ClientConfig.tar

步骤16 执行以下命令，安装客户端到新的目录，例如“/opt/Bigdata/client”。安装时自动生成目录。

sh /opt/MRS_Services_ClientConfig/install.sh /opt/Bigdata/client

查看安装输出信息，如有以下结果表示客户端安装成功：

```
Components client installation is complete.
```

步骤17 验证弹性云服务器节点是否与集群Master节点的IP是否连通?

例如，执行以下命令：**ping** *Master节点IP地址*

- 是，执行**步骤18**。
- 否，检查VPC、安全组是否正确，是否与MRS集群在相同VPC和安全组，然后执行**步骤18**。

步骤18 执行以下命令配置环境变量：

source /opt/Bigdata/client/bigdata_env

步骤19 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit*MRS集群用户*

例如, **kinit admin**

步骤20 执行组件的客户端命令。

例如，执行以下命令查看HDFS目录：

hdfs dfs -ls /

----结束

12.29.3.3 更新客户端（3.x 及之后版本）

集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户如果在Manager修改了服务配置参数并重启了服务，已安装的客户端需要重新下载并安装，或者使用配置文件更新客户端。

更新客户端配置

方法一：

步骤1 访问FusionInsight Manager（MRS 3.x及之后版本），在“集群”下拉列表中单击需要操作的集群名称。

步骤2 选择“更多 > 下载客户端 > 仅配置文件”。

此时生成的压缩文件包含所有服务的配置文件。

步骤3 是否在集群的节点中生成配置文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。然后执行**步骤4**。
- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，然后执行**步骤4**。

步骤4 使用WinSCP工具，以客户端安装用户将压缩文件保存到客户端安装的目录，例如“/opt/hadoopclient”。

步骤5 解压软件包。

例如下载的客户端文件为“FusionInsight_Cluster_1_Services_Client.tar”执行如下命令进入客户端所在目录，解压文件到本地目录。

cd /opt/hadoopclient

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤6 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar: OK
```

步骤7 解压获取配置文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar

步骤8 在客户端安装目录下执行如下命令，使用配置文件更新客户端。

sh refreshConfig.sh *客户端安装目录 配置文件所在目录*

例如，执行以下命令：

sh refreshConfig.sh /opt/hadoopclient /opt/hadoopclient/FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 以**root**用户登录客户端安装节点。

步骤2 进入客户端安装的目录，例如“/opt/hadoopclient”，执行以下命令更新配置文件：

cd /opt/hadoopclient

sh autoRefreshConfig.sh

步骤3 按照提示输入FusionInsight Manager管理员用户名，密码以及FusionInsight Manager界面浮动IP。

步骤4 输入需要更新配置的组件名，组件名之间使用“,”分隔。如需更新所有组件配置，可直接单击回车键。

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

12.29.3.4 更新客户端（3.x 之前版本）

说明

本章节适用于MRS 3.x之前版本的集群。MRS 3.x及之后版本，请参考**更新客户端（3.x及之后版本）**。

更新客户端配置文件

操作场景

MRS集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户使用MRS的客户端时，如果在MRS Manager修改了服务配置参数并重启了服务或者重启了服务，需要先下载并更新客户端配置文件。

用户创建集群时，默认在集群所有节点的“/opt/client”目录安装保存了原始客户端。集群创建完成后，仅Master节点的客户端可以直接使用，Core节点客户端在使用前需要更新客户端配置文件。

操作步骤

方法一：

步骤1 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 3.x之前版本）**，然后选择“服务管理”。

步骤2 单击“下载客户端”。

“客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤3 查询并登录主Master节点。

步骤4 若在集群内使用客户端，执行以下命令切换到omm用户，若在集群外使用客户端，请切换到root用户：

sudo su - omm

步骤5 执行以下命令切换客户端目录，例如“**/opt/Bigdata/client**”：

cd /opt/Bigdata/client

步骤6 执行以下命令，更新客户端配置：

sh refreshConfig.sh *客户端安装目录客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/Bigdata/client /tmp/MRS-client/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 集群安装完成之后，执行以下命令切换到omm用户，若在集群外使用客户端，请切换到root用户。

sudo su - omm

步骤2 执行以下命令切换客户端目录，例如“**/opt/Bigdata/client**”。

cd /opt/Bigdata/client

步骤3 执行以下命令并按照提示输入MRS Manager有下载权限的用户名和密码（例如，用户名为admin，密码为创建集群时设置的密码），更新客户端配置。

sh autoRefreshConfig.sh

步骤4 命令执行后显示如下信息，其中XXX表示集群安装的组件名称，如需更新全部组件配置，单击“Enter”键，如需更新部分组件配置，请输入需要更新的组件名称，多个组件名称以逗号相隔。

```
Components "xxx" have been installed in the cluster. Please input the comma-separated names of the components for which you want to update client configurations. If you press Enter without inputting any component name, the client configurations of all components will be updated:
```

界面显示以下信息表示配置更新成功：

```
Succeed to refresh components client config.
```

界面显示以下信息表示用户名或者密码错误：

```
login manager failed,Incorrect username or password.
```

说明

- 该脚本会自动连接到集群并调用refreshConfig.sh脚本下载并刷新客户端配置文件。
- 客户端默认使用安装目录下文件Version中的“wsom=xxx”所配置的浮动IP刷新客户端配置，如需刷新为其他集群的配置文件，请执行本步骤前修改Version文件中“wsom=xxx”的值为对应集群的浮动IP地址。

----结束

全量更新主 Master 节点的原始客户端

场景描述

用户创建集群时，默认在集群所有节点的“/opt/client”目录安装保存了原始客户端。以下操作以“/opt/Bigdata/client”为例进行说明。

- MRS普通集群，在console页面提交作业时，会使用master节点上预置安装的客户端进行作业提交。

- 用户也可使用master节点上预置安装的客户端来连接服务端、查看任务结果或管理数据等

对集群安装补丁后，用户需要重新更新master节点上的客户端，才能保证继续使用内置客户端功能。

操作步骤

步骤1 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 3.x之前版本）**，然后选择“服务管理”。

步骤2 单击“下载客户端”。

“客户端类型”选择“完整客户端”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤3 查询并登录主Master节点。

步骤4 在弹性云服务器，切换到**root**用户，并将安装包复制到目录“/opt”。

sudo su - root

cp /tmp/MRS-client/MRS_Services_Client.tar /opt

步骤5 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Services_Client.tar

步骤6 执行以下命令，校验文件包。

sha256sum -c MRS_Services_ClientConfig.tar.sha256

界面显示如下：

```
MRS_Services_ClientConfig.tar: OK
```

步骤7 执行以下命令，解压“MRS_Services_ClientConfig.tar”。

tar -xvf MRS_Services_ClientConfig.tar

步骤8 执行以下命令，移走原来老的客户端到/opt/Bigdata/client_bak目录下

mv /opt/Bigdata/client /opt/Bigdata/client_bak

步骤9 执行以下命令，安装客户端到新的目录，客户端路径必须为“/opt/Bigdata/client”。

sh /opt/MRS_Services_ClientConfig/install.sh /opt/Bigdata/client

查看安装输出信息，如有以下结果表示客户端安装成功：

```
Components client installation is complete.
```

步骤10 执行以下命令，修改/opt/Bigdata/client目录的所属用户和用户组。

chown omm:wheel /opt/Bigdata/client -R

步骤11 执行以下命令配置环境变量：

source /opt/Bigdata/client/bigdata_env

步骤12 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤13 执行组件的客户端命令。

例如，执行以下命令查看HDFS目录：

hdfs dfs -ls /

----结束

全量更新备 Master 节点的原始客户端

步骤1 参见**步骤1**~**步骤3**登录备Master节点，执行如下命令切换到omm用户。

sudo su - omm

步骤2 在备master节点上执行如下命令，从主master节点拷贝下载的客户端包。

scp omm@*master1节点IP地址*:/tmp/MRS-client/MRS_Services_Client.tar /tmp/MRS-client/

说明

- 该命令以master1节点为主master节点为例。
- 目的路径以备master节点的/tmp/MRS-client/目录为例，请根据实际路径修改。

步骤3 参见**步骤4**~**步骤13**，更新备Master节点的客户端。

----结束

13 安全性说明

13.1 集群（未启用 Kerberos 认证）安全配置建议

Hadoop社区版本提供两种认证方式Kerberos认证（安全模式）和Simple认证（普通模式），在创建集群时，MRS支持配置是否启用Kerberos认证。

在安全模式下MRS集群统一使用Kerberos认证协议进行安全认证。

而普通模式下MRS集群各组件使用原生开源的认证机制，一般为Simple认证方式。而Simple认证，在客户端连接服务端的过程中，默认以客户端执行用户（例如操作系统用户“root”等）自动完成认证，管理员或业务用户不显示感知认证。而且客户端在运行时，甚至可以通过注入UserGroupInformation来伪装成任意用户（包括superuser），集群资源管理接口和数据控制接口在服务端无认证和鉴权控制，很容易被黑客利用和攻击。

所以在普通模式下，必须通过严格限定网络访问权限来保障集群的安全。操作建议如下：

- 尽量将业务应用程序部署在同VPC和子网下的ECS中，避免通过外网访问MRS集群。
- 配置严格限制访问范围的安全组规则，禁止对MRS集群的入方向端口配置允许Any或0.0.0.0的访问规则。
- 如需从集群外访问集群内组件原生页面，请参考**创建连接MRS集群的SSH隧道并配置浏览器**进行配置。

13.2 安全认证原理和认证机制

功能

开启了 Kerberos认证的安全模式集群，进行应用开发时需要进行安全认证。

Kerberos作为安全认证的概念，使用Kerberos的系统在设计上采用“客户端/服务器”结构与AES等加密技术，并且能够进行相互认证（即客户端和服务器端均可对对方进行身份认证）。可以用于防止窃听、防止replay攻击、保护数据完整性等场合，是一种应用对称密钥体制进行密钥管理的系统。

结构

Kerberos的原理架构如图13-1所示，各模块的说明如表13-1所示。

图 13-1 原理架构

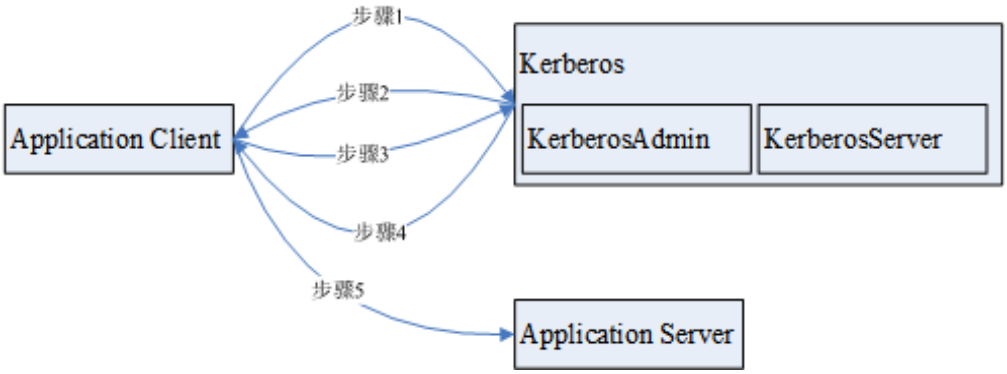


表 13-1 模块说明

模块	说明
Application Client	应用客户端，通常是需要提交任务（或者作业）的应用程序。
Application Server	应用服务端，通常是应用客户端需要访问的应用程序。
Kerberos	提供安全认证的服务。
KerberosAdmin	提供认证用户管理的进程。
KerberosServer	提供认证票据分发的进程。

步骤原理说明：

应用客户端（Application Client）可以是集群内某个服务，也可以是客户二次开发的一个应用程序，应用程序可以向应用服务提交任务或者作业。

1. 应用程序在提交任务或者作业前，需要向Kerberos服务申请TGT（Ticket-Granting Ticket），用于建立和Kerberos服务器的安全会话。
2. Kerberos服务在收到TGT请求后，会解析其中的参数来生成对应的TGT，使用客户端指定的用户名的密钥进行加密响应消息。
3. 应用客户端收到TGT响应消息后，解析获取TGT，此时，再由应用客户端（通常是rpc底层）向Kerberos服务获取应用服务端的ST（Server Ticket）。
4. Kerberos服务在收到ST请求后，校验其中的TGT合法后，生成对应的应用服务的ST，再使用应用服务密钥将响应消息进行加密处理。
5. 应用客户端收到ST响应消息后，将ST打包到发给应用服务的消息里面传输给对应的应用服务端（Application Server）。
6. 应用服务端收到请求后，使用本端应用服务对应的密钥解析其中的ST，并校验成功后，本次请求合法通过。

基本概念

以下为常见的基本概念，可以帮助用户减少学习Kerberos框架所花费的时间，有助于更好的理解Kerberos业务。以HDFS安全认证为例：

TGT

票据授权票据（ Ticket-Granting Ticket ），由Kerberos服务生成，提供给应用程序与Kerberos服务器建立认证安全会话，该票据的默认有效期为24小时，24小时后该票据自动过期。

TGT申请方式(以HDFS为例)：

1. 通过HDFS提供的接口获取。

```
/**
 * login Kerberos to get TGT, if the cluster is in security mode
 * @throws IOException if login is failed
 */
 private void login() throws IOException {
 // not security mode, just return
   if (! "kerberos".equalsIgnoreCase(conf.get("hadoop.security.authentication"))) {
       return;
   }

   //security mode
   System.setProperty("java.security.krb5.conf", PATH_TO_KRB5_CONF);

   UserGroupInformation.setConfiguration(conf);
   UserGroupInformation.loginUserFromKeytab(PRNCIPAL_NAME, PATH_TO_KEYTAB);
 }
```

2. 通过客户端shell命令以**kinit**方式获取。

ST

服务票据（ Server Ticket ），由Kerberos服务生成，提供给应用程序与应用服务建立安全会话，该票据一次性有效。

ST的生成在FusionInsight产品中，基于hadoop-rpc通信，由rpc底层自动向Kerberos服务端提交请求，由Kerberos服务端生成。

认证代码实例讲解

```
package com.xxx.bigdata.hdfs.examples;

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileStatus;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.security.UserGroupInformation;

public class KerberosTest {
    private static String PATH_TO_HDFS_SITE_XML = KerberosTest.class.getClassLoader().getResource("hdfs-site.xml")
            .getPath();
    private static String PATH_TO_CORE_SITE_XML = KerberosTest.class.getClassLoader().getResource("core-site.xml")
            .getPath();
    private static String PATH_TO_KEYTAB = KerberosTest.class.getClassLoader().getResource("user.keytab").getPath();
    private static String PATH_TO_KRB5_CONF = KerberosTest.class.getClassLoader().getResource("krb5.conf").getPath();
    private static String PRNCIPAL_NAME = "develop";
    private FileSystem fs;
```

```
    private Configuration conf;

    /**
     * initialize Configuration
     */
    private void initConf() {
        conf = new Configuration();

        // add configuration files
        conf.addResource(new Path(PATH_TO_HDFS_SITE_XML));
        conf.addResource(new Path(PATH_TO_CORE_SITE_XML));
    }

    /**
     * login Kerberos to get TGT, if the cluster is in security mode
     * @throws IOException if login is failed
     */
    private void login() throws IOException {
        // not security mode, just return
        if (! "kerberos".equalsIgnoreCase(conf.get("hadoop.security.authentication"))) {
            return;
        }

        //security mode
        System.setProperty("java.security.krb5.conf", PATH_TO_KRB5_CONF);

        UserGroupInformation.setConfiguration(conf);
        UserGroupInformation.loginUserFromKeytab(PRNCIPAL_NAME, PATH_TO_KEYTAB);
    }

    /**
     * initialize FileSystem, and get ST from Kerberos
     * @throws IOException
     */
    private void initFileSystem() throws IOException {
        fs = FileSystem.get(conf);
    }

    /**
     * An example to access the HDFS
     * @throws IOException
     */
    private void doSth() throws IOException {
        Path path = new Path("/tmp");
        FileStatus fStatus = fs.getFileStatus(path);
        System.out.println("Status of " + path + " is " + fStatus);
        //other thing
    }


    public static void main(String[] args) throws Exception {
        KerberosTest test = new KerberosTest();
        test.initConf();
        test.login();
        test.initFileSystem();
        test.doSth();
    }
}
```

说明

1. Kerberos认证时需要配置Kerberos认证所需要的文件参数，主要包含keytab路径，Kerberos认证的用户名称，Kerberos认证所需要的客户端配置krb5.conf文件。
2. 方法login()为调用hadoop的接口执行Kerberos认证，生成TGT票据。
3. 方法doSth()调用hadoop的接口访问文件系统，此时底层RPC会自动携带TGT去Kerberos认证，生成ST票据。

14 高危操作一览表

禁用操作

表14-1中描述了在集群操作与维护阶段，观察进行日常操作时应注意的禁用操作。

表 14-1 禁用操作

类别	操作风险
严禁删除ZooKeeper相关数据目录	ClickHouse/HDFS/Yarn/HBase/Hive等很多组件都依赖于ZooKeeper，在ZooKeeper中保存元数据信息。删除ZooKeeper中相关数据目录将会影响相关组件的正常运行。
严禁JDBCServer主备节点频繁倒换	频繁主备倒换将导致业务中断。
严禁删除Phoenix系统表或系统表数据(SYSTEM.CATALOG、SYSTEM.STATS、SYSTEM.SEQUENCE、SYSTEM.FUNCTION)	删除系统表将导致无法正常进行业务操作。
严禁手动修改Hive元数据库的数据（ hivemeta数据库 ）	修改Hive元数据可能会导致Hive数据解析错误，Hive无法正常提供服务。
严禁修改Hive私有文件目录hdfs:///tmp/hive-scratch的权限	修改该目录权限可能会导致Hive服务不可用。
严禁修改Kafka配置文件中broker.id	修改Kafka配置文件中broker.id将会导致该节点数据失效。
严禁修改节点主机名	主机名修改后会导致该主机上相关实例和上层组件无法正常提供服务，且无法修复。
禁止重装节点OS	该操作会导致MRS集群进入异常状态，影响MRS集群使用。
禁止使用私有镜像	该操作会导致MRS集群进入异常状态，影响MRS集群使用。

以下各表分别描述了各组件在操作与维护阶段，进行日常操作时应注意的高危操作。

集群高危操作

表 14-2 集群高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
随意修改omm用户下的文件目录或者文件权限	该操作会导致MRS集群服务不可用	★ ★ ★ ★ ★	请勿执行该操作	观察MRS集群服务是否可用
绑定弹性公网IP	该操作会将集群的manager所在的master节点暴露在公网，会增大来自互联网的网络攻击风险可能性	★ ★ ★ ★ ★	请确认绑定的弹性公网IP为可信任的公网访问IP	无
开放集群22端口安全组规则	该操作会增大用户利用22进行漏洞攻击的风险	★ ★ ★ ★ ★	针对开放的22端口进行设置安全组规则，只允许可信的IP可以访问该端口，入方向规则不推荐设置允许0.0.0.0可以访问。	无
删除集群或删除集群数据	该操作会导致数据丢失	★ ★ ★ ★ ★	删除前请务必再次确认该操作的必要性，同时要保证数据已完成备份	无
缩容集群	该操作会导致数据丢失	★ ★ ★ ★ ★	缩容前请务必再次确认该操作的必要性，同时要保证数据已完成备份	无
卸载磁盘或格式化数据盘	该操作会导致数据丢失	★ ★ ★ ★ ★	操作前请请务必再次确认该操作的必要性，同时要保证数据已完成备份	无

Manager 高危操作

表 14-3 Manager 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改OMS密码	该操作会重启OMS各进程，影响集群的管理维护	★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常
导入证书	该操作会重启OMS进程和整个集群，影响集群的管理维护和业务	★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
升级	该操作会重启Manager和整个集群，影响集群的管理维护和业务 分配集群管理权限的用户，需要严格管控，以防范可能的安全风险	★ ★ ★	修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
恢复OMS	该操作会重启Manager和整个集群，影响集群的管理维护和业务	★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改IP	该操作会重启Manager和整个集群，影响集群的管理维护和业务	★ ★ ★	修改时确保同一时间无其它管理维护操作，且修改的IP填写正确无误	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改日志级别	如果修改为DEBUG，会导致Manager运行速度明显降低	★ ★	修改前确认操作的必要性，并及时修改回默认设定	无
更换控制节点	该操作会导致部署在该节点上的服务中断，且当该节点同时为管理节点时，更换节点会导致重启OMS各进程，影响集群的管理维护	★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
更换管理节点	该操作会导致部署在该节点上的服务中断，会导致重启OMS各进程，影响集群的管理维护	★ ★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
重启下层服务时，如果勾选同时重启上层服务	该操作会导致上层服务业务中断，影响集群的管理维护和业务	★ ★ ★ ★	操作前确认操作的必要性，操作时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改OLDAP端口	修改该参数时，会重启LdapServer和Kerberos服务和其关联的所有服务，会影响业务运行	★ ★ ★ ★ ★	操作前确认操作的必要性，操作时确保同一时间无其它管理维护操作	无
用户删除supergroup组	删除supergroup组导致相关用户权限变小，影响业务访问	★ ★ ★ ★ ★	修改前确认需要添加的权限，确保用户绑定的supergroup权限删除前，相关权限已经添加，不会对业务造成影响	无
重启服务	重启过程中会中断服务，如果勾选同时重启上层服务会导致依赖该服务的上层服务中断	★ ★ ★	操作前确认重启的必要性	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改节点SSH默认端口	修改默认端口（22）将导致创建集群、添加服务/实例、添加主机、重装主机等功能无法正常使用，并且会导致集群健康检查结果中节点互信、omm/ommdba用户密码过期等检查项不准确	★ ★ ★	执行相关操作前将SSH端口改回默认值	无

ClickHouse 高危操作

表 14-4 ClickHouse 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除数据目录	该操作将会导致业务信息丢失	★ ★ ★	请勿手动删除数据目录	观察数据目录是否正常
缩容ClickHouseServer实例	该操作需要关注同分片中的ClickHouseServer实例节点需要同时退服缩容，否则会造成逻辑集群拓扑信息错乱；该操作执行前需检查逻辑集群内各节点的数据库和数据表信息，进行缩容预分析，保证缩容退服过程中数据迁移成功，避免数据丢失	★ ★ ★ ★ ★	进行缩容操作前，提前收集信息进行ClickHouse逻辑集群及实例节点状态判断	观察ClickHouse逻辑集群拓扑信息，各ClickHouseServer中数据库和数据表信息，以及数据量
扩容ClickHouseServer实例	该操作需要关注新扩容节点是否需要创建老节点上同名的数据库或数据表，否则会造成后续数据迁移、数据均衡以及缩容退服失败	★ ★ ★ ★ ★	进行扩容操作前，确认新扩容ClickHouseServer实例作用和目的，是否需要同步创建相关数据库和数据表	观察ClickHouse逻辑集群拓扑信息，各ClickHouseServer中数据库和数据表信息，以及数据量
退服ClickHouseServer实例	该操作需要关注同分片中的ClickHouseServer实例节点需要同时退服，否则会造成逻辑集群拓扑信息错乱；该操作执行前需检查逻辑集群内各节点的数据库和数据表信息，进行预分析，保证退服过程中数据迁移成功，避免数据丢失	★ ★ ★ ★ ★	进行退服操作前，提前收集信息进行Clickhouse逻辑集群及实例节点状态判断	观察ClickHouse逻辑集群拓扑信息，各ClickHouseServer中数据库和数据表信息，以及数据量
入服ClickHouseServer实例	该操作需要关注入服时必须选择原有分片中的所有节点入服，否则会造成逻辑集群拓扑信息错乱	★ ★ ★ ★ ★	进行入服操作前，对于待入服节点的分片归属信息需要确认	观察ClickHouse逻辑集群拓扑信息

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改数据目录下内容（创建文件、文件夹）	该操作将会导致该节点上的ClickHouse的实例故障	★ ★ ★	请勿手动在数据目录下创建或修改文件及文件夹	观察数据目录是否正常
单独启停基础组件	该操作将会影响服务的一些基础功能导致业务失败	★ ★ ★	请勿单独启停ZooKeeper/Kerberos/LDAP等基础组件，启停基础组件请勾选关联服务	观察服务状态是否正常
重启/停止服务	该操作将会导致业务中断	★ ★	确保在必要时重启/停止服务	观察服务是否运行正常

DBService 高危操作

表 14-5 DBService 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改DBService密码	修改密码需要重启服务，服务在重启过程中无法访问。	★ ★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作。	观察是否有未恢复的告警产生，观察集群的管理维护是否正常
恢复DBService数据	数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。	★ ★ ★ ★	恢复前确认操作的必要性，恢复时确保同一时间无其它管理维护操作。	观察是否有未恢复的告警产生，观察集群的管理维护是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
DBService主备倒换	倒换DBServer过程中，DBService无法提供服务。	★ ★	操作前确认该操作的必要性，操作时确保同一时间无其它管理维护操作。	无
修改DBService浮动IP配置	需要重启DBService服务使配置生效，服务在重启无法访问。 如果浮动IP已被使用过，将会导致配置失败，DBService启动失败。	★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效。	观察服务能否正常启动

Flink 高危操作

表 14-6 Flink 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改日志级别	如果修改为DEBUG，会影响任务运行性能	★ ★	修改前确认操作的必要性，并及时修改回默认设定	无
修改文件权限	该操作可能导致任务运行失败	★ ★ ★	修改前确认操作的必要性	观察相关业务操作是否正常

Flume 高危操作

表 14-7 Flume 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Flume实例的启动参数GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
修改HDFS的副本数目dfs.replication，将默认值由3改为1	导致： 1. 存储可靠性下降，磁盘故障时，会发生数据丢失 2. NameNode重启失败，HDFS服务不可用	★ ★ ★ ★	修改相关配置项时，请仔细查看参数说明。保证数据存储的副本数不低于2	观察默认的副本值是否不为1，HDFS服务是否可以正常提供服务

HBase 高危操作

表 14-8 HBase 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改加密的相关配置项 ● hbase.regionserver.wal.encryption ● hbase.crypto.keyprovider.parameters.uri ● hbase.crypto.keyprovider.parameters.encryptedtext	导致服务启动异常	★ ★ ★ ★	修改相关配置项时请严格按照提示描述，加密相关配置项是有关联的，确保修改后的值有效	观察服务能否正常启动
已使用加密的情况下关闭或者切换加密算法，关闭主要指修改hbase.regionserver.wal.encryption为false，切换主要指AES和SMS4的切换	导致服务启动失败，数据丢失	★ ★ ★ ★	加密HFile和WAL内容的时候，如果已经使用一种加密算法加密并且已经建表，请不要随意关闭或者切换加密算法。 未建加密表（ENCRYPTION=>AES/SMS4）的情况下可以切换，否则禁止操作	无

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改HBase实例的启动参数GC_OPTS、HBASE_HEAPSIZE	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效，且GC_OPTS与HBASE_HEAPSIZE参数值无冲突	观察服务能否正常启动
使用OfflineMetaRepair工具	导致服务启动异常	★ ★ ★ ★	必须在HBase下线的情况下才可以使用该命令，而且不能在数据迁移的场景中使用该命令	观察HBase服务是否可以正常启动。

HDFS 高危操作

表 14-9 HDFS 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改HDFS的NameNode的数据存储目录dfs.namenode.name.dir、DataNode的数据配置目录dfs.datanode.data.dir	导致服务启动异常	★ ★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
执行**hadoop distcp**命令时，使用**-delete**参数	Distcp拷贝时，源集群没有而目的集群存在的文件，会在目的集群删除	★ ★	在使用Distcp的时候，确保是否保留目的集群多余的文件，谨慎使用**-delete**参数	Distcp数据拷贝后，查看目的的数据是否按照参数配置保留或删除
修改HDFS实例的启动参数GC_OPTS、HADOOP_HEAPSIZE和GC_PROFILE	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效，且GC_OPTS与HADOOP_HEAPSIZE参数值无冲突	观察服务能否正常启动
修改HDFS的副本数目dfs.replication，将默认值由3改为1	导致： 1. 存储可靠性下降，磁盘故障时，会发生数据丢失 2. NameNode重启失败，HDFS服务不可用	★ ★ ★ ★	修改相关配置项时，请仔细查看参数说明。保证数据存储的副本数不低于2	观察默认的副本值是否不为1，HDFS服务是否可以正常提供服务
修改Hadoop中各模块的RPC通道的加密方式hadoop.rpc.protection	导致服务故障及业务异常	★ ★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察HDFS及其他依赖HDFS的服务能否正常启动，并提供服务

Hive 高危操作

表 14-10 Hive 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Hive实例的启动参数GC_OPTS	修改该参数可能会导致Hive实例无法启动	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
删除MetaStore所有实例	Hive元数据丢失，Hive无法提供服务	★ ★ ★	除非确定丢弃Hive所有表信息，否则不要执行该操作	观察服务能否正常启动
使用HDFS文件系统接口或者HBase接口删除或修改Hive表对应的文件	该操作会导致Hive业务数据丢失或被篡改	★ ★	除非确定丢弃这些数据，或者确保该修改操作符合业务需求，否则不要执行该操作	观察Hive数据是否完整
使用HDFS文件系统接口或者HBase接口修改Hive表对应的文件或目录访问权限	该操作可能会导致相关业务场景不可用	★ ★ ★	请勿执行该操作	观察相关业务操作是否正常
使用HDFS文件系统接口删除或修改文件hdfs:///apps/templeton/hive-3.1.0.tar.gz	该操作可能会导致WebHCat无法正常执行业务	★ ★	请勿执行该操作	观察相关业务操作是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
导出表数据覆盖写入本地目录，例如将**t1**表中数据导出，覆盖到“/opt/dir”路径下： **insert overwrite local directory '/opt/dir' select * from t1;**	该操作会删除目标目录，如果设置错误，会导致软件或者操作系统无法启动	★ ★ ★ ★ ★	确认需要写入的路径下不要包含任何文件；或者不要使用overwrite关键字	观察目标路径是否有文件丢失
将不同的数据库、表或分区文件指定至相同路径，例如默认仓库路径“/user/hive/warehouse”。	执行创建操作后数据可能会紊乱，如果删除其中一个数据库、表或分区，会导致其他对象数据丢失	★ ★ ★ ★ ★	请勿执行该操作	观察目标路径是否有文件丢失

Kafka 高危操作

表 14-11 Kafka 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除Topic	该操作将会删除已有的主题和数据	★ ★ ★	采用Kerberos认证，保证合法用户具有操作权限，并确保主题名称正确	观察主题是否正常处理

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除数据目录	该操作将会导致业务信息丢失	★ ★ ★	请勿手动删除数据目录	观察数据目录是否正常
修改数据目录下内容（创建文件、文件夹）	该操作将会导致该节点上的Broker实例故障	★ ★ ★	请勿手动在数据目录下创建或修改文件及文件夹	观察数据目录是否正常
修改磁盘自适应功能“disk.adapter.enable”参数	该操作会在磁盘使用空间达到阈值时调整Topic数据保存周期，超出保存周期的历史数据可能被清除	★ ★ ★	若个别Topic不能做保存周期调整，将该Topic配置在“disk.adapter.topic.blacklist”参数中	在KafkaTopic监控页面观察数据的存储周期
修改数据目录“log.dirs”配置	该配置不正确将会导致进程故障	★ ★ ★	确保所修改或者添加的数据目录为空目录，且权限正确	观察数据目录是否正常
减容Kafka集群	该操作将会导致部分Topic数据副本数量减少，可能会导致Topic无法访问	★ ★	请先做好数据副本转移工作，然后再进行减容操作	观察分区所在备份节点是否都存活，确保数据安全
单独启停基础组件	该操作将会影响服务的一些基础功能导致业务失败	★ ★ ★	请勿单独启停ZooKeeper/Kerberos/LDAP等基础组件，启停基础组件请勾选关联服务	观察服务状态是否正常
重启/停止服务	该操作将会导致业务中断	★ ★	确保在必要时重启/停止服务	观察服务是否运行正常
修改配置参数	该操作将需要重启服务使得配置生效	★ ★	确保在必要时修改配置	观察服务是否运行正常
删除/修改元数据	修改或者删除ZooKeeper上Kafka的元数据可能导致Topic或者Kafka服务不可用	★ ★ ★	请勿删除或者修改Kafka在ZooKeeper上保存的元数据信息	观察Topic或者Kafka服务是否可用

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改元数据备份文件	修改Kafka元数据备份文件，并被使用进行Kafka元数据恢复成功后，可能导致Topic或者Kafka服务不可用	★ ★ ★	请勿修改Kafka元数据备份文件	观察Topic或者Kafka服务是否可用

KrbServer 高危操作

表 14-12 KrbServer 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改KrbServer的参数KADMIN_PORT	修改该参数后，若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
修改KrbServer的参数kdc_ports	修改该参数后，若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
修改KrbServer的参数KPASSWD_PORT	修改该参数后，若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
修改Manager系统域名	若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
配置跨集群互信	该操作会重启KrbServer服务和其关联的所有服务，影响集群的管理维护和业务	★ ★ ★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常

LdapServer 高危操作

表 14-13 LdapServer 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改LdapServer的参数LDAP_SERVER_PORT	修改该参数后，若没有及时重启LdapServer服务和其关联的所有服务，会导致集群内部LdapClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启LdapServer服务和其关联的所有服务	无
恢复LdapServer数据	该操作会重启Manager和整个集群，影响集群的管理维护和业务	★ ★ ★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
更换LdapServer所在节点	该操作会导致部署在该节点上的服务中断，且当该节点为管理节点时，更换节点会导致重启OMS各进程，影响集群的管理维护	★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改LdapServer密码	修改密码需要重启LdapServer和Kerberos服务，影响集群的管理维护和业务	★ ★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	无
节点重启导致LdapServer数据损坏	如果未停止LdapServer服务，直接重启LdapServer所在节点，可能导致LdapServer数据损坏	★ ★ ★ ★ ★	使用LdapServer备份数据进行恢复	无

Loader 高危操作

表 14-14 Loader 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Loader实例的浮动IP地址loader.float.ip	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察Loader UI是否可以正常连接
修改Loader实例的启动参数LOADER_GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
往HBase导入数据时，选择清空表数据	目标表的原数据被清空	★ ★	选择时，确保目标表的数据可以清空	选择前，需确认目标表数据是否可以清空

Spark2x 高危操作

说明

MRS 3.x之前版本，服务名称为为Spark。

表 14-15 Spark2x 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
配置项的修改（spark.yarn.queue）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
配置项的修改（spark.driver.extraJavaOptions）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
配置项的修改（spark.yarn.cluster.driver.extraJavaOptions）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
配置项的修改（spark.eventLog.dir）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
配置项的修改（SPARK_DAEMON_JAVA_OPTS）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
删除所有JobHistory2x实例	导致历史应用的event log丢失	★ ★	至少保留一个JobHistory2x实例	观察JobHistory2x中是否可以查看历史应用信息
删除或修改HDFS上的/user/spark2x/jars/8.0.0/spark-archive-2x.zip	导致JDBCServer2x启动异常及业务功能异常	★ ★ ★	删除/user/spark2x/jars/8.0.0/spark-archive-2x.zip，等待10-15分钟，zip包自动恢复	观察服务能否正常启动

Storm 高危操作

表 14-16 Storm 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改插件相关的配置项： ● storm.scheduler ● nimbus.authorizer ● storm.thrift.transport ● nimbus.blobstore.class ● nimbus.topology.validator ● storm.principal.tolocal	导致服务启动异常	★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的类名是存在并有效的	观察服务能否正常启动

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Storm实例的启动参数 GC_OPTS NIMBUS_GC_OPTS SUPERVISOR_GC_OPTS UI_GC_OPTS LOGVIEWER_GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
修改用户资源池配置参数 resource.aware.scheduler.user.pools	导致业务提交后无法正常运行	★ ★ ★	修改相关配置项时请严格按照提示描述，确保给每个用户分配的资源合理有效	观察服务能否正常启动并且业务能否正常运行
修改数据目录	该操作不当会导致服务异常，无法提供服务	★ ★ ★ ★	请勿手动操作数据目录	观察数据目录是否正常
重启服务/实例	该操作会导致服务有短暂中断，如果有业务运行也会引起业务短暂中断	★ ★ ★	确保在必要时重启服务	观察服务是否运行正常，业务是否恢复
同步配置（重启服务）	该操作会引起服务重启，导致服务短暂中断，若引起Supervisor重启会导致所运行业务短暂中断	★ ★ ★	确保在必要时修改配置	观察服务是否运行正常，业务是否恢复
停止服务/实例	该操作会导致服务停止，业务中断	★ ★ ★	确保在必要时停止服务	观察服务是否正常停止
删除/修改元数据	删除Nimbus元数据会导致服务异常，并且已运行业务丢失	★ ★ ★ ★ ★	请勿手动删除Nimbus元数据文件	观察Nimbus元数据文件是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改文件权限	修改元数据目录和日志目录权限不当会引起服务异常	★★★★	请勿手动修改文件权限	观察数据目录和日志目录权限是否正常
删除拓扑	该操作会删除正在运行中的拓扑	★★★★	确保在必要时删除拓扑	观察拓扑是否删除成功

Yarn 高危操作

表 14-17 Yarn 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除或者修改数据目录 yarn.nodemanager.local-dirs 和 yarn.nodemanager.log-dirs	该操作将会导致业务信息丢失	★★★	请勿手动删除数据目录	观察数据目录是否正常

ZooKeeper 高危操作

表 14-18 ZooKeeper 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除或者修改ZooKeeper的数据目录	该操作将会导致业务信息丢失	★ ★ ★	修改ZooKeeper目录时候，严格按照扩容指导操作	观察服务能否正常启动，关联组件能否正常启动
修改ZooKeeper实例的启动参数GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
设置ZooKeeper中znode的ACL信息	修改ZooKeeper中znode的权限，可能会导致其他用户无权限访问该znode，导致系统功能异常	★ ★ ★ ★	修改相关配置项时请严格按照“ZooKeeper权限设置指南”章节操作，确保修改ACL信息，不会影响其他组件正常使用ZooKeeper	观察项目观察其他依赖ZooKeeper的组件能否正常启动，并提供服务

15 常见问题

15.1 产品咨询类

15.1.1 MRS 可以做什么?

MapReduce服务（MapReduce Service）为客户提供ClickHouse、Spark、Flink、Kafka、HBase等Hadoop生态的高性能大数据引擎，支持数据湖、数据仓库、BI、AI融合等能力，完全兼容开源，快速帮助客户上云构建低成本、灵活开放、安全可靠、全栈式的云原生大数据平台，满足客户业务快速增长和敏捷创新诉求。

15.1.2 MRS 支持什么类型的分布式存储?

提供目前主流的Hadoop，目前支持Hadoop 3.1.x版本，并且随社区更新版本。MRS支持的组件版本请参考**表15-1**。

表 15-1 MRS 组件版本信息

MRS支持的组件	MRS 1.9.2（适用于MRS 1.9.x）	MRS 3.1.0
Alluxio	2.0.1	-
CarbonData	1.6.1	2.0.1
DBService	1.0.0	2.7.0
Flink	1.7.0	1.12.0
Flume	1.6.0	1.9.0
HBase	1.3.1	2.2.3
HDFS	2.8.3	3.1.1
Hive	2.3.3	3.1.0
Hudi	-	0.7.0

MRS支持的组件	MRS 1.9.2（适用于MRS 1.9.x）	MRS 3.1.0
Hue	3.11.0	4.7.0
Impala	-	3.4.0
Kafka	1.1.0	2.11-2.4.0
KafkaManager	1.3.3.1	-
KrbServer	1.15.2	1.17
Kudu	-	1.12.1
LdapServer	1.0.0	2.7.0
Loader	2.0.0	-
MapReduce	2.8.3	3.1.1
Oozie	-	5.1.0
Opentsdb	2.3.0	-
Presto	0.216	333
Phoenix（集成在HBase中）	-	5.0.0
Ranger	1.0.1	2.0.0
Spark	2.2.2	-
Spark2x	-	2.4.5
Sqoop	-	1.4.7
Storm	1.2.1	-
Tez	0.9.1	0.9.2
YARN	2.8.3	3.1.1
ZooKeeper	3.5.1	3.5.6
MRS Manager	1.9.2	-
FusionInsight Manager	-	8.1.0

15.1.3 如何使用自定义安全组创建 MRS 集群?

用户购买集群时，如果选择使用自己创建的安全组，则需要放开9022端口，或者在界面上购买集群时，安全组选择"自动创建"。

15.1.4 如何使用 MRS?

MRS是一个在云上部署和管理Hadoop系统的服务，一键即可部署Hadoop集群。MRS提供租户完全可控的企业级大数据集群云服务，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件。

MRS使用简单，通过使用在集群中连接在一起的多台计算机，您可以运行各种任务，处理或者存储（PB级）巨量数据。MRS的基本使用流程如下：

1. 上传程序和数据文件到对象存储服务（OBS）中，用户需要先将本地的程序和数据文件上传至OBS中。
2. 创建集群，用户可以指定集群类型用于离线数据分析和流处理任务，指定集群中预置的弹性云服务器实例规格、实例数量、数据盘类型（普通IO、高 IO、超高IO）、要安装的组件（Hadoop、Spark、HBase、Hive、Kafka、Storm等）。用户可以使用引导操作在集群启动前（或后）在指定的节点上执行脚本，安装其他第三方软件或修改集群运行环境等自定义操作。
3. 管理作业，MRS为用户提供程序执行平台，程序由用户自身开发，MRS负责程序的提交、执行和监控。
4. 管理集群，MRS为用户提供企业级的大数据集群的统一管理平台，帮助用户快速掌握服务及主机的健康状态，通过图形化的指标监控及定制及时的获取系统的关键信息，根据实际业务的性能需求修改服务属性的配置，对集群、服务、角色实例等实现一键启停等操作。
5. 删除集群，如果作业执行结束后不需要集群，可以删除MRS集群。

15.1.5 如何保证数据和业务运行安全?

MRS作为一个海量数据管理和分析平台，具备高安全性。主要从以下几个方面保障数据和业务运行安全：

- 网络隔离

  整个公有云网络划分为2个平面，即业务平面和管理平面。两个平面采用物理隔离的方式进行部署，保证业务、管理各自网络的安全性。
  - 业务平面：主要是集群组件运行的网络平面，支持为用户提供业务通道，对外提供数据存取、任务提交及计算能力。
  - 管理平面：主要是公有云管理控制台，用于购买和管理MRS。
- 主机安全

  用户可以根据自己业务的需要部署第三方的防病毒软件。针对操作系统和端口部分，MRS提供如下安全措施：
  - 操作系统内核安全加固
  - 更新操作系统最新补丁
  - 操作系统权限控制
  - 操作系统端口管理
  - 操作系统协议与端口防攻击
- 数据安全

  MRS支持数据存储在OBS上，保障客户数据安全。
- 数据完整性

  MRS处理完数据后，通过SSL加密传输数据至OBS，保证客户数据的完整性。

15.1.6 如何配置 Phoenix 连接池?

Phoenix不支持连接池设置，建议用户自己写代码实现一个管理连接的工具类，模拟连接池。

15.1.7 MRS 是否支持更换网段?

MRS支持更换网段，请在集群详情页“默认生效子网”右侧单击“切换子网”，选择当前集群所在VPC下的其他子网，实现可用子网IP的扩充。新增子网不会影响当前已有节点的IP地址和子网。

15.1.8 MRS 服务集群节点是否执行降配操作?

MRS服务暂不支持降级配置操作，如果有诉求，建议客户联系技术支持处理。

15.1.9 Hive 与其他组件有什么关系?

- Hive与HDFS间的关系

  Hive是Apache的Hadoop项目的子项目，Hive利用HDFS作为其文件存储系统。Hive通过解析和计算处理结构化的数据，Hadoop HDFS则为Hive提供了高可靠性的底层存储支持。Hive数据库中的所有数据文件都可以存储在Hadoop HDFS文件系统上，Hive所有的数据操作也都是通过Hadoop HDFS接口进行。

- Hive与MapReduce间的关系

  Hive所有的数据计算都依赖于MapReduce。MapReduce也是Apache的Hadoop项目的子项目，它是一个基于Hadoop HDFS分布式并行计算框架。Hive进行数据分析时，会将用户提交的HiveQL语句解析成相应的MapReduce任务并提交MapReduce执行。

- Hive与DBService间的关系

  Hive的MetaStore（元数据服务）处理Hive的数据库、表、分区等的结构和属性信息，这些信息需要存放在一个关系型数据库中，由MetaStore维护和处理。在MRS中，这个关系型数据库由DBService组件维护。

- Hive与Spark间的关系

  Hive的数据计算也可以运行在Spark上。Spark也是Apache的一个项目，它是基于内存的分布式计算框架。Hive进行数据分析时，会将用户提交的HiveQL语句解析成相应的Spark任务并提交Spark执行。

15.1.10 MRS 集群是否支持 Hive on Spark?

- MRS 1.9.x版本集群支持Hive on Spark。
- MRS 3.x及之后版本的集群支持Hive on Spark。
- 其他版本可使用Hive on Tez替代。

15.1.11 Hive 版本之间是否兼容?

Hive 3.1版本与Hive 1.2版本相比不兼容内容如下：

- 字段类型约束：Hive 3.1不支持String转成int
- UDF不兼容：Hive 3.1版本UDF内的Date类型改为Hive内置
- 索引功能废弃

- 时间函数问题：Hive 3.1版本为UTC时间，Hive 1.2版本为当地时区时间
- 驱动不兼容：Hive 3.1和Hive 1.2版本的jdbc驱动不兼容
- Hive 3.1对ORC文件列名大小写，下划线敏感
- Hive 3.1版本列中不能有名为time的列

15.1.12 MRS 集群哪个版本支持建立 Hive 连接且有用户同步功能?

MRS 2.0.5以上版本支持DGC建立hive连接且有IAM用户同步功能。

15.1.13 数据存储在 OBS 和 HDFS 有什么区别?

MRS集群处理的数据源来源于OBS或HDFS，HDFS是Hadoop分布式文件系统（Hadoop Distributed File System），OBS（Object Storage Service）即对象存储服务，是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于OBS服务 Web界面和OBS客户端对数据进行浏览、管理和使用，同时可以通过REST API接口方式单独或集成到业务程序进行管理和访问数据。

- 数据存储在OBS：数据存储和计算分离，集群存储成本低，存储量不受限制，并且集群可以随时删除，但计算性能取决于OBS访问性能，相对HDFS有所下降，建议在数据计算不频繁场景下使用。
- 数据存储在HDFS：数据存储和计算不分离，集群成本较高，计算性能高，但存储量受磁盘空间限制，删除集群前需将数据导出保存，建议在数据计算频繁场景下使用。

15.1.14 Hadoop 压力测试工具如何获取?

请从如下URL中下载：**https://github.com/Intel-bigdata/HiBench**

15.1.15 Impala 与其他组件有什么关系?

- Impala与HDFS间的关系

  Impala默认利用HDFS作为其文件存储系统。Impala通过解析和计算处理结构化的数据，Hadoop HDFS则为Impala提供了高可靠性的底层存储支持。使用Impala将无需移动HDFS中的数据并且提供更快的访问。

- Impala与Hive间的关系

  Impala使用Hive的元数据、ODBC驱动程序和SQL语法。与Hive不同，Impala不基于MapReduce算法，它实现了一个基于守护进程的分布式架构，它负责在同一台机器上运行的查询执行的所有方面。因此，它减少了使用MapReduce的延迟，这使Impala比Hive快。

- Impala与MapReduce间的关系

  无

- Impala与Spark间的关系

  无

- Impala与Kudu间的关系

  Kudu与Impala紧密集成，替代Impala+HDFS+Parquet组合。允许使用Impala的SQL语法从Kudu tablets插入、查询、更新和删除数据。此外，还可以用 JDBC或ODBC，Impala作为代理连接Kudu进行数据操作。

- Impala与HBase间的关系

  默认的Impala表使用存储在HDFS上的数据文件，这对于使用全表扫描的批量加载和查询是理想的。但是，HBase可以提供对OLTP样式组织的数据的便捷高效查询。

15.1.16 关于 MRS 服务集成的开源第三方 SDK 中包含的公网 IP 地址声明

MRS服务集成的开源组件所依赖的开源三方包中包含SDK使用示例，其中涉及“12.1.2.3”、“54.123.4.56”、“203.0.113.0”、“203.0.113.12”等公网IP均为示例IP，MRS服务进程不会主动发起与该公网IP的连接，也不会与该公网IP进行任何数据交换。

15.1.17 Kudu 和 HBase 间的关系?

Kudu的设计有参考HBase的结构，也能够实现HBase擅长的快速的随机读写、更新功能。二者的结构差别不大，主要差别在于：

- Kudu不依赖Zookeeper，通过自身实现Raft来保证一致性。
- Kudu持久化数据不依赖HDFS，TServer实现数据的强一致性和可靠性。

15.1.18 MRS 是否支持 Hive on Kudu?

MRS不支持Hive on Kudu。

目前MRS只支持两种方式访问kudu：

- 通过impala表访问kudu。
- 通过客户端应用程序访问操作kudu表。

15.1.19 10 亿级数据量场景的解决方案

- 有数据更新、联机事务处理OLTP、复杂分析的场景，建议使用云数据库GaussDB(for MySQL)。
- MRS的Impala + Kudu也能满足该场景，Impala + Kudu可以在join操作时，把当前所有的join表都加载到内存中来实现。

15.1.20 如何修改 DBService 的 IP?

MRS不支持修改DBService的IP。

15.1.21 MRS sudo log 能否清理?

MRS sudo log文件是omm用户的操作记录，是为了方便问题的定位 ，可以清理。因为日志占用了一部分存储空间，建议客户可以清除比较久远的操作日志释放资源空间。

1. 日志文件较大，可以将此文件目录添加到/etc/logrotate.d/syslog中，让系统做日志老化 ，定时清理久远的日志 。

   方法：更改文件日志目录：sed -i '3 a/var/log/sudo/sudo.log' /etc/logrotate.d/syslog
2. 可以根据日志个数和大小进行设置/etc/logrotate.d/syslog，超过设置的日志会自动删除掉。一般默认按照存档大小和个数进行老化的，可以通过size和rotate分别

是日志大小限制和个数限制，默认没有时间周期的限制，如需进行周期设置可以增加daily/weekly/monthly指定清理日志的周期为每天/每周/每月。

15.1.22 MRS 2.1.0 集群版本对 Storm 日志也有 20G 的限制么

MRS 2.1.0集群版本对Storm日志仍然有20G的大小限制，超出后会循环删除。因日志是保存在系统盘上，还是有空间限制的。若如需长期保存，日志需要挂载出来，方便长期保存。

15.1.23 Spark ThriftServer 是什么

ThriftServer是一个JDBC接口，用户可以通过JDBC连接ThriftServer来访问SparkSQL的数据。因此Spark组件中可以看到JDBCServer进程，不会看到ThriftServer。

15.1.24 Kafka 目前支持的访问协议类型

当前支持四种协议类型的访问：PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL

15.1.25 zstd 的压缩比怎么样

zstd的压缩比orc好一倍，是开源的，具体请参见**https://github.com/L-Angel/compress-demo**，CarbonData不支持lzo，MRS里面有集成zstd。

15.1.26 创建 MRS 集群时，找不到 HDFS、Yarn、MapReduce 组件

HDFS、Yarn和MapReduce组件包含在Hadoop组件中，当创建MRS集群时无法看到HDFS、Yarn和MapReduce组件，勾选Hadoop组件并等待集群创建完成后即可在“组件管理”页签看到HDFS、Yarn和MapReduce组件。

15.1.27 创建 MRS 集群时，找不到 ZooKeeper 组件

创建MRS 3.x之前版本集群时，ZooKeeper组件为默认安装的组件，不在创建集群的界面上显示。

创建MRS 3.x及之后版本集群时，可以在创建集群的界面看到ZooKeeper组件，并默认勾选。

集群创建完成后可在集群“组件管理”页签看到ZooKeeper组件。

15.1.28 MRS 3.1.0 集群版本，Spark 任务支持 python 哪些版本?

MRS 3.1.0集群版本，Spark任务支持的python建议使用2.7或3.x版本。

15.1.29 如何让不同的业务程序分别用不同的 Yarn 队列?

在Manager页面上创建一个新的租户即可。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择父租户节点然后单击，打开添加子租户的配置页面，参见**表15-2**为子租户配置属性。

表 15-2 子租户参数一览

参数名	描述
集群	显示上级父租户所在集群。
父租户资源	显示上级父租户的名称。
名称	● 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（ _ ）。 ● 根据业务需求规划子租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
租户类型	指定租户是否是一个叶子租户： ● 选择“叶子租户”：当前租户为叶子租户，不支持添加子租户。 ● 选择“非叶子租户”：当前租户为非叶子租户，支持添加子租户，但租户层级不能超过5层。
计算资源	为当前租户选择动态计算资源。 ● 选择“Yarn”时，系统自动在Yarn中以子租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 ● 不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量 （%）	配置当前租户使用的计算资源百分比，基数为父租户的资源总量。
默认资源池最大容量（%）	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。

参数名	描述
存储空间配额	配置当前租户使用的HDFS存储空间配额。 ● 当存储空间配额单位设置为MB时，范围为1～8796093022208，当“存储空间配额单位”设置为GB时，范围为1～8589934592。 ● 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 ● 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 ● 如果此配额大于父租户的配额，实际存储量不超过父租户配额。
存储路径	配置租户在HDFS中的存储目录。 ● 系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“/tenant/ta1”，系统默认自动配置此参数值为“/tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。 ● 支持在父目录中自定义存储路径。
描述	配置当前租户的描述信息

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。
- 子租户可以将当前租户的资源进一步分配。每一级别父租户下，直接子租户的资源百分比之和不能超过100%。所有一级租户的计算资源百分比之和也不能超过100%。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：
   - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
   - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

15.1.30 MRS 管理控制台和集群 Manager 页面区别与联系

用户可以通过MRS管理控制台页面登录到MRS的Manager页面。

Manager分为MRS Manager和FusionInsight Manager，其中：

- MRS 2.x及之前版本集群的Manager界面称为MRS Manager。
- MRS 3.x及之后版本集群的Manager界面称为FusionInsight Manager。

管理控制台与FusionInsight Manager页面的区别和联系请参考下表：

常用操作	MRS Console	FusionInsight Manager
切换子网、添加安全组规则、OBS权限控制、管理委托、IAM用户同步	支持	不支持
新增节点组、扩容、缩容、升级规格	支持	不支持
隔离主机、启动所有角色、停止所有角色	支持	支持
下载客户端、启动服务、停止服务、滚动重启服务	支持	支持
查看服务实例状态、参数配置、同步配置	支持	支持
查看清除告警、查看事件	支持	支持
查看告警帮助	不支持	支持
阈值设置	不支持	支持
添加消息订阅规格	支持	不支持
文件管理	支持	不支持
作业管理	支持	不支持
租户管理	支持	支持

常用操作	MRS Console	FusionInsight Manager
标签管理	支持	不支持
权限（添加删除用户、用户组、角色、修改密码）	不支持	支持
备份恢复	不支持	支持
审计	不支持	支持
资源监控、日志	支持	支持

15.1.31 MRS 如何解绑 EIP?

问题现象

控制台页面绑定了EIP后无法在VPC服务EIP模块进行解绑。

弹框提示公网IP被MapReduce服务使用，不能执行该操作。

操作步骤

步骤1 单击导航栏我的虚拟私有云，找到对应vpc。

步骤2 单击进入子网页面，找到对应集群所属子网。

步骤3 找到对应的公网IP，单击后面的解绑弹性公网IP按钮即可解绑。

----结束

15.2 帐号密码类

15.2.1 登录 Manager 帐号的是什么?

系统默认登录**Manager**的帐号为admin，密码为创建集群时用户自己设置的密码。

15.2.2 帐号密码的过期时间如何查询和修改

查询密码有效期

查询组件运行用户（人机用户、机机用户）密码有效期：

步骤1 以客户端安装用户，登录安装了客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，输入kadmin/admin用户密码后进入kadmin控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“Admin@123”，首次登录后需修改密码，请按照提示修改并妥善保存。

步骤5 执行如下命令，可以查看用户的信息。

getprinc *系统内部用户名*

例如：**getprinc user1**

```
kadmin:  getprinc user1
......
Expiration date: [never]
Last password change: Sun Oct 09 15:29:54 CST 2022
Password expiration date: [never]
......
```

----结束

查询操作系统用户密码有效期：

步骤1 以**root**用户登录集群任一Master节点。

步骤2 执行以下命令查看用户密码有效期（“Password expires”参数值）。

chage -l *用户名*

例如查看**root**用户密码有效期，则执行**chage -l root**，执行后结果如下：

```
[root@xxx ~]#chage -l root
Last password change                                    : Sep 12, 2021
Password expires                                        : never
Password inactive                                       : never
Account expires                                         : never
Minimum number of days between password change          : 0
Maximum number of days between password change          : 99999
Number of days of warning before password expires       : 7
```

----结束

修改密码有效期

- “机机”用户密码随机生成，密码默认永不过期。
- “人机”用户密码的有效期可以在Manager页面通过修改密码策略进行修改。

15.3 帐号权限类

15.3.1 如果不开启 Kerberos 认证，MRS 集群能否支持访问权限细分?

MRS 2.1.0及之前版本：在MRS Manager页面选择“系统设置”>“配置”>“权限配置”查询。

MRS 3.x及之后版本：在FusionInsight Manager页面选择“系统 > 权限”查询。

15.3.2 如何给新建的帐号添加租户管理权限?

分析集群和混合集群支持添加租户管理权限，流式集群不支持添加租户管理权限。给新建帐号添加租户管理权限方法如下：

MRS 3.x之前版本：

步骤1 用admin帐号登录MRS Manager.

步骤2 在“系统设置 > 用户管理”中选择新建的帐号，单击“操作”列中的“修改”.

步骤3 在“分配角色权限”中单击“选择并绑定角色”。

- 绑定Manager_tenant角色，则该帐号拥有租户管理的查看权限。
- 绑定Manager_administrator角色，则该帐号拥有租户管理的查看和操作权限。

步骤4 单击“确定”完成修改。

----结束

MRS 3.x及之后版本：

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 用户”。

步骤2 在要修改信息的用户所在行，单击“修改”。

根据实际情况，修改对应参数。

绑定Manager_tenant角色，则该帐号拥有租户管理的查看权限。绑定Manager_administrator角色，则该帐号拥有租户管理的查看和操作权限。

说明

修改用户的用户组，或者修改用户的角色权限，最长可能需要3分钟时间生效。

步骤3 单击“确定”完成修改操作。

----结束

15.3.3 如何自定义配置 MRS 服务策略?

1. 在IAM控制台，单击左侧导航栏的“权限”，在右上角选择“创建自定义策略”。
2. 策略名称：自定义策略的名称。
3. 作用范围：根据服务的属性填写，MRS为项目级服务，选择“项目级服务”。
4. 策略配置方式。
   - 可视化视图：通过可视化视图创建自定义策略，无需了解JSON语法，按可视化视图导航栏选择云服务、操作、资源、条件等策略内容，可自动生成策略。

- JSON视图：通过JSON视图创建自定义策略，可以在选择策略模板后，根据具体需求编辑策略内容；也可以直接在编辑框内编写JSON格式的策略内容。

  也可以从“策略内容”区域，单击“从已有策略复制”选择已有策略作为模板进行修改。

5. 输入“策略描述”（可选）。
6. 单击“确定”，自定义策略创建完成。
7. 将新创建的自定义策略授予用户组，使得用户组中的用户具备自定义策略中的权限。

15.3.4 在 MRS Manager 页面“系统设置”中找不到用户管理，什么原因?

该用户没有Manager_administrator角色权限，所以在MRS Manager页面“系统设置”中不显示“用户管理”。

15.3.5 Hue 有没有配置帐号权限的功能?

Hue服务没有配置帐号权限的功能，可以通过在Manager的“系统设置”中配置用户角色和用户组来配置帐号权限，从而实现Hue权限的配置。

15.4 客户端使用类

15.4.1 如何使用组件客户端?

1. 以**root**用户登录任意一个Master节点。
2. 执行**su - omm**命令，切换到omm用户。
3. 执行**cd /opt/client**命令，切换到客户端。
4. 执行**source bigdata_env**命令，配置环境变量。

   如果当前集群已启用Kerberos认证，执行**kinit** *组件业务用户*认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。
5. 环境变量配置成功后，即可执行组件的客户端命令。例如查看组件的相关信息，可执行HDFS客户端命令**hdfs dfs -ls /**查看HDFS根目录文件。

15.4.2 怎么关闭 ZooKeeper SASL 认证

登录FusionInsight Manager，选择“集群 > 服务 > ZooKeeper > 配置 > 全部配置”，在左侧导航栏选择“quorumpeer > 自定义”添加参数名称和值：zookeeper.sasl.disable = false。保存配置后，重启ZooKeeper服务。

15.4.3 在 MRS 集群外客户端中执行 kinit 报错

问题现象

在MRS集群外节点上安装了客户端后并执行**kinit**命令报错如下：

```
-bash kinit Permission denied
```

执行**java**命令也报错如下：

```
-bash: /xxx/java: Permission denied
```

执行**ll /*java安装路径*/JDK/jdk/bin/java**命令查看该文件执行权限信息正常。

原因分析

执行**mount | column -t**查看挂接的分区状态，发现java执行文件所在的挂载点的分区状态是“noexec”。当前环境中将安装MRS客户端所在的数据盘配置成“noexec”，即禁止二进制文件执行，从而无法使用java命令。

解决方法

1. 以**root**用户登录MRS客户端所在节点。
2. 移除“/etc/fstab”文件中MRS客户端所在的数据盘的配置项“noexec”。
3. 执行**umount**命令卸载数据盘，然后再执行**mount -a**重新挂载数据盘。

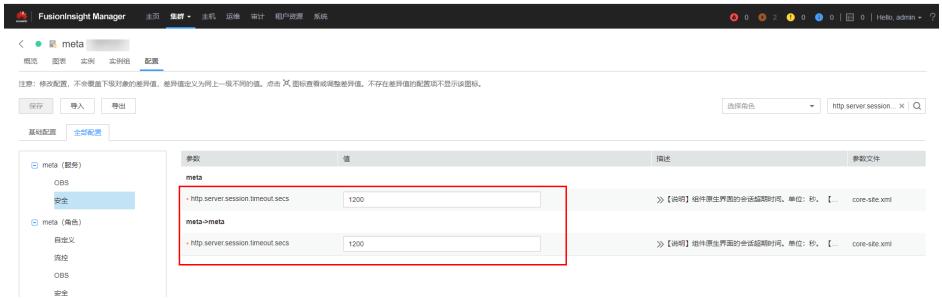
15.5 Web 页面访问类

15.5.1 修改开源组件 Web 页面会话超时时间

请合理设置Web页面超时时间，避免由于Web页面长时间暴露造成的信息泄露。

确定集群是否支持调整会话时长

- MRS 3.x之前版本集群：在集群详情页面，选择“组件管理 > meta > 服务配置”，切换“基础配置”为“全部配置”，搜索配置项“http.server.session.timeout.secs”，如果有该配置项请参考如下步骤修改，如果没有该配置项则版本不支持动态调整会话时长。
- MRS 3.x及之后版本集群：登录FusionInsight Manager，选择“集群 > 服务 > meta”，单击“配置”，选择“全部配置”。搜索配置项“http.server.session.timeout.secs”，如果有该配置项请参考如下步骤修改，如果没有该配置项则版本不支持动态调整会话时长。



所有超时时长的值请设置为统一值，避免时长设置不一致导致的页面实际生效的超时时长和设置值的冲突。

修改 Manager 页面及认证中心的超时时长

MRS 3.x之前版本集群：

1. 分别登录集群所有Master节点，在所有Master节点中执行2-4的修改。
2. 修改“/opt/Bigdata/apache-tomcat-7.0.78/webapps/cas/WEB-INF/web.xml”中的“<session-timeout>20</session-timeout>”，其中20为会话超时时间请根据需要修改，单位为分钟，超时时间最长不要超过480分钟。

3. 修改“/opt/Bigdata/apache-tomcat-7.0.78/webapps/web/WEB-INF/web.xml”中的“<session-timeout>20</session-timeout>”其中20为会话超时时间请根据需要修改，单位为分钟，超时时间最长不要超过480分钟。
4. 修改“/opt/Bigdata/apache-tomcat-7.0.78/webapps/cas/WEB-INF/spring-configuration/ticketExpirationPolicies.xml”中的“p:maxTimeToLiveInSeconds="${tgt.maxTimeToLiveInSeconds:1200}"”和“p:timeToKillInSeconds="${tgt.timeToKillInSeconds:1200}"”，其中1200为认证中心的有效时长请根据需要修改，单位为秒，有效时长不要超过28800秒。
5. 在主管理节点重启Tomcat节点。
   a. 在主master节点上用omm用户执行**netstat -anp |grep 28443 |grep LISTEN | awk '{print $7}'**查询Tomcat的进程号。
   b. 执行**kill -9 {pid}**，其中{pid}为**5.a**中获得的Tomcat进程号。
   c. 等待进程自动重启。可以执行**netstat -anp |grep 28443 |grep LISTEN**查看进程是否重启成功，如果可以查到进程说明已经重启成功，如果未查到请稍后再次查询。

MRS 3.x及之后版本集群：

1. 分别登录集群所有Master节点，在所有Master节点中执行**2**-**3**的修改。
2. 修改“/opt/Bigdata/om-server_xxx/apache-tomcat-xxx/webapps/web/WEB-INF/web.xml”中的“<session-timeout>20</session-timeout>”，其中20为会话超时时间请根据需要修改，单位为分钟，超时时间最长不要超过480分钟。
3. 修改“/opt/Bigdata/om-server_xxx/apache-tomcat-8.5.63/webapps/cas/WEB-INF/classes/config/application.properties”文件，在文件中新增配置ticket.tgt.timeToKillInSeconds=28800，其中28800为认证中心的有效时长请根据需要修改，单位为秒，有效时长不要超过28800秒。
4. 在主管理节点重启Tomcat节点。
   a. 在主master节点上用omm用户执行**netstat -anp |grep 28443 |grep LISTEN | awk '{print $7}'**查询Tomcat的进程号。
   b. 执行**kill -9 {pid}**，其中{pid}为**4.a**中获得的Tomcat进程号。
   c. 等待进程自动重启。可以执行**netstat -anp |grep 28443 |grep LISTEN**查看进程是否重启成功，如果可以查到进程说明已经重启成功，如果未查到请稍后再次查询。

修改开源组件 Web 页面的超时时间

1. 进入服务全部配置界面。

   MRS 3.x之前版本集群：在集群详情页面，选择“组件管理 > meta > 服务配置”。

   MRS 3.x及之后版本集群：登录FusionInsight Manager，选择“集群 > 服务 > meta”，单击“配置”，选择“全部配置”。
2. 根据需要修改“meta”下的“http.server.session.timeout.secs”值，单位为秒。
3. 保存配置，不勾选“重新启动受影响的服务或实例”并单击“确定”。

   重启会影响业务，建议在业务空闲时执行重启操作。
4. （可选）若需要使用Spark的Web页面，则需要Spark“全部配置”页面，搜索并修改配置项“spark.session.maxAge”为合适的值，单位为秒。

   保存配置，不勾选“重新启动受影响的服务或实例”并单击“确定”。

5. 重启meta服务及需要使用Web界面的服务，或者在业务空闲时重启集群。

   重启会影响业务，建议在业务空闲时执行重启操作，或使用滚动重启功能，在不影响业务的情况下重启服务。

15.5.2 MRS 租户管理中的动态资源计划页面无法刷新

步骤1 以root用户分别登录Master1和Master2节点。

步骤2 执行**ps -ef |grep aos**命令检查aos进程号。

步骤3 执行**kill -9 *aos进程号*** 结束aos进程。

步骤4 等待aos进程自动重启成功，可通过**ps -ef |grep aos**命令查询进程是否存在，若存在则重启成功，若不存在请稍后再查询。

----结束

15.5.3 Kafka Topic 监控页签在 Manager 页面不显示

步骤1 分别登录集群Master节点，并切换用户为omm。

步骤2 进入目录“/opt/Bigdata/apache-tomcat-7.0.78/webapps/web/WEB-INF/lib/components/Kafka/”。

步骤3 拷贝zookeeper包到该目录**cp /opt/share/zookeeper-3.5.1-mrs-2.0/zookeeper-3.5.1-mrs-2.0.jar ./**。

步骤4 重启Tomcat。

sh /opt/Bigdata/apache-tomcat-7.0.78/bin/shutdown.sh

sh /opt/Bigdata/apache-tomcat-7.0.78/bin/startup.sh

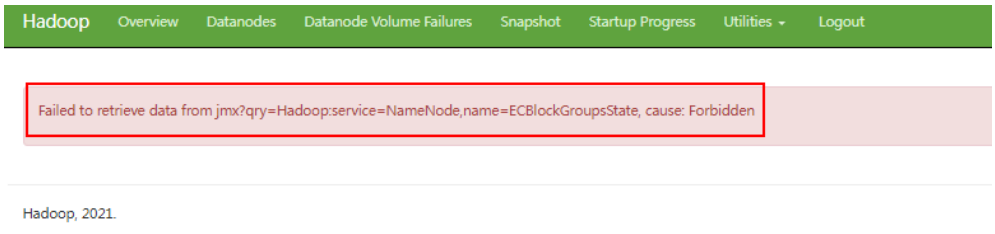
----结束

15.5.4 访问 HDFS、Hue、Yarn、Flink 等组件的 WebUI 界面报错，或部分功能不可用

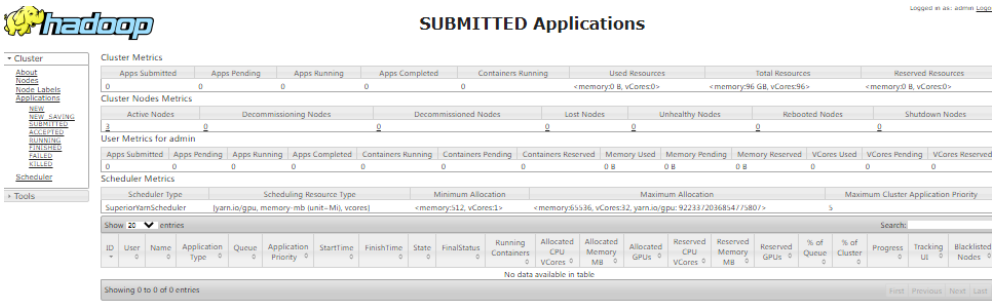
访问HDFS、Hue、Yarn、Flink等组件的WebUI的用户不具备对应组件的管理权限，导致界面报错或部分功能不可用，例如：

- 使用当前用户登录Flink WebUI后，部分内容不能正常显示，且没有权限创建应用、创建集群连接、创建数据连接等：

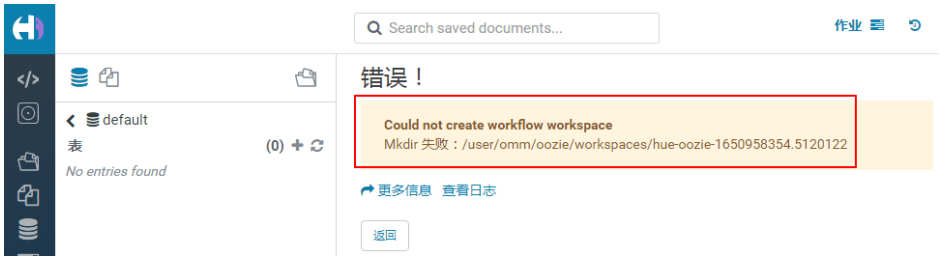
- 使用当前用户访问HDFS WebUI报错：Failed to retrieve data from /jmx?qry=java.lang:type=Memory, cause: Forbidden



- 使用当前用户访问Yarn WebUI界面，无法查看作业信息：



- 使用当前用户登录Hue WebUI后，在界面左侧导航栏单击  ，选择“Workflow”后报错：



建议使用新建的具有对于组件管理权限的用户访问，创建一个业务用户，例如创建一个具有HDFS管理权限的用户登录并访问HDFS WebUI界面。

15.6 监控告警类

15.6.1 在 MRS 流式集群中，Kafka topic 监控是否支持发送告警?

暂不支持Kafka topic监控发送邮件和短信告警，目前用户可以在Manager界面看到告警信息。

15.6.2 产生告警“ALM-18022 Yarn 队列资源不足”时，在哪里可以看到在运行的资源队列

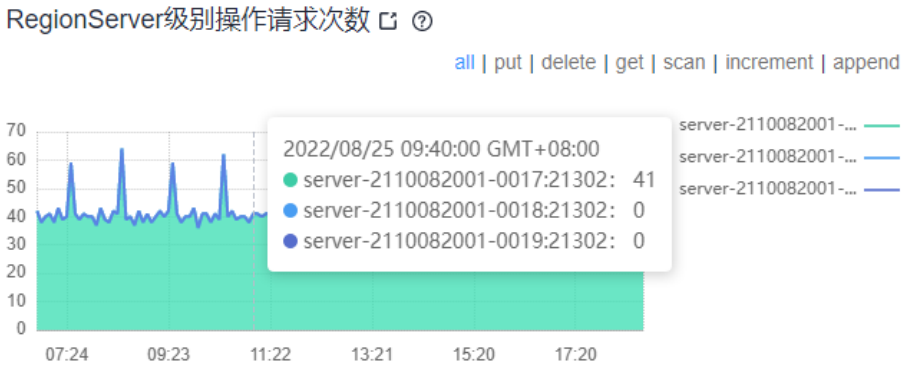
Yarn资源队列可以登录Manager界面，选择“集群 > 服务 > Yarn > ResourceManager(主)”，登录Yarn的原生页面进行查看。

具体的告警处理方法可查看该告警的联机帮助文档进行处理。

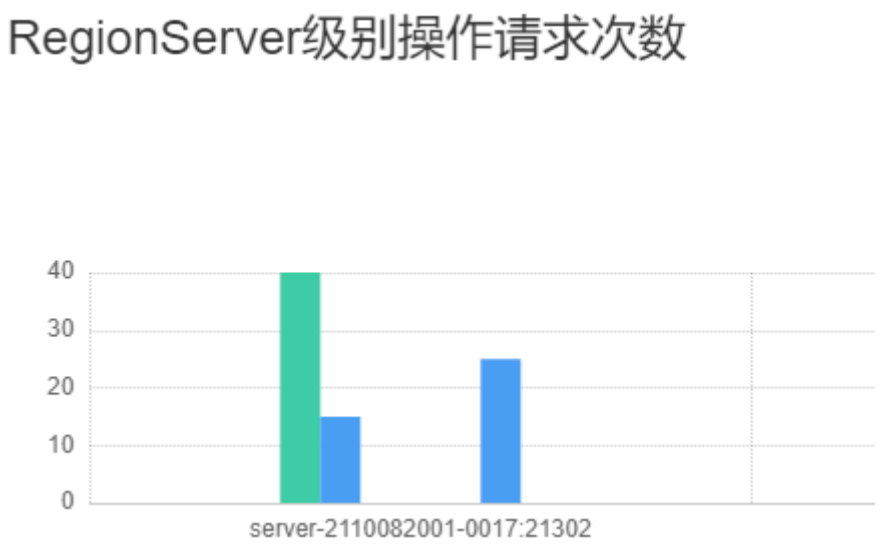
15.6.3 HBase 操作请求次数指标中的多级图表统计如何理解

以“RegionServer级别操作请求次数”监控项为例：

1. 登录FusionInsight Manager，选择“集群 > 服务 > HBase > 资源”，在该界面即可查看“RegionServer级别操作请求次数”图表，选中“all”，则显示当前集群所有RegionServer的所有操作请求次数总和排Top10的值，统计时间间隔为5分钟。

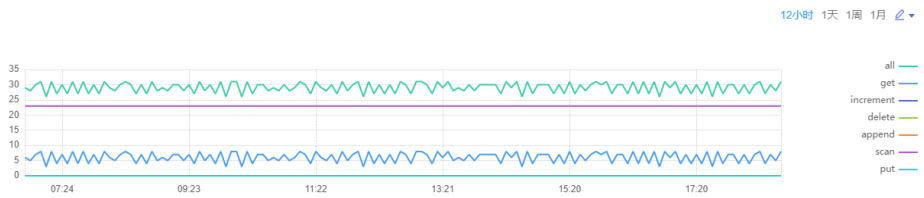


2. 单击表格中某一统计点，即可进入二级图表，表示该时刻前5分钟内统计的所有RegionServer的操作请求数。



3. 再单击某一个操作统计柱状图即可进入三级图表，表示该时间段内各个Region相应操作的分布情况。

4. 单击某个Region名称，进入该Region在12小时内每5分钟做的操作数统计分布图，可查看具体的操作在该时间段内执行的次数。



15.7 性能优化类

15.7.1 MRS 集群是否支持重装系统?

MRS集群不支持重装系统。

15.7.2 MRS 集群是否支持切换操作系统?

MRS集群节点不支持切换操作系统。

15.7.3 如何提高集群 Core 节点的资源使用率?

1. 搜索并修改“yarn.nodemanager.resource.memory-mb”的值，请根据集群的节点内存实际情况调大该值。
2. 保存配置并重启受影响的服务或实例。

15.7.4 如何关闭防火墙服务?

步骤1 以root用户登录集群的各个节点。

步骤2 检查防火墙服务是否启动。

例如，EulerOS环境下执行**systemctl status firewalld.service**命令。

步骤3 关闭防火墙服务。

例如，EulerOS环境下执行**systemctl stop firewalld.service**命令。

----结束

15.8 作业开发类

15.8.1 如何准备 MRS 的数据源?

MRS既可以处理OBS中的数据，也可以处理HDFS中的数据。在使用MRS分析数据前，需要先准备数据。

1. 将本地数据上传OBS。
   a. 登录OBS管理控制台。
   b. 在OBS上创建userdata并行文件系统，然后在userdata文件系统下创建program、input、output和log文件夹。
      i. 单击“并行文件系统 > 创建并行文件系统”，创建一个名称为userdata的文件系统。

ii. 在OBS文件系统列表中单击文件系统名称userdata，选择“文件 > 新建文件夹”，分别创建program、input、output和log目录。

c. 上传数据至userdata文件系统。

i. 进入program文件夹，单击“上传文件”。

ii. 单击“添加文件”并选择用户程序。

iii. 单击“上传”。

iv. 使用同样方式将用户数据文件上传至input目录。

2. 将OBS数据导入至HDFS。

当“Kerberos认证”为“关闭”，且运行中的集群，可执行将OBS数据导入至HDFS的操作。

a. 登录MRS管理控制台。

b. 单击集群名称进入集群详情页面。

c. 单击“文件管理”，选择“HDFS文件列表”。

d. 进入数据存储目录，如“bd_app1”。

“bd_app1”目录仅为示例，可以是界面上的任何目录，也可以通过“新建”创建新的目录。

e. 单击“导入数据”，通过单击“浏览”选择OBS和HDFS路径。

f. 单击“确定”。

文件上传进度可在“文件操作记录”中查看。

15.8.2 集群支持提交哪些形式的 Spark 作业?

当前在MRS页面，集群支持提交Spark、Spark Script和Spark SQL形式的Spark作业。

15.8.3 MRS 集群的租户资源最小值改为 0 后，只能同时跑一个 Spark 任务吗?

MRS集群的租户资源最小值改为0后，只能同时跑一个Spark任务。

15.8.4 Spark 作业 Client 模式和 Cluster 模式的区别

理解YARN-Client和YARN-Cluster深层次的区别之前先清楚一个概念：Application Master。

在YARN中，每个Application实例都有一个ApplicationMaster进程，它是Application启动的第一个容器。它负责和ResourceManager打交道并请求资源，获取资源之后告诉NodeManager为其启动Container。从深层次的含义讲YARN-Cluster和YARN-Client模式的区别其实就是ApplicationMaster进程的区别。

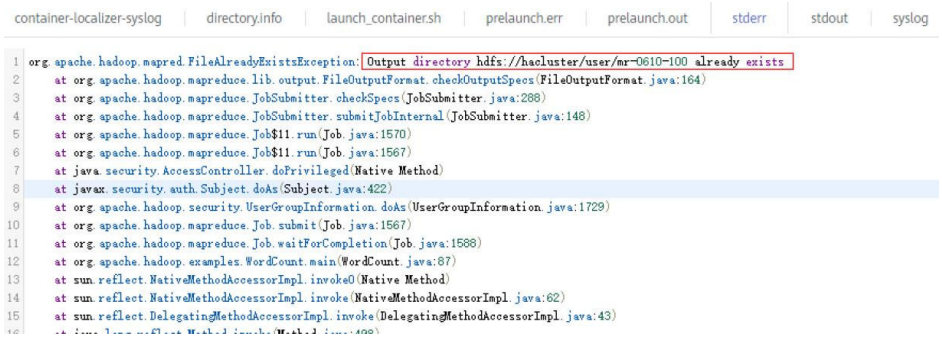
YARN-Cluster模式下，Driver运行在AM(Application Master)中，它负责向YARN申请资源，并监督作业的运行状况。当用户提交了作业之后，就可以关掉Client，作业会继续在YARN上运行，因而YARN-Cluster模式不适合运行交互类型的作业。

YARN-Client模式下，Application Master仅仅向YARN请求Executor，Client会和请求的Container通信来调度他们工作，也就是说Client不能离开。

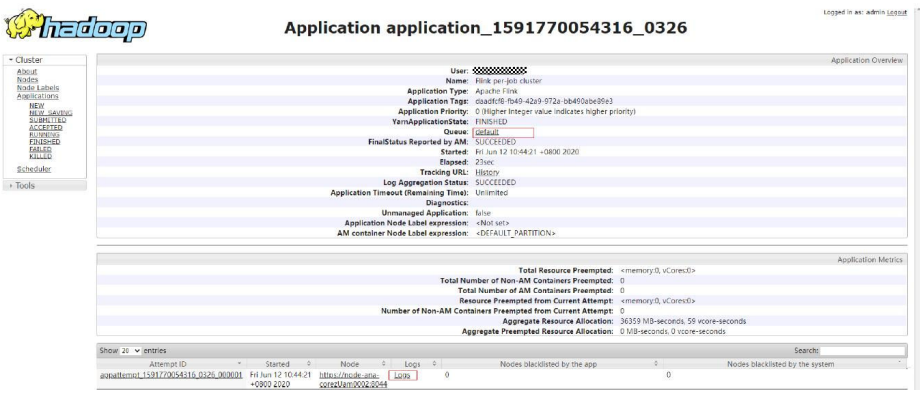
15.8.5 如何查看 MRS 作业日志?

步骤1 MRS Console页面作业管理，每一条作业支持查看日志，包含launcherJob日志和realJob日志。

- launcherJob作业的日志，一般会在stderr和stdout中打印错误日志，如下图所示：



- realJob的日志，可以通过MRS Manager中 Yarn服务提供的ResourceManager Web UI查看。



步骤2 登录集群Master节点，可获取步骤1作业的日志文件 ，具体hdfs路径为“/tmp/logs/*{submit_user}*/logs/*{application_id}*”。

步骤3 提交作业后，在Yarn的WEB UI未找到对应作业的application_id，说明该作业没有提交成功，可登录集群主Master节点，查看提交作业进程日志“/var/log/executor/logs/exe.log”。

----结束

15.8.6 报错提示“当前用户在 MRS Manager 不存在，请先在 IAM 给予该用户足够的权限，再在概览页签进行 IAM 用户同步”

安全集群使提交作业时，未进行IAM用户同步，会出现“当前用户在MRS Manager不存在，请先在IAM给予该用户足够的权限，再在概览页签进行IAM用户同步”错误。

需要在提交作业之前，先在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，然后再提交作业。

15.8.7 LauncherJob 作业执行结果为 Failed. 报错信息为：jobPropertiesMap is null.

launcher作业失败的原因为：提交作业用户无“hdfs /mrs/job-properties”目录的写权限。

该问题在2.1.0.6的补丁中修复，也可通过在MRS Manager页面给同步的提交作业用户赋予该目录“/mrs/job-properties”的写入权限。

15.8.8 MRS Console 页面 Flink 作业状态与 Yarn 上的作业状态不一致

为了节约存储空间，用户修改了Yarn的配置项yarn.resourcemanager.max-completed-applications，减小yarn上历史作业的记录保存个数。由于Flink是长时作业，在yarn上realJob还在运行，但launcherJob已经被删除，导致因从Yarn上查不到launcherJob，从而更新作业状态失败。该问题在2.1.0.6补丁中解决。

规避方法：终止找不到launcherJob的作业，后续提交的作业状态就会更新。

15.8.9 提交长时作业 SparkStreaming，运行几十个小时后失败，报 OBS 访问 403

当用户提交作业需要读写OBS时，提交作业程序会默认为用户添加访问OBS的临时accesskey和secretkey，但是临时accesskey和secretkey有过期时间。

如果需要运行像Flink和SparkStreaming这样的长时作业时，用户可通过“服务配置参数”选项框传入永久的accesskey和secretkey，以保证作业不会在运行过程中因密钥过期而执行失败。

15.8.10 ClickHouse 客户端执行 SQL 查询时报内存不足问题

问题现象

ClickHouse会限制group by使用的内存量，在使用ClickHouse客户端执行SQL查询时报如下错误：

```
Progress: 1.83 billion rows, 85.31 GB (68.80 million rows/s., 3.21 GB/s.)  6%Received exception from server:
Code: 241. DB::Exception: Received from localhost:9000, 127.0.0.1.
DB::Exception: Memory limit (for query) exceeded: would use 9.31 GiB (attempt to allocate chunk of 1048576 bytes), maximum: 9.31 GiB:
(while reading column hits):
```

解决方法

- 在执行SQL语句前，执行如下命令。注意执行前保证集群有足够内存可以设置。
  ```
  SET max_memory_usage = 128000000000; #128G
  ```
- 如果没有上述大小内存可用，ClickHouse可以通过如下设置将“溢出”数据到磁盘。建议将max_memory_usage设置为max_bytes_before_external_group_by大小的两倍。
  ```
  set max_bytes_before_external_group_by=20000000000; #20G
  set max_memory_usage=40000000000; #40G
  ```

15.8.11 Spark 运行作业报错：java.io.IOException: Connection reset by peer

问题现象

Spark作业运行一直不结束，查看日志报错：java.io.IOException: Connection reset by peer

解决方法

修改提交参数，加上参数“executor.memoryOverhead”。

15.8.12 Spark 作业访问 OBS 报错：requestId=4971883851071737250

问题现象

Spark作业访问OBS报错：requestId=4971883851071737250

解决方法

登录Spark客户端节点，进入conf目录，修改配置文件“core-site.xml”中的“fs.obs.metrics.switch”参数值为“false”。

15.8.13 DataArts Studio 调度 spark 作业，偶现失败，重跑失败

问题现象

DataArts Studio调度spark作业，偶现失败，重跑失败，作业报错：

```
Caused by: org.apache.spark.SparkException: Application application_1619511926396_2586346 finished with failed status
```

解决方法

使用root用户登录Spark客户端节点，调高“spark-defaults.conf”文件中“spark.driver.memory”参数值。

15.8.14 Flink 任务运行失败，报错：java.lang.NoSuchFieldError: SECURITY_SSL_ENCRYPT_ENABLED

问题现象

Flink任务运行失败，报错：

```
Caused by: java.lang.NoSuchFieldError: SECURITY_SSL_ENCRYPT_ENABLED
```

解决方法

客户代码里面打包的第三方依赖包和集群包冲突，提交到MRS集群运行失败，需修改相关的依赖包，并将pom文件中的开源版本的Hadoop包和Flink包的作用域设置为provide，添加完成后重新打包运行任务。

15.8.15 提交的 Yarn 作业在界面上查看不到

创建完Yarn作业后，以admin用户登录界面查看不到运行的作业。

- admin用户为集群管理页面用户，检查是否有supergroup权限，一般需要使用具有supergroup权限的用户才可以查看作业。
- 一般使用提交作业的用户登录查看Yarn上的作业。不使用admin管理帐号查看。

15.8.16 如何修改现有集群的 HDFS NameSpace(fs.defaultFS)

当前不建议在服务端修改或者新增集群的HDFS NameSpace(fs.defaultFS)，如果只是为了客户端更好的识别，则一般可以通过修改客户端的“core-site.xml”，“hdfs-site.xml”两个文件进行实现。

15.8.17 通过管控面提交 Flink 任务时 launcher-job 因 heap size 不够被 Yarn 结束

问题现象

管控面提交Flink任务时launcher-job被Yarn结束。

解决方法

调大launcher-job的heap size值。

1. 使用**omm**用户登录主OMS节点。
2. 修改“/opt/executor/webapps/executor/WEB-INF/classes/servicebroker.xml”中参数“job.launcher.resource.memory.mb”的值为“2048”。
3. 使用**sh /opt/executor/bin/restart-executor.sh**重启executor进程。

15.8.18 Flink 作业提交时报错 slot request timeout

问题现象

Flink作业提交时，jobmanager启动成功，但taskmanager一直是启动中直到超时，报错如下：

```
org.apache.flink.runtime.jobmanager.scheduler.NoResourceAvailableException: Could not allocate the required slot within slot request timeout. Please make sure that the cluster has enough resources
```

可能原因

1. Yarn队列中资源不足，导致创建taskmanager启动不成功。
2. 用户的Jar包与环境中的Jar包冲突导致，可以通过执行wordcount程序是否成功来判断。
3. 若集群为安全集群，可能是Flink的SSL证书配置错误，或者证书过期。

解决方法

1. 增加队列的资源。
2. 排除用户Jar包中的Flink和Hadoop依赖，依靠环境中的Jar包。
3. 重新配置Flink的SSL证书。

15.8.19 DistCP 类型作业导入导出数据问题

- DistCP类型作业导入导出数据时，是否会对比数据的一致性?

  DistCP类型作业导入导出数据时不会对比数据的一致性，只是对数据进行拷贝，不会修改数据。
- DistCP类型作业在导出时，遇到OBS里已经存在的文件是如何处理的?

  DistCP类型作业在导出时，遇到OBS里已经存在的文件时会覆盖原始文件。

15.9 集群升级/补丁

15.9.1 MRS 版本如何进行升级?

MRS目前还无法实现低版本到高版本的平滑升级。目前只能重新创建一个新版本的集群，然后将老版本集群的数据迁移到新的集群。

15.9.2 MRS 是否支持修改版本?

MRS不支持修改版本，建议删除集群之后重新创建集群。

15.10 集群访问类

15.10.1 MRS 登录集群节点的两种方式能够切换么?

不可以。创建集群时选择登录方式后不能更改登录方式。

15.10.2 如何获取 ZooKeeper 的 IP 地址和端口?

ZooKeeper的IP地址和端口可以通过MRS控制台或登录Manager界面获取。

方法一：通过MRS控制台获取

1. 在MRS集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。
2. 选择“组件管理 > ZooKeeper > 实例”，获取ZooKeeper实例的“业务IP”地址。
3. 选择“服务配置”页签，搜索“clientPort”参数，该参数的值即为ZooKeeper的端口。

方法二：通过Manager界面获取

1. 登录Manager界面。
2. 在Manager界面获取ZooKeeper的IP地址和端口。

- 针对MRS 3.x之前版本集群
  i. 选择“服务管理 > ZooKeeper > 实例”，获取ZooKeeper实例的“业务IP”地址。
  ii. 选择“服务配置”页签，搜索“clientPort”参数，该参数的值即为ZooKeeper的端口。
- 针对MRS 3.x及之后版本集群
  i. 选择“集群 > 服务 > ZooKeeper > 实例”，获取ZooKeeper实例的“业务IP”地址。
  ii. 选择“配置”页签，搜索参数“clientPort”值，该参数的值即为ZooKeeper的端口。

15.10.3 如何通过集群外的节点访问 MRS 集群?

创建集群外 Linux 操作系统 ECS 节点访问 MRS 集群

步骤1 创建一个集群外ECS节点。

ECS节点的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。

步骤2 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

步骤3 配置集群安全组规则。

1. 在集群“概览”界面，选择“添加安全组规则 > 管理安全组规则”。
2. 在“入方向规则”页签，选择“添加规则”，在“添加入方向规则”配置ECS节点的IP和放开所有端口。
3. 安全组规则添加完成后，可以直接下载并安装客户端到集群外ECS节点。
4. 使用客户端。

   使用客户端安装用户，登录客户端节点，执行以下命令切换到客户端目录。

   cd /opt/hadoopclient

   执行以下命令加载环境变量。

   source bigdata_env

   如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则通常情况下无需认证。

   kinit *MRS集群用户*

   例如：

   kinit admin

   直接执行组件的客户端命令。

   例如：

   使用HDFS客户端命令查看HDFS根目录文件。

   hdfs dfs -ls /

```
Found 15 items
drwxrwx--x   - hive      hive              0 2021-10-26 16:30 /apps
drwxr-xr-x   - hdfs      hadoop            0 2021-10-18 20:54 /datasets
drwxr-xr-x   - hdfs      hadoop            0 2021-10-18 20:54 /datastore
drwxrwx---+  - flink     hadoop            0 2021-10-18 21:10 /flink
```

```
drwxr-x---   - flume      hadoop              0 2021-10-18 20:54 /flume
drwxrwx--x   - hbase      hadoop              0 2021-10-30 07:31 /hbase
...
```

----结束

15.11 大数据业务开发

15.11.1 MRS 是否支持同时运行多个 Flume 任务?

Flume客户端可以包含多个独立的数据流，即在一个配置文件properties.properties中配置多个Source、Channel、Sink。这些组件可以链接以形成多个流。

例如在一个配置中配置两个数据流，示例如下：

```
server.sources = source1 source2
server.sinks = sink1 sink2
server.channels = channel1 channel2

#dataflow1
server.sources.source1.channels = channel1
server.sinks.sink1.channel = channel1

#dataflow2
server.sources.source2.channels = channel2
server.sinks.sink2.channel = channel2
```

15.11.2 如何修改 FlumeClient 的日志为标准输出日志?

1. 登录Flume客户端安装节点。
2. 进入Flume客户端安装目录，假设Flume客户端安装路径为“/opt/FlumeClient”，可以执行以下命令。

   cd /opt/FlumeClient/fusioninsight-flume-1.9.0/bin
3. 执行**./flume-manage.sh stop**命令，停止FlumeClient。
4. 执行**vi /log4j.properties**命令，打开log4j.properties文件，修改“flume.root.logger”的取值为“${flume.log.level},console”。
5. 执行**vim /flume-manager.sh**命令，打开flume安装目录bin目录下的启动脚本flume-manager.sh。
6. 修改flume-manager.sh脚本，注释如下内容。

   >/dev/null 2>&1 &
7. 执行**./flume-manage.sh start**命令，重启FlumeClient。
8. 修改完成后，请检查docker配置信息是否正确。

15.11.3 Hadoop 组件 jar 包位置和环境变量的位置在哪里?

- hadoopstreaming.jar位置在**/opt/share/hadoop-streaming-***目录下。其中*由Hadoop版本决定。
- jdk环境变量：**/opt/client/JDK/component_env**
- Hadoop组件的环境变量位置：**/opt/client/HDFS/component_env**
- Hadoop客户端路径：**/opt/client/HDFS/hadoop**

15.11.4 HBase 支持的压缩算法有哪些?

HBase目前支持的压缩算法有snappy、lz4和gz。

15.11.5 MRS 是否支持通过 Hive 的 HBase 外表将数据写入到 HBase?

不支持。Hive on HBase只支持查询，不支持更改数据。

15.11.6 如何查看 HBase 日志?

1. 使用**root**用户登录集群的Master节点。
2. 执行**su - omm**命令，切换到**omm**用户。
3. 执行**cd /var/log/Bigdata/hbase/**命令，进入到“/var/log/Bigdata/hbase/”目录，即可查看HBase日志信息。

15.11.7 HBase 表如何设置和修改数据保留期?

- 创建表时指定

  创建t_task_log表，列族f, TTL设置86400秒过期

  ```
  create 't_task_log',{NAME => 'f', TTL=>'86400'}
  ```

- 在已有表的基础上指定：

  ```
  disable "t_task_log"  #禁用表（这个需要停止业务）
  alter "t_task_log",NAME=>'data',TTL=>'86400' #设置TTL值，作用于列族data
  enable  "t_task_log"  #恢复表
  ```

15.11.8 HDFS 如何进行数据均衡?

1. 登录集群的Master节点，并执行以下命令配置环境变量。其中“/opt/client”为客户端安装目录，具体以实际为准。

   source /opt/client/bigdata_env

   kinit *组件业务用户*（如果集群已开启kerberos认证，则执行该命令进行用户认证。未开启kerberos认证的集群无需执行该命令。）

2. 执行如下命令启动balancer。

   /opt/client/HDFS/hadoop/sbin/start-balancer.sh -threshold 5

3. 查看日志。

   balance任务执行时会在客户端安装目录“/opt/client/HDFS/hadoop/logs”目录下生成名为hadoop-root-balancer-主机名.log日志。

4. （可选）若不想再进行数据均衡，可执行如下命令停止balancer。

   source /opt/client/bigdata_env

   kinit *组件业务用户*（如果集群已开启kerberos认证，则执行该命令进行用户认证。未开启kerberos认证的集群无需执行该命令。）

   /opt/client/HDFS/hadoop/sbin/stop-balancer.sh -threshold 5

15.11.9 如何修改 HDFS 的副本数?

1. 搜索并修改“dfs.replication”的值，合理修改这个数值，该参数取值范围为1~16，重启HDFS实例。

15.11.10 如何使用 Python 远程连接 HDFS 的端口?

HDFS开源HTTP端口3.0.0以前版本的默认端口为50070，3.0.0及以后的默认端口为9870。HDFS开源组件的端口号如**HDFS常用端口**所示。

HDFS 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
dfs.namenode.rpc.port	• 9820（MRS 3.x之前版本） • 8020（MRS 3.x及之后版本）	NameNode RPC 端口。 该端口用于： 1. HDFS客户端与Namenode间的通信。 2. Datanode与NameNode之间的连接。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
dfs.namenode.http.port	9870	HDFS HTTP端口(NameNode)。 该端口用于： 1. 点对点的NameNode检查点操作。 2. 远程Web客户端连接NameNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是
dfs.namenode.https.port	9871	HDFS HTTPS端口(NameNode)。 该端口用于： 1. 点对点的NameNode检查点操作。 2. 远程Web客户端连接NameNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 • 安装时是否缺省启用：是 • 安全加固后是否启用：是

配置参数	默认端口	端口说明
dfs.datanode.ipc.port	9867	Datanode IPC 服务器端口。 该端口用于： 客户端连接DataNode用来执行RPC操作。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.port	9866	Datanode数据传输端口。 该端口用于： 1. HDFS客户端从DataNode传输数据或传输数据到DataNode。 2. 点对点的Datanode传输数据。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.http.port	9864	Datanode HTTP端口。 该端口用于： 安全模式下，远程Web客户端连接DataNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.https.port	9865	Datanode HTTPS端口。 该端口用于： 安全模式下，远程Web客户端连接DataNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口	端口说明
dfs.JournalNode.rpc.port	8485	JournalNode RPC端口。 该端口用于： 客户端通信用于访问多种信息。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.journalnode.http.port	8480	JournalNode HTTP端口。 该端口用于： 安全模式下，远程Web客户端链接JournalNode。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.journalnode.https.port	8481	JournalNode HTTPS端口。 该端口用于： 安全模式下，远程Web客户端链接JournalNode。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
httpfs.http.port	14000	HttpFS HTTP服务器侦听的端口。 该端口用于： 远程REST接口连接HttpFS。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

15.11.11 如何修改 HDFS 主备倒换类?

当MRS 3.x版本集群使用HDFS连接NameNode报类org.apache.hadoop.hdfs.server.namenode.ha.AdaptiveFailoverProxyProvider无法找到时，是由于MRS 3.x版本集群HDFS的主备倒换类默认为该类，可通过如下方式解决。

- 方式一：添加hadoop-plugins-xxx.jar到程序的classpath或者lib目录中。

  hadoop-plugins-xxx.jar包一般在HDFS客户端目录下：$HADOOP_HOME/share/hadoop/common/lib/hadoop-plugins-8.0.2-302023.jar
- 方式二：将HDFS的如下配置项修改为开源类：

  dfs.client.failover.proxy.provider.hacluster=org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider

15.11.12 DynamoDB 的 number 在 Hive 表中用什么类型比较好?

Hive支持smallint，推荐使用smallint类型。

15.11.13 Hive Driver 是否支持对接 dbcp2?

Hive driver不支持对接dbcp2数据库连接池。dbcp2数据库连接池调用isValid方法检查连接是否可用，而Hive对于这个方法的实现就是直接报错。

15.11.14 用户 A 如何查看用户 B 创建的 Hive 表?

MRS 3.x之前版本：

1. 登录MRS Manager，选择“系统设置 > 权限配置 > 角色管理”。
2. 单击“添加角色”，输入“角色名称”和“描述”。
3. 在“权限”的表格中选择“Hive > Hive Read Write Privileges”。
4. 在数据库列表中单击用户B创建的表所在的数据库名称，显示用户B创建的表。
5. 在用户B创建的表的“权限”列，勾选“Select”。
6. 单击“确定”，返回“角色”。
7. 选择“系统设置 > 用户管理”，在用户A所在的行，单击“修改”，为用户A绑定新创建的角色，单击“确定”，等待5分钟左右即可访问到用户B创建的表。

MRS 3.x及之后版本：

1. 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 更多”，查看“启用Ranger鉴权”是否置灰。
   - 是，执行**9**。
   - 否，执行**2**-**8**。
2. 登录FusionInsight Manager，选择“系统 > 权限 > 角色”
3. 单击“添加角色”，输入“角色名称”和“描述”。
4. 在“配置资源权限”的表格中选择“*待操作集群的名称* > Hive > Hive读写权限”。
5. 在数据库列表中单击用户B创建的表所在的数据库名称，显示用户B创建的表。
6. 在用户B创建的表的“权限”列，勾选“查询”。
7. 单击“确定”，返回“角色”
8. 单击“用户”，在用户A所在行，单击“修改”，为用户A绑定新创建的角色，单击“确定”，等待5分钟左右即可访问到用户B创建的表。
9. 添加Hive的Ranger访问权限策略：

   a. 使用Hive管理员用户登录FusionInsight Manager，选择“集群 > 服务 > Ranger”，单击“Ranger WebUI”右侧的链接进入Ranger管理界面。

b. 在首页中单击“HADOOP SQL”区域的组件插件名称，例如“Hive”。
c. 在“Access”页签单击“Add New Policy”，添加Hive权限控制策略。
d. 在“Create Policy”页面填写如下内容：
   - Policy Name：策略名称，例如：table_test_hive。
   - database：填写或选择用户B创建的表所在的数据库，例如：default。
   - table：填写或选择用户B创建的表，例如：test。
   - column：填写并选择对应的列，例如：*。
   - 在“Allow Conditions”区域，单击“Select User”下选择框选择用户A，单击“Add Permissions”，勾选“select”。
   - 单击“Add”。

10. 添加HDFS的Ranger访问权限策略：
    a. 使用**rangeradmin**用户登录FusionInsight Manager，选择“集群 > 服务 > Ranger”，单击“Ranger WebUI”右侧的链接进入Ranger管理界面。
    b. 在首页中单击“HDFS”区域的组件插件名称，例如“hacluster”。
    c. 单击“Add New Policy”，添加HDFS权限控制策略.
    d. 在“Create Policy”页面填写如下内容：
       - Policy Name：策略名称，例如：tablehdfs_test。
       - Resource Path：配置用户B创建的表所在的HDFS路径，例如：/user/hive/warehouse/*数据库名称/表名*
       - 在“Allow Conditions”区域，单击“Select User”下选择框选择用户A，单击“Add Permissions”，勾选“Read”和“Execute”。
       - 单击“Add”。
11. 在策略列表可查看策略的基本信息。等待策略生效后，用户A即可查看用户B创建的表。

15.11.15 Hive 查询数据是否支持导出?

Hive查询数据支持导出，请参考如下语句进行导出：

```
insert overwrite local directory "/tmp/out/" row format delimited fields terminated by "\t" select * from table;
```

15.11.16 Hive 使用 beeline -e 执行多条语句报错

MRS 3.x版本Hive使用beeline执行beeline -e " use default;show tables;"报错：Error while compiling statement: FAILED: ParseException line 1:11 missing EOF at ';' near 'default' (state=42000,code=40000)。

处理方法：

- 方法一：使用beeline --entirelineascommand=false -e "use default;show tables;"。
- 方法二：

a. 在Hive客户端如/opt/Bigdata/client/Hive目录下修改component_env文件，修改**export CLIENT_HIVE_ENTIRELINEASCOMMAND=true**为**export CLIENT_HIVE_ENTIRELINEASCOMMAND=false**。

图 15-1 修改 component_env 文件

b. 执行如下命令验证配置。

source /opt/Bigdata/client/bigdata_env

beeline -e " use default;show tables;"

15.11.17 添加 Hive 服务后，提交 hivesql/hivescript 作业失败

该问题是由于提交作业的用户所在用户组绑定的MRS CommonOperations策略权限在同步到Manager中后没有Hive相关权限，处理方法如下：

1. 添加Hive服务完成后。
2. 登录IAM服务控制台，创建一个用户组，该用户组所绑定策略和提交作业用户所在用户组权限相同。
3. 将提交作业的用户添加到新用户组中。
4. 刷新MRS控制台集群详情页面，“IAM用户同步”会显示“未同步”。
5. 单击“IAM用户同步”右侧的“同步”。同步状态在MRS控制台页面选择“操作日志”查看当前用户是否被修改。
   - 是，则可以重新提交hive作业，
   - 否，则检视上述步骤是否全部已执行完成。
     - 是，请联系运维人员处理。
     - 否，请等待执行完成后再提交hive作业。

15.11.18 Hue 下载 excel 无法打开

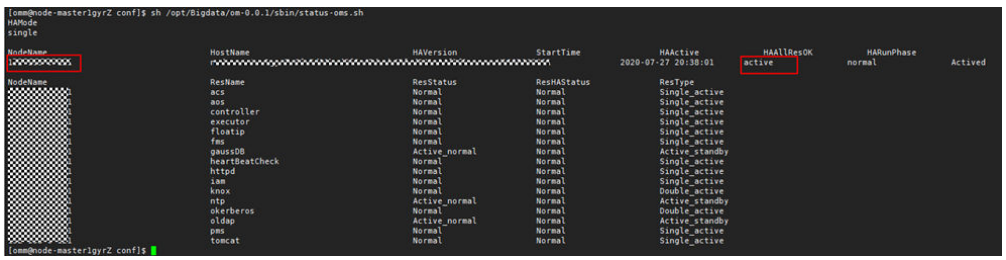
1. 以root用户登录任意一个Master节点，切换到omm用户。

   su - omm
2. 使用如下命令查看当前节点是否为oms主节点。

   sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh

   回显**active**即为主节点，否则请登录另一个Master节点。

图 15-2 oms 主节点



3. 进入“${BIGDATA_HOME}/Apache-httpd-*/conf”目录。

   cd ${BIGDATA_HOME}/Apache-httpd-*/conf

4. 打开httpd.conf文件。

   vim httpd.conf

5. 在文件中搜索21201，并删除文件中的如下内容。proxy_ip和proxy_port对应实际环境中的值。

   ProxyHTMLEnable On
   SetEnv PROXY_PREFIX=https://[proxy_ip]:[proxy_port]
   ProxyHTMLURLMap (https?:\/\/[^:]*:[0-9]*.*) ${PROXY_PREFIX}/proxyRedirect=$1 RV

   图 15-3 待删除内容

6. 退出并保存修改。
7. 再次打开httpd.conf文件，搜索**proxy_hue_port**，并删除如下内容。

   ProxyHTMLEnable On
   SetEnv PROXY_PREFIX=https://[proxy_ip]:[proxy_port]
   ProxyHTMLURLMap (https?:\/\/[^:]*:[0-9]*.*) ${PROXY_PREFIX}/proxyRedirect=$1 RV

   图 15-4 待删除内容

8. 退出并保存修改。
9. 执行如下命令重启httpd。

   sh ${BIGDATA_HOME}/Apache-httpd-*/setup/restarthttpd.sh

10. 检查备Master节点上的httpd.conf文件是否已修改，若已修改则处理完成，若未修改，参考上述步骤进行修改备Master节点的httpd.conf文件，无需重启httpd。
11. 重新下载excel即可打开。

15.11.19 Hue 连接 hiveserver，不释放 session，报错 over max user connections 如何处理?

适用版本：MRS 3.1.0及之前的MRS 3.x版本。

1. 修改两个Hue节点的以下文件：

   /opt/Bigdata/FusionInsight_Porter_8.*/install/FusionInsight-Hue-*/hue/apps/beeswax/src/beeswax/models.py

2. 修改文件中的396和404行的值

   q = self.filter(owner=user, application=application).exclude(guid='').exclude(secret='')改为q = self.filter(owner=user, application=application).exclude(guid=None).exclude(secret=None)

15.11.20 如何重置 Kafka 数据?

删除Kafka topic信息即重置Kafka数据，具体命令请参考：

- 删除topic：**kafka-topics.sh --delete --zookeeper** *ZooKeeper集群业务IP*:**2181/kafka --topic** ***topicname***
- 查询所有topic：**kafka-topics.sh --zookeeper** *ZooKeeper集群业务IP*:**2181/kafka --list**

执行删除命令后topic数据为空则此topic会立刻被删除，如果有数据则会标记删除，后续Kafka会自行进行实际删除。

15.11.21 MRS Kafka 如何查看客户端版本信息?

用如下命令 --bootstrap-server 查看新版本客户端信息。

15.11.22 Kafka 目前支持的访问协议类型有哪些?

当前支持4种协议类型的访问：PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL。

15.11.23 消费 kafka topic，报错：Not Authorized to access group xxx

该问题由于由于集群的Ranger鉴权和集群自带的ACL鉴权冲突导致。Kafka集群使用自带的ACL进行权限访问控制，且集群的Kafka服务也开启Ranger鉴权控制时，该组件所有鉴权将由Ranger统一管理，原鉴权插件设置的权限将会失效，导致ACL权限授权未生效。可通过关闭Kafka的Ranger鉴权并重启Kafka服务来处理该问题。操作步骤如下：

1. 登录FusionInsight Manager页面，选择“集群 > Kafka”。
2. 在服务“概览”页面右上角单击“更多”，选择“停用Ranger鉴权”。在弹出的对话框中输入密码，单击“确定”，操作成功后单击“完成”。
3. 在服务“概览”页面右上角单击“更多”，选择“重启服务”。重启Kafka服务。

15.11.24 Kudu 支持的压缩算法有哪些?

Kudu目前支持的压缩算法有**snappy**、**lz4**和**zlib**，默认是**lz4**。

15.11.25 如何查看 Kudu 日志?

1. 登录集群的Master节点。
2. 执行**su - omm**命令，切换到omm用户。
3. 执行**cd /var/log/Bigdata/kudu/**命令，进入到“/var/log/Bigdata/kudu/”目录，即可查看Kudu日志信息。

15.11.26 新建集群 Kudu 服务异常处理

查看 Kudu 服务异常日志

1. 登录MRS管理控制台。
2. 单击集群名称进入集群详情页面。
3. 选择“组件管理 > Kudu > 实例”，找到异常实例所属的IP。

   若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
4. 登录异常实例IP所在节点，查看Kudu日志。

   ```
   cd /var/log/Bigdata/Kudu
   [root@node-master1AERu kudu]# ls
   healthchecklog  runninglog  startlog
   ```

   其中healthchecklog 目录保存Kudu健康检查日志， startlog保存启动日志，runninglog保存Kudu进程运行日志。

   ```
   [root@node-master1AERu logs]# pwd
   /var/log/Bigdata/kudu/runninglog/master/logs
   [root@node-master1AERu logs]# ls -al
   kudu-master.ERROR   kudu-master.INFO   kudu-master.WARNING
   ```

   运行日志分ERROR, INFO, WARNING三类, 每类会单独打印到相应的文件中，通过**cat**命令即可查看。

已知 Kudu 服务异常处理

日志/var/log/Bigdata/kudu/runninglog/master/logs/kudu-master.INFO 出现异常打印

```
"Unable to init master catalog manager: not found: Unable to initialize catalog manager: Failed to initialize sys tables async: Unable to load consensus metadata for tablet 00000000000000000000000000: xxx"
```

如果该异常是Kudu 服务初次安装时出现，可能是KuduMaster没能同时启动，造成数据不一样导致启动失败。可以通过如下步骤清空数据目录，重启Kudu服务解决。若非初次安装，清空数据目录会造成数据丢失，请先进行数据迁移再进行数据目录清空操作慎重操作。

1. 查找数据目录 fs_data_dir, fs_wal_dir, fs_meta_dir。

   find /opt -name master.gflagfile

   cat /opt/Bigdata/FusionInsight_Kudu_*/*_KuduMaster/etc/master.gflagfile | grep fs_
2. 在集群详情页面选择“组件管理 > Kudu”，单击“停止服务”。
3. 在所有KuduMaster, KuduTserver的节点清空 Kudu 数据目录，如下命令以两个数据盘为例，具体命令请以实际情况为准。

   rm -Rvf /srv/Bigdata/data1/kudu, rm -Rvf /srv/Bigdata/data2/kudu
4. 在集群详情页面选择“组件管理 > Kudu”，单击“更多 > 重启服务”。

5. 查看Kudu服务状态和日志。

15.11.27 OpenTSDB 是否支持 python 的接口?

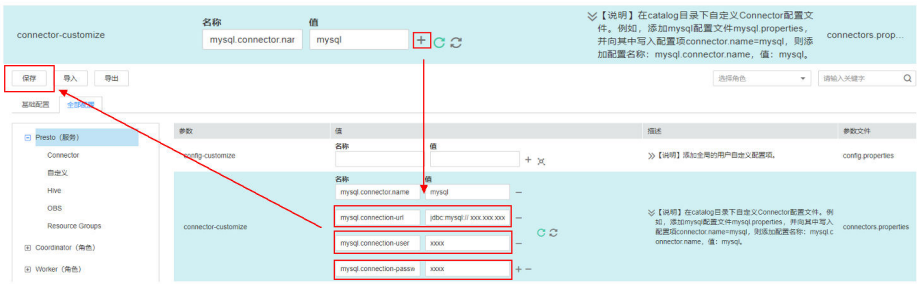
OpenTSDB基于HTTP提供了访问其的RESTful接口，而RESTful接口本身具有语言无关性的特点，凡是支持HTTP请求的语言都可以对接OpenTSDB，所以OpenTSDB支持python的接口。

15.11.28 Presto 如何配置其他数据源?

本指导以mysql为例。

- MRS 1.x及MRS 3.x版本。
  a. 登录MRS管理控制台。
  b. 单击集群名称进入集群详情页面。
  c. 选择“组件管理 > Presto”。设置“参数类别”为“全部配置”，进入Presto配置界面修改参数配置。
  d. 搜索“connector-customize”配置。
  e. 按照配置项说明填写对应参数。
  名称：mysql.connector.name
  值：mysql

  f. 填写connector-customize参数名称和参数值。



名称	值	参数说明
mysql.connection-url	jdbc:mysql:// xxx.xxx.xxx.xxx:3306	数据库连接池
mysql.connection-user	xxxx	数据库登录用户名
mysql.connection-password	xxxx	数据库密码

  g. 重启Presto服务。

h. 启用Kerberos认证的集群，执行以下命令连接本集群的Presto Server。

presto_cli.sh --*krb5-config-path* {krb5.conf文件路径} --*krb5-principal* {用户principal} --*krb5-keytab-path* {user.keytab文件路径} --*user* {presto用户名}

i. 登录Presto后执行**show catalogs**命令，确认可以查询Presto的数据源列表mysql。

```
[root@node-master2uoHG bin]# ./presto_cli.sh
--server http://15[redacted]20
        show catalogs;
 Catalog
---------
 hive
 jmx
 mysql
 system
 tpcds
 tpch
(6 rows)

Query 20220422_121338_00002_ra2vb, FINISHED, 3 nodes
Splits: 53 total, 53 done (100.00%)
0:00 [0 rows, 0B] [0 rows/s, 0B/s]
```

执行**show schemas from mysql**命令即可查询mysql数据库。

- MRS 2.x版本。

a. 创建mysql.properties配置文件，内容如下：

connector.name=mysql

connection-url=jdbc:mysql://mysqlIp:3306

connection-user=用户名

connection-password=密码

说明

- mysqlIp为mysql实例ip，需要和mrs网络互通 。
- 用户名和密码为登录mysql的用户名和密码。

b. 分别上传配置文件到master节点（Coordinator实例所在节点）的/opt/Bigdata/MRS_Current/1_14_Coordinator/etc/catalog/和core节点的/opt/Bigdata/MRS_Current/1_14_Worker/etc/catalog/目录下（路径以集群实际路径为准），文件属组改为omm:wheel。

c. 重启Presto服务。

15.11.29 MRS 如何连接 spark-shell

1. 用root用户登录集群Master节点。
2. 配置环境变量。

   source /opt/client/bigdata_env
3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如：

   - 开发用户为“机机”用户时请执行：kinit -kt user.keytab sparkuser
   - 开发用户为“人机”用户时请执行：kinit sparkuser
4. 执行如下命令连接Spark组件的客户端。

   spark-shell

15.11.30 MRS 如何连接 spark-beeline

1. 用root用户登录集群Master节点。
2. 配置环境变量。

   source /opt/client/bigdata_env
3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如：
   - 开发用户为“机机”用户时请执行：kinit -kt user.keytab sparkuser
   - 开发用户为“人机”用户时请执行：kinit sparkuser
4. 执行如下命令连接Spark组件的客户端。

   spark-beeline
5. 在spark-beeline中执行命令，例如在obs://mrs-word001/table/目录中创建表test。

   create table test(id int) location 'obs://mrs-word001/table/';
6. 执行如下命令查询所有表，返回结果中存在表test，即表示访问OBS成功。

   show tables;

   图 15-5 Spark 验证返回已创建的表名

   ```
   0: jdbc:hive2://ha-cluster/default> create table test(id int) location 'obs://mrs-word001/table/';
   +---------+--+
   | Result  |
   +---------+--+
   +---------+--+
   No rows selected (2.515 seconds)
   0: jdbc:hive2://ha-cluster/default> show tables;
   +-----------+------------+--------------+--+
   | database  | tableName  | isTemporary  |
   +-----------+------------+--------------+--+
   | default   | test       | false        |
   | default   | test_obs   | false        |
   +-----------+------------+--------------+--+
   2 rows selected (0.127 seconds)
   ```
7. 使用“Ctrl + C”退出spark beeline。

15.11.31 spark job 对应的执行日志保存在哪里?

- spark job没有完成的任务日志保存在Core节点的/srv/BigData/hadoop/data1/nm/containerlogs/
- spark job完成的任务日志保存在HDFS的/tmp/logs/*用户名*/logs

15.11.32 MRS 的 Storm 集群提交任务时如何指定日志路径?

客户可以根据自己的需求，修改MRS的流式Core节点上的/opt/Bigdata/MRS_XXX / 1_XX _Supervisor/etc/worker.xml文件，将标签filename的值设定为客户需要的路径，然后在Manager页面重启对应实例。

建议客户尽量不要修改MRS默认的日志配置，可能会造成日志系统异常。

15.11.33 Yarn 的 ResourceManager 配置是否正常?

步骤1 登录MRS Manager页面，选择“服务管理 > Yarn > 实例”。

步骤2 分别单击两个ResourceManager名称，选择“更多 > 同步配置”，并选择不勾选“重启配置过期的服务或实例。”。

步骤3 单击“是”进行配置同步。

步骤4 以**root**用户分别登录Master节点。

步骤5 执行**cd /opt/Bigdata/MRS_Current/*_*_ResourceManager/etc_UPDATED/**命令进入**etc_UPDATED**目录。

步骤6 执行**grep '\.queues' capacity-scheduler.xml -A2**找到配置的所有队列，并检查队列和Manager页面上看到的队列是否一一对应。

root-default在Manager页面隐藏，在页面看不到属于正常现象。

```
[omm@node-master1IlZA etc]$
[omm@node-master1IlZA etc]$ grep '\.queues' capacity-scheduler.xml  -A2
<name>yarn.scheduler.capacity.root.queues</name>
<value>default,root-default,launcher-job,test1,test2,test3,test4</value>
</property>
[omm@node-master1IlZA etc]$
[omm@node-master1IlZA etc]$
```

步骤7 执行**grep '\.capacity</name>' capacity-scheduler.xml -A2**找出各队列配置的值，检查每个队列配置的值是否和Manager上看到的一致。并检查所有队列配置的值的总和是否是100。

- 是，则说明配置正常。
- 否，则说明配置异常，请执行后续步骤修复。

```
[omm@node-master1IlZA etc]$
[omm@node-master1IlZA etc]$ grep '\.capacity</name>' capacity-scheduler.xml  -A2
<name>yarn.scheduler.capacity.root.root-default.accessible-node-labels.zhaolu.capacity</name>
<value>0.0</value>
</property>
--
<name>yarn.scheduler.capacity.root.launcher-job.capacity</name>
<value>10</value>
</property>
--
<name>yarn.scheduler.capacity.root.accessible-node-labels.zhaolu.capacity</name>
<value>100</value>
</property>
--
<name>yarn.scheduler.capacity.root.test1.capacity</name>
<value>10</value>
</property>
--
<name>yarn.scheduler.capacity.root.test2.capacity</name>
<value>10</value>
</property>
--
<name>yarn.scheduler.capacity.root.test3.capacity</name>
<value>10</value>
</property>
--
<name>yarn.scheduler.capacity.root.capacity</name>
<value>100</value>
</property>
--
<name>yarn.scheduler.capacity.root.root-default.capacity</name>
<value>40.0</value>
</property>
--
<name>yarn.scheduler.capacity.root.test4.accessible-node-labels.zhaolu.capacity</name>
<value>100</value>
</property>
--
<name>yarn.scheduler.capacity.root.test4.capacity</name>
<value>0</value>
</property>
--
<name>yarn.scheduler.capacity.root.default.capacity</name>
<value>20</value>
</property>
[omm@node-master1IlZA etc]$
```

步骤8 登录MRS Manager页面，选择“主机管理”。

步骤9 查找主Master节点，主机名称前带实心五角星的Master节点即为主Master节点。

步骤10 以**root**用户登录主Master节点。

步骤11 执行**su - omm**切换到omm用户。

步骤12 执行**sh /opt/Bigdata/om-0.0.1/sbin/restart-controller.sh**重启Controller。

请在Manager页面没有其他操作后重启Controller，重启Controller对大数据组件业务无影响。

步骤13 重新执行步骤1~步骤7同步ResourceManager的配置并检查配置是否正常。

配置同步完成后Manager页面可能显示配置过期，该显示不影响业务，是由于组件没有加载最新的配置，待后续组件重启的时会自动加载。

----结束

15.11.34 如何修改 Clickhouse 服务的 allow_drop_detached 配置项?

步骤1 用root用户登录Clickhouse客户端所在节点。

步骤2 进入客户端目录，配置环境变量。

cd /opt/*客户端安装目录*

source bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

说明

该用户必须具有Clickhouse管理员权限。

步骤4 执行命令**clickhouse client --host** *192.168.42.90* **--secure -m**，其中192.168.42.90为ClickHouseServer实例节点IP，执行结果如下：

```
[root@server-2110082001-0017 hadoopclient]# clickhouse client --host 192.168.42.90 --secure -m
ClickHouse client version 21.3.4.25.
Connecting to 192.168.42.90:21427.
Connected to ClickHouse server version 21.3.4 revision 54447.
```

步骤5 执行命令修改allow_drop_detached的值。

例如：设置allow_drop_detached=1

set allow_drop_detached=1;

步骤6 执行如下命令查看allow_drop_detached的值：

SELECT * FROM system.settings WHERE name = 'allow_drop_detached';

```
server-2110081635-0001 :) SELECT *  FROM system.settings  WHERE name = 'allow_drop_detached';

SELECT *
FROM system.settings
WHERE name = 'allow_drop_detached'

Query id: 8211d1ff-5717-49af-929f-8e4170c6e1d1

┌─name──────────────────┬─value─┬─changed─┬─description──────────────────────────────────────────┬─min──┬─max──┬─readonly─┬─type─┐
│ allow_drop_detached │ 1     │       1 │ Allow ALTER TABLE ... DROP DETACHED PART[ITION] ... queries │ NULL │ NULL │        0 │ Bool │
└───────────────────────┴───────┴─────────┴──────────────────────────────────────────────────────┴──────┴──────┴──────────┴──────┘

1 rows in set. Elapsed: 0.004 sec.
```

步骤7 执行命令**q;**退出clickhouse client。

----结束

15.11.35 执行 Spark 任务报内存不足告警

问题现象

执行Spark任务就会报内存不足告警，告警id：18022，可用内存会陡降到0。

处理步骤

在SQL脚本前设置executor参数，限制executor的核数和内存。

例如设置如下：

```
set hive.execution.engine=spark;
set spark.executor.cores=2;
set spark.executor.memory=4G;
set spark.executor.instances=10;
```

参数值大小请根据实际业务情况调整。

15.11.36 ClickHouse 占用大量 CPU，一直不下降

问题现象

客户使用ClickHouse，执行了大量的update操作。ClickHouse集群使用此操作会比较占用资源，而且如果失败了会不断重试，大量的失败语句在不断重试导致占用大量的CPU。

处理步骤

在ZooKeeper中把存在的数据删除掉，然后释放掉update语句。

15.11.37 ClickHouse 如何开启 Map 类型?

步骤1 使用**root**用户登录主Master节点 。

步骤2 修改“/opt/Bigdata/components/current/ClickHouse/configurations.xml”配置文件，开启用户参数自定义：

vim /opt/Bigdata/components/current/ClickHouse/configurations.xml

修改“hidden”为“advanced”保存退出，如下加粗部分：

```
<property type="hidden" scope="all" classification="Customization"
classdesc="RESID_CLICKHOUSE_CONF_0056">
       <name>_clickhouse.custom_content.key</name>
       <value>_user-xml-content</value>
</property>
<property type="advanced" scope="all" classification="Customization"
classdesc="RESID_CLICKHOUSE_CONF_0056">
      <name>_user-xml-content</name>
      <value vType="text" checker="clickhouse.xmlformat">&lt;yandex&gt;&lt;/yandex&gt;</value>
      <description>RESID_CLICKHOUSE_CONF_0025</description>
</property>
```

步骤3 切换为omm用户，重启controller服务。

su - omm

sh /opt/Bigdata/om-server/om/sbin/restart-controller.sh

步骤4 登录FusionInsight Manager页面，选择“集群 > 服务 > ClickHouse > 配置 > 全部配置 > ClickHouseServer (角色) > 自定义”，在“_user-xml-content”配置项中添加如下内容：

```
<yandex>
  <profiles>
    <default>
   <allow_experimental_map_type>1</allow_experimental_map_type>
    </default>
  </profiles>
</yandex>
```

步骤5 单击“保存”，保存配置。

步骤6 选择“集群 > 服务 > ClickHouse”，单击右上角的“更多> 重启服务”，重启Clickhouse服务。

----结束

15.11.38 SparkSQL 访问 hive 分区表大量调用 OBS 接口

问题背景

使用SparkSql访问hive的一个数据存放于OBS的一个分区表，但是运行速度却很慢，并且会大量调用OBS的查询接口。

SQL样例：

```
select a,b,c from test where b=xxx
```

原因分析

按照设定，任务应该只扫描b=xxx的分区，但是查看任务日志可以发现，实际上任务却扫描了所有的分区再来计算b=xxx的数据，因此任务计算的很慢。并且因为需要扫描所有文件，会有大量的OBS请求发送。

MRS默认开启基于分区统计信息的执行计划优化，相当于自动执行Analyze Table（默认开启的设置方法为spark.sql.statistics.fallBackToHdfs=true，可通过配置为false关闭）。开启后，SQL执行过程中会扫描表的分区统计信息，并作为执行计划中的代价估算，例如对于代价评估中识别的小表，会广播小表放在内存中广播到各个节点上，进行join操作，大大节省shuffle时间。 此开关对于Join场景有较大的性能优化，但是会带来OBS调用量的增加。

处理步骤

在SparkSQL中设置以下参数后再运行：

```
set spark.sql.statistics.fallBackToHdfs=false;
```

或者在启动之前使用--conf设置这个值为false：

```
--conf spark.sql.statistics.fallBackToHdfs=false
```

15.12 API 使用类

15.12.1 使用调整集群节点接口时参数 node_id 如何配置?

使用调整集群节点接口时，参数node_id的值固定为node_orderadd，直接填固定值即可。

15.13 集群管理类

15.13.1 如何查看所有集群?

MRS所有的集群都展示在“集群列表”页面中，进入“集群列表”页面，可查看所有集群。集群数量较多时，可采用翻页显示，您可以查看任何状态下的集群。

- 现有集群：包括除了“失败”和“已删除”状态以外的所有集群。
- 历史集群：仅包含“已删除”状态的集群，目前界面只显示6个月内创建且已删除的集群，若需要查看6个月以前删除的集群，请联系技术支持人员。
- 失败任务管理：仅包含“失败”状态的任务。
  - 集群创建失败的任务
  - 集群删除失败的任务
  - 集群扩容失败的任务
  - 集群缩容失败的任务

15.13.2 如何查看日志信息?

“操作日志”页面记录了用户对集群和作业的操作的日志信息。目前，MRS界面记录的日志信息分为以下几类：

- 集群操作
  - 创建集群、删除集群、扩容集群和缩容集群等操作
  - 创建目录、删除目录和删除文件等操作
- 作业操作：创建作业、停止作业和删除作业等操作
- 数据操作：IAM用户任务、新增用户、新增用户组等操作

记录用户操作的日志信息如图15-6所示：

图 15-6 日志信息

操作类型	操作IP	操作内容	时间
集群操作	10.63.167.82	创建id为0bb2a919-666d-40c0-8cb1-a3486431aae6，名字为bigdata_xq318的集群	2016-03-18 17:17:46
集群操作	10.57.99.128	删除id为e92e5dc7-34c1-449d-b353-3651853e7631,名字为bigdata_DVwu的集群	2016-03-10 16:45:24
作业操作	10.63.167.82	提交作业,作业id:f591520b-c632-4f33-9d2f-063e942c93a2,作业名:distcp,集群Id:e92e5dc7-34c1-449d-b353-3651853e7631	2016-03-10 10:26:28
作业操作	10.63.167.82	提交作业,作业id:d8a58879-72d4-4ebb-84fb-0eca09b1c981,作业名:job_spark,集群Id:e92e5dc7-34c1-449d-b353-3651853e7631	2016-03-07 11:02:28
作业操作	10.63.167.82	提交作业,作业id:bab88cc1-df9e-4735-b6f8-db190f303295,作业名:mr_01,集群Id:e92e5dc7-34c1-449d-b353-3651853e7631	2016-03-07 10:52:37
作业操作	10.63.195.73	提交作业,作业id:f346875e-9bd9-42e1-a7ff-422133605b3d,作业名:sparkSql,集群Id:e92e5dc7-34c1-449d-b353-3651853e7631	2016-02-23 11:23:22
集群操作	10.63.195.73	创建id为e92e5dc7-34c1-449d-b353-3651853e7631，名字为bigdata_DVwu的集群	2016-02-23 11:05:24

15.13.3 如何查看集群配置信息?

- 集群创建完成后在MRS控制台单击集群名称进入集群基本信息页面，可以查看到集群的基本配置信息。其中，节点的实例规格和容量决定了该集群对数据的分析处理能力。节点实例规格越高，容量越大，集群运行速度越快，分析处理能力越强，相应的成本也越高。
- 在基本信息页面，单击“前往Manager”，跳转至MRS集群管理页面。用户可在集群管理页面查看和处理告警信息、修改集群配置等。

15.13.4 如何在 MRS 集群中安装 Kafka，Flume 组件?

已经创建的MRS 3.1.0及之前版本集群不支持安装组件。Kafka和Flume为流式集群的组件，如果要安装Kafka和Flume组件，则需要创建流式集群或者混合集群并选择该组件。

15.13.5 如何停止 MRS 集群?

如果想停止MRS集群，可以在“节点管理”页面，单击各个节点名称，进入“弹性云服务器”页面，选择“关机”即可。

15.13.6 MRS 支持数据盘扩容吗?

MRS支持数据盘扩容，建议选择业务量较低时进行数据扩容，云硬盘扩容以后，仅扩大了云硬盘的存储容量，还需要登录云服务器自行扩展分区和文件系统。MRS节点均使用公共镜像安装，支持“正在使用”状态云硬盘扩容。

15.13.7 现有集群如何增加组件?

已经创建的MRS 3.1.0及之前版本集群暂不支持组件的添加和移除，建议重新创建MRS集群并包含所需组件。

15.13.8 MRS 集群中安装的组件能否删除?

已经创建的MRS 3.1.0及之前版本集群中的组件不可以删除，如果不使用的话可以登录Manager页面在服务管理中找到对应的组件将其停止。

15.13.9 MRS 是否支持变更 MRS 集群节点?

MRS管理控制台不支持变更集群节点，也不建议用户在ECS管理控制台直接修改MRS集群节点。如果手动在ECS管理控制台对集群节点执行停止ECS、删除ECS、修改或重装ECS操作系统，以及修改ECS规格的操作，可能影响集群稳定运行。

如果您对MRS集群节点进行了上述操作，MRS会自动识别并直接删除发生变更的集群节点。您可以登录MRS管理控制台，通过扩容恢复已经删除的节点。请勿在扩容过程中对正在扩容的节点进行操作。

15.13.10 如何取消集群风险告警

1. 登录MRS服务控制台。
2. 单击集群名称进入集群详情页面。
3. 选择“告警管理 > 消息订阅规则”。
4. 在待修改的规则所在行的“操作”列单击“编辑”，在“订阅规则”中取消对应风险告警。

5. 单击“确定”完成修改。

15.13.11 为什么 MRS 集群显示的资源池内存小于实际集群内存?

在MRS集群中，MRS默认为Yarn服务分配集群内存的50%，用户从逻辑上对Yarn服务的节点按照资源池进行分区管理，所以集群中显示的资源池总内存仅有集群总内存的50%。

15.13.12 如何配置 knox 内存?

步骤1 以**root**用户登录集群Master节点。

步骤2 在Master节点执行如下命令打开gateway.sh文件。

su omm

vim /opt/knox/bin/gateway.sh

步骤3 将“APP_MEM_OPTS=""” 修改为 “APP_MEM_OPTS="-Xms256m -Xmx768m"”保存并退出文件。

步骤4 在Master节点执行如下命令重启knox进程。

sh /opt/knox/bin/gateway.sh stop

sh /opt/knox/bin/gateway.sh start

步骤5 在其他Master节点上重复执行如上步骤。

步骤6 执行**ps -ef |grep knox**命令可查看已设置的内存信息。

图 15-7 knox 内存

----结束

15.13.13 MRS 集群安装的 Python 版本是多少?

以**root**用户登录任意一个Master节点，然后执行**Python3**即可获取MRS集群安装的python版本。

15.13.14 如何查看各组件配置文件路径?

常用组件配置文件路径如下所示：

组件	配置文件目录
ClickHouse	*客户端安装路径*/ClickHouse/clickhouse/config
Flink	*客户端安装路径*/Flink/flink/conf
Flume	*Flume客户端安装目录*/fusioninsight-flume-*xxx*/conf
HBase	*客户端安装路径*/HBase/hbase/conf
HDFS	*客户端安装路径*/HDFS/hadoop/etc/hadoop

组件	配置文件目录
Hive	*客户端安装路径*/Hive/config
Hudi	*客户端安装路径*/Hudi/hudi/conf
Kafka	*客户端安装路径*/Kafka/kafka/config
Loader	• *客户端安装路径*/Loader/loader-tools-*xxx*/loader-tool/conf • *客户端安装路径*/Loader/loader-tools-*xxx*/schedule-tool/conf • *客户端安装路径*/Loader/loader-tools-*xxx*/shell-client/conf • *客户端安装路径*/Loader/loader-tools-*xxx*/sqoop-shell/conf
Oozie	*客户端安装路径*/Oozie/oozie-client-*xxx*/conf
Spark2x	*客户端安装路径*/Spark2x/spark/conf
Yarn	*客户端安装路径*/Yarn/config
ZooKeeper	*客户端安装路径*/Zookeeper/zookeeper/conf

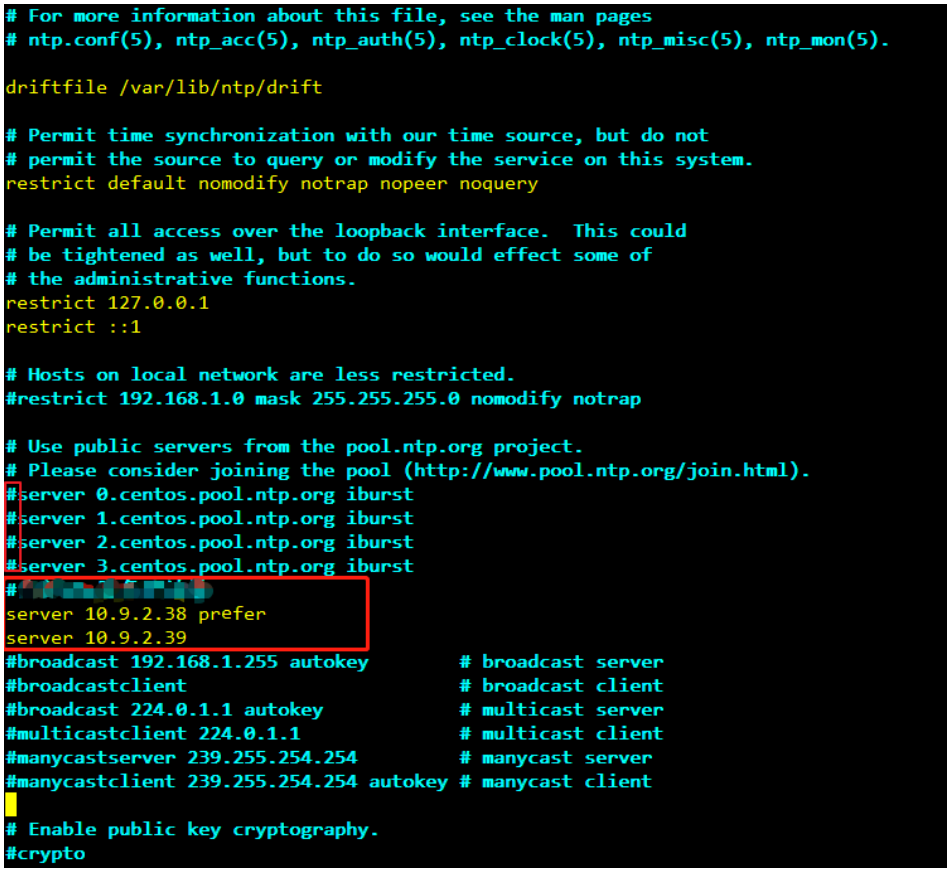
15.13.15 MRS 节点时间不正确

- 若集群内节点时间不正确，请分别登录集群内时间不正确的节点，并从步骤2开始执行。
- 若集群内节点与集群外节点时间不同步，请登录集群外节点，并从步骤1开始执行。

1. 执行**vi /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

```
server master1_ip prefer
server master2_ip
```

图 15-8 增加 Master 节点的 IP



2. 执行**service ntpd stop**命令关闭NTP服务。
3. 执行**/usr/sbin/ntpdate** *主Master节点的IP地址* 命令手动同步一次时间。
4. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。
5. 执行**ntpstat**命令查看时间同步结果。

15.13.16 如何查询 MRS 节点的启动时间

登录当前节点，执行如下命令查询节点启动时间：

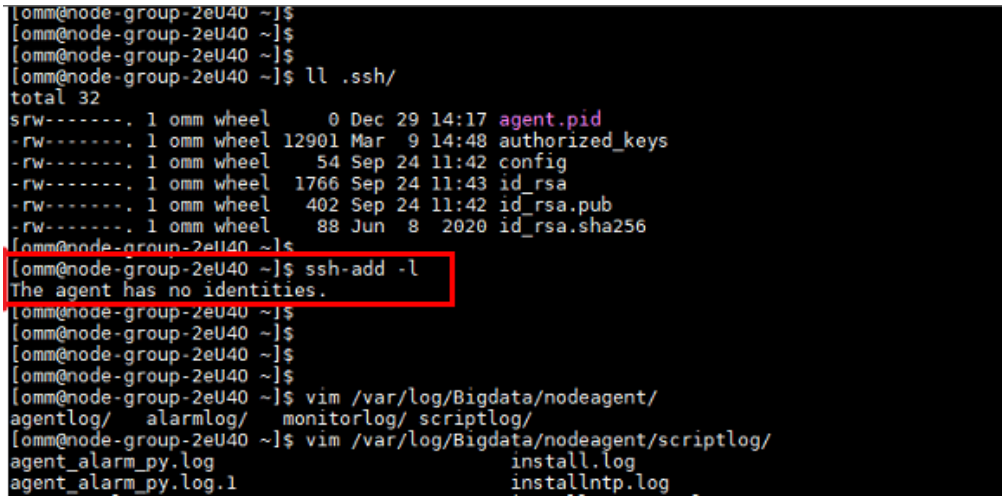
date -d "$(awk -F. '{print $1}' /proc/uptime) second ago" +"%Y-%m-%d %H:%M:%S"

```
[root@server-2110082001-0018 ~]#date -d "$(awk -F. '{print $1}' /proc/uptime) second ago" +"%Y-%m-%d %H:%M:%S"
2021-12-13 15:56:23
```

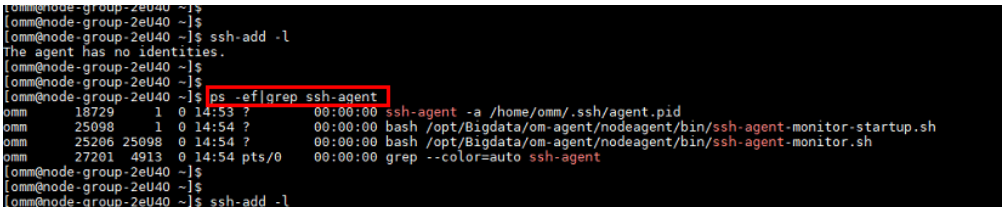
15.13.17 节点互信异常如何处理?

当Manager报“ALM-12066 节点间互信失效”告警，或者发现节点间无ssh互信时，可参考如下步骤操作。

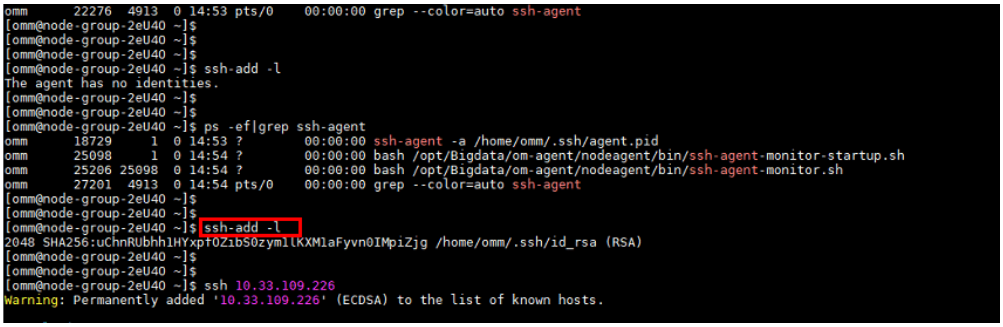
1. 分别在互信集群的两端节点执行**ssh-add -l** 确认是否有identities信息。



2. 如果没有identities信息，执行**ps -ef|grep ssh-agent**找到**ssh-agent**进程，并**kill**该进程等待该进程自动重启。



3. 执行**ssh-add -l** 查看是否已经添加identities信息，如果已经添加，请手动**ssh**确认互信是否正常。



4. 如果有identities信息，需要确认/home/omm/.ssh/authorized_keys中是否有对端节点/home/omm/.ssh/id_rsa.pub文件中的信息，如果没有手动添加对端节点信息。
5. 检查/home/omm/.ssh目录下的文件权限是否正确。
6. 排查日志文件 “/var/log/Bigdata/nodeagent/scriptlog/ssh-agent-monitor.log”，
7. 如果用户把omm的家目录删除了，需要联系MRS支撑人员修复。

15.13.18 如何调整 manager-executor 进程内存?

问题现象

MRS服务在集群的Master1和Master2节点上部署了manager-executor进程，该进程主要用于将管控面对集群的操作进行封装，比如作业的提交、心跳上报、部分告警信息上报、集群创扩缩等操作。当客户从MRS管控面提交作业，随着任务量的增大或者

任务并发较高，有可能会导致Executor内存不足，导致占用CPU较高、Executor进程OOM等问题。

解决步骤

1. 通过**root**用户登录Master1或Master2其中任意一节点，执行以下命令切换到**omm**用户。

   su - omm

2. 执行以下命令，修改“catalina.sh”脚本，搜索“JAVA_OPTS” ，找到类似如下的配置JAVA_OPTS=“-Xms1024m -Xmx4096m”,将其修改为需要的值大小，保存修改。

   vim /opt/executor/bin/catalina.sh

   ```
   JAVA_OPTS="-Xms1024m -Xmx4096m"
   JAVA_OPTS="$JAVA_OPTS $JSSE_OPTS"
   LOG4J_PROPERTIES_PATH="${CATALINA_HOME}/lib/log4j.properties"
   CATALINA_OPTS="-XX:+PrintGC -XX:+PrintGCDetails -XX:+PrintGCTimeStamps -XX:+PrintGCApplicationStoppedTime \
   -XX:+PrintHeapAtGC -Xloggc:/var/log/executor/logs/gc.log -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 \
   -XX:GCLogFileSize=20M -XX:OnOutOfMemoryError=\"kill -9 %p\" -XX:+HeapDumpOnOutOfMemoryError \
   -XX:HeapDumpPath=/var/log/executor/logs/executor-dump.hprof"
   ```

3. 重启manager-executor进程，该进程在Master1和Master2节点上为主备部署，同一时刻只有一个节点上运行该进程，首先确认节点上是否有该进程，如果有该进程则进行重启操作。

   a. 分别登录Master1节点和Master2节点，执行以下命令确认当前节点是否存在该进程。当有输出时则进程存在。

      ps -ef | grep "/opt/executor" | grep -v grep

   b. 重启命令如下。

      sh /opt/executor/bin/shutdown.shsh /opt/executor/bin/startup.sh

15.14 Kerberos 使用

15.14.1 已创建的 MRS 集群如何修改 Kerberos 状态?

MRS服务暂不支持集群创建完成后开启和关闭Kerberos服务，需要在创建MRS服务集群的时候选择开启或者关闭kerberos服务，建议重新创建集群。

15.14.2 Kerberos 认证服务的端口有哪些?

Kerberos认证服务的端口有21730(TCP)、21731(TCP/UDP)、21732(TCP/UDP)。

15.14.3 如何在运行中的集群中部署 Kerberos 服务?

MRS服务集群不支持客户自定义安装部署Kerberos服务，且部署的Kerberos认证无法在组件之间建立认证方式。故建议客户另外创建一套Kerberos开启的集群，然后进行数据迁移。

15.14.4 开启 Kerberos 认证的集群如何访问 Hive?

1. 用root用户登录集群Master节点。

2. 配置环境变量。

   source /opt/client/bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建Hive表的权限。

   kinit MRS集群用户

   例如，kinit hiveuser

4. 执行Hive组件的客户端命令。

   beeline

5. 在beeline中运行Hive命令，例如：

   create table test_obs(a int, b string) row format delimited fields terminated by "," stored as textfile location "obs://test_obs";

6. 使用“Ctrl + C”退出hive beeline。

15.14.5 开启 Kerberos 认证的集群如何访问 Presto?

1. 用root用户登录集群Master节点。
2. 配置环境变量。

   source /opt/client/bigdata_env

3. 开启Kerberos认证的安全集群访问Presto。
   a. 登录Manager创建一个拥有“Hive Admin Privilege”权限的角色，例如prestorole。
   b. 创建一个属于“Presto”和“Hive”组的用户，同时为该用户绑定3.a中创建的角色，例如presto001。
   c. 认证当前用户。

      kinit presto001

   d. 下载用户认证凭据。

      - MRS Manager界面操作：登录MRS Manager，选择“系统设置 > 用户管理 ”，单击新增用户所在行的“更多 > 下载认证凭据”。

        图 15-9 下载 Presto 用户认证凭据

      - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“系统 > 权限 > 用户”，单击新增用户所在行的“更多 > 下载认证凭据”。

   e. 解压下载的用户凭证文件，得到“krb5.conf”和“user.keytab”两个文件并放入客户端目录，例如“/opt/client/Presto/”。
   f. 执行如下命令获取用户principal。

      klist -kt /opt/client/Presto/user.keytab

g. 启用Kerberos认证的集群，执行以下命令连接本集群的Presto Server。

presto_cli.sh --krb5-config-path {krb5.conf文件路径} --krb5-principal {用户principal} --krb5-keytab-path {user.keytab文件路径} --user {presto用户名}

- krb5.conf文件路径：请替换为3.e中设置的文件存放路径，例如“/opt/client/Presto/krb5.conf”。
- user.keytab文件路径：请替换为3.e中设置的文件存放路径，例如“/opt/client/Presto/user.keytab”。
- 用户principal：请替换为3.f中返回的结果
- presto用户名：请替换为3.b中创建的用户名，例如“presto001”

例如：presto_cli.sh --krb5-config-path /opt/client/Presto/krb5.conf --krb5-principal prest001@xxx_xxx_xxx_xxx.COM --krb5-keytab-path /opt/client/Presto/user.keytab --user presto001

h. 在Presto客户端中执行语句创建schema，例如：

CREATE SCHEMA hive.demo01 WITH (location = 'obs://presto-demo002/');

i. 在该schema中建表，该表的数据即存储在OBS桶内，例如：

CREATE TABLE hive.demo01.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

图 15-10 安全集群 Presto 验证返回结果

j. 执行**exit**退出客户端。

15.14.6 开启 Kerberos 认证的集群如何访问 Spark?

1. 用**root**用户登录集群Master节点。
2. 配置环境变量。

   source /opt/client/bigdata_env
3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。

   kinit *MRS集群用户*

   例如：

   开发用户为“机机”用户时执行：**kinit -kt user.keytab sparkuser**

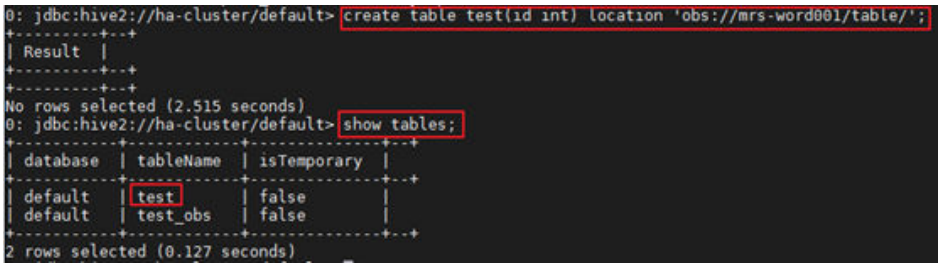
   开发用户为“人机”用户时执行：**kinit sparkuser**
4. 执行如下命令连接Spark组件客户端。

   spark-beeline
5. 在spark-beeline中执行命令，例如在obs://mrs-word001/table/目录中创建表test。

   create table test(id int) location 'obs://mrs-word001/table/';
6. 执行如下命令查询所有表，返回结果中存在表test，即表示访问OBS成功。

   show tables;

图 15-11 Spark 验证返回已创建的表名



7. 使用“Ctrl + C ”退出spark beeline。

15.14.7 如何避免 Kerberos 认证过期?

- 对于JAVA应用

  在连接HBase、HDFS或者其他大数据组件前，先调用loginUserFromKeytab()创建UGI，然后启动一个定时线程进行检查是否过期并在过期前重新登录。

```
private static void startCheckKeytabTgtAndReloginJob() {
    //10分钟循环 达到距离到期时间一定范围就会更新凭证
    ThreadPool.updateConfigThread.scheduleWithFixedDelay(() -> {
        try {
            UserGroupInformation.getLoginUser().checkTGTAndReloginFromKeytab();
            logger.warn("get tgt:{}", UserGroupInformation.getLoginUser().getTGT());
            logger.warn("Check Kerberos Tgt And Relogin From Keytab Finish.");
        } catch (IOException e) {
            logger.error("Check Kerberos Tgt And Relogin From Keytab Error", e);
        }
    }, 0, 10, TimeUnit.MINUTES);
    logger.warn("Start Check Keytab TGT And Relogin Job Success.");
}
```

- 对于shell方式执行的任务

  a. 先执行kinit命令认证用户。

  b. 通过操作系统定时任务或者其他定时任务方式定时执行kinit命令认证用户。

  c. 提交作业执行大数据任务。

- 对于Spark作业

  通过spark-shell、spark-submit、spark-sql方式提交作业，可以直接在命令行中指定Keytab和Principal以获取认证，定期更新登录凭证和授权tokens，避免认证过期, 例如：

  spark-shell --principal spark2x/hadoop.<系统域名>@<系统域名> --keytab ${BIGDATA_HOME}/FusionInsight_Spark2x_8.1.0.1/install/FusionInsight-Spark2x-2.4.5/keytab/spark2x/SparkResource/spark2x.keytab --master yarn

15.15 元数据管理

15.15.1 Hive 元数据在哪里查看?

- Hive的元数据存放在MRS服务集群的GaussDB中，可以登录到集群的DBServer主节点上并切换到omm用户，然后执行**gsql -p 20051 -U {USER} -W {PASSWD} -d hivemeta**查看。
- Hive元数据存放在外部的关系型数据库存储时，请通过如下步骤获取信息：

a. 集群详情页的“数据连接”右侧单击“单击管理”。
b. 在弹出页面中查看“数据连接ID”。
c. 在MRS控制台，单击“数据连接”。
d. 在数据连接列表中根据集群所关联的数据连接ID查找对应数据连接。
e. 在对应数据连接的“操作”列单击“编辑”，查看该数据连接所连接的RDS实例及数据库。

16 故障排除

16.1 Web 页面访问类

16.1.1 无法访问 MRS 集群管理页面（MRS Manager 界面）

问题现象

集群创建完成后，无法访问集群管理页面，即无法访问MRS Manager界面。

原因分析

- 需要对MRS节点绑定弹性IP才可访问
- 需要添加安全组规则，放开9022端口

处理步骤

步骤1 登录MRS的Console页面，在现有集群列表中找到需要访问的集群，单击集群名称。

步骤2 在节点信息中单击需要访问的节点名称，选择“弹性公网IP” > “绑定弹性公网IP”。

步骤3 在“绑定弹性公网”IP页面，“选择网卡”下拉框中选择需要绑定的网卡，“选择弹性公网IP”中选择需要绑定的弹性公网IP，单击“确定”。

步骤4 弹性IP绑定成功后，需要将安全组规则中9022端口放开。

选择“安全组”页签，单击“更改安全组”。

可以选择添加已有的安全组，或者单击“新建安全组”，进入安全组管理界面进行创建，添加用户访问公网IP地址9022端口的安全组规则。

步骤5 添加成功后，可通过*https://弹性ip:9022/mrsmanager/*访问MRS。如果还未能访问，请联系技术支持。

----结束

16.1.2 升级 Python 后，无法登录 MRS Manager 页面

用户问题

升级Python后，登录不进去MRS Manager页面。

问题现象

自行升级Python后，使用admin帐号且密码正确的情况下登录不进去MRS Manager页面。

原因分析

用户升级Python版本到Python3.x的过程中，修改了openssl的文件目录权限，导致LdapServer服务无法正常启动，从而引起登录认证失败。

处理步骤

步骤1 以root用户登录集群的Master节点。

步骤2 执行**chmod 755 /usr/bin/openssl**命令，修改**/usr/bin/openssl**的文件目录权限为755。

步骤3 执行**su omm**命令，切换到omm用户。

步骤4 执行**openssl**命令，查看是否能够进入openssl模式。

如果能够成功进入，则表示权限修改成功，如果不能进入，则表示权限未修改成功。

如果权限未修改成功，请检查执行的命令是否正确，或者联系运维人员。

步骤5 权限修改成功后会重启LdapServer服务，请等待LdapServer服务重启成功后，重新登录MRS Manager。

----结束

建议与总结

自行安装的软件建议和系统的分开，系统软件升级可能造成兼容性问题。

16.1.3 用户修改域名后无法登录 MRS Manager 页面

问题现象

用户修改域名后，通过console页面无法登录MRS Manager页面，或者登录MRS Manager页面异常。

问题原因

用户修改域名后，没有刷新executor用户的keytab文件，导致executor进程认证失败后不断循环认证，导致了acs进程内存溢出。

解决方案

步骤1 重启acs进程。

1. 使用root用户登录主管理节点（即MRS集群详情页面“节点管理”页签下实心五角星所在的Master节点）。
2. 执行如下命令重启进程：

   su - omm

   ps -ef|grep =acs（查找acs进程PID）

   kill -9 *PID*（PID替换为实际的ID，结束acs进程）
3. 等待几分钟后执行命令**ps -ef|grep =acs**查询进程是否已经自动启动。

步骤2 替换executor用户的keytab文件。

1. 登录MRS Manager页面，选择“系统 > 用户”，在executor用户所在的“操作”列，单击“下载认证凭据”，解压后获取keytab文件。
2. 使用root用户登录主管理节点，将获取的keytab替换“/opt/executor/webapps/executor/WEB-INF/classes/user.keytab”文件。

步骤3 替换knox用户的keytab和conf文件。

1. 登录MRS Manager页面，选择“系统 > 用户”，在knox用户所在的“操作”列，单击“下载认证凭据”，解压后获取keytab和conf文件。
2. 使用root用户登录主管理节点，将获取的keytab替换“/opt/knox/conf/user.keytab”文件。
3. 修改/opt/knox/conf/krb5JAASLogin.conf中的principal的值，把域名修改为更改后的域名。
4. 将获取的krb5.conf 替换“/opt/knox/conf/krb5.conf”文件。

步骤4 备份原有客户端目录

mv *{客户端目录}* /opt/client_init

步骤5 重新安装客户端。

步骤6 使用root用户登录主备管理节点，执行如下命令，重启knox进程。

su - omm

ps -ef | grep gateway | grep -v grep（查找knox进程PID）

kill -9 *PID*（PID替换为实际的ID，结束knox进程）

/opt/knox/bin/restart-knox.sh（启动knox进程）

步骤7 使用root用户登录主备管理节点，执行如下命令，重启executor进程。

su - omm

netstat -anp |grep 8181 |grep LISTEN（查找executor进程PID）

kill -9 *PID*（PID替换为实际的ID，结束executor进程）

/opt/executor/bin/startup.sh（启动executor进程）

----结束

16.1.4 登录 Manager，页面空白不显示

用户问题

登录到FusionInsight Manager界面后，页面空白不显示。

问题现象

登录到FusionInsight Manager界面后，页面空白不显示。

原因分析

Manager无法登录，需要清除浏览器缓存。

处理步骤

步骤1 切换至浏览器窗口（以Chrome为例），通过键盘按下“Ctrl+Shift+Delete”弹出“清除浏览数据”对话框。

步骤2 勾选待清除的浏览记录，单击“清除数据”，完成浏览器缓存清理。

----结束

16.1.5 用户名过长时下载认证凭据失败

用户问题

MRS 3.0.2~MRS 3.1.0版本集群，当用户名超过20位时（添加用户时最长限制为32位），下载Keytab文件会下载失败，状态代码：400 Bad Request。

问题现象

MRS 3.0.2~MRS 3.1.0版本集群，当用户名超过20位时（添加用户时最长限制为32位），下载Keytab文件会下载失败，状态代码：400 Bad Request。

原因分析

需要在主Master节点的“/opt/Bigdata/om-server_*/apache-tomcat-*/webapps/web/WEB-INF/validate”路径下，修改validate-common-config.xml、validate-rule-session.xml、 validate-rule-user.xml三个配置文件。

处理步骤

步骤1 以**omm**用户登录主Master节点的“/opt/Bigdata/om-server_*/apache-tomcat-*/webapps/web/WEB-INF/validate”路径。

cd /opt/Bigdata/om-server_*/apache-tomcat-*/webapps/web/WEB-INF/validate

步骤2 修改validate-common-config.xml文件。

vi validate-common-config.xml

将用户名的“maxLength”参数的值从“32”修改为“64”：

```
<!-- 用户名 -->
<validators alias="USER_NAME">
      <validator name="RANGE_LENGTH_VALIDATOR" minLength="3"
         maxLength="64" />
      <validator name="REGEXP_VALIDATOR" rule="^[_a-zA-Z0-9\- ]+$"
</validators>
```

步骤3 修改validate-rule-session.xml文件。

vi validate-rule-session.xml

将“下载当前用户凭据”的参数“rule”的值从“20”改为“64”：

```
 <!-- 下载当前用户凭据 -->
 <param_validator url="/api/v2/session/user/keytab/download" method="get"
errorHandler="com.xxx.bigdata.om.web.api.validate.SpecialValidatorErrorHandler" dataPattern="form">
            <!-- 参数名：文件名 -->
            <!--校验规则：userName_13位数字_keytab.tar；区分大小写-->
            <parameter name="file_name" required="true" errorKey="13-4000005"
errorMessage="RESID_OM_API_SESSION_0013">
                  <validator name="REGEXP_VALIDATOR" rule="[\-\w ]{3,64}_\d{13}_keytab\.tar"
caseSensitive="true" />
 </parameter>
```

步骤4 修改validate-rule-user.xml文件。

vi validate-rule-user.xml

将“下载当前用户凭据”的参数“rule”的值从“20”改为“64”：

```
<!--下载用户凭据  -->
   <param_validator url="/api/v2/permission/users/keytab/download" method="get"
errorHandler="com.xxx.bigdata.om.web.api.validate.SpecialValidatorErrorHandler" dataPattern="form">
          <!--必输；userName_13位数字_keytab.tar；区分大小写-->
          <parameter name="file_name" required="true" errorKey="12-4000005"
errorMessage="RESID_OM_API_AUTHORITY_0005">
                <validator name="REGEXP_VALIDATOR" rule="[\-\w ]{3,64}_\d{13}_keytab\.tar"
caseSensitive="true" />
          </parameter>
    </param_validator>
```

步骤5 重启Tomcat，并等待启动成功。

1. 以**omm**用户执行以下命令，查询出Tomcat进程的PID号。
   ps -ef|grep apache-tomcat
2. 使用**kill -9** *PID*命令强制停止查询出来的Tomcat进程，例如**:**
   kill -9 1203
3. 执行以下命令进行重启。
   sh ${BIGDATA_HOME}/om-server/tomcat/bin/startup.sh

步骤6 重新下载认证凭据。

----结束

16.2 集群管理类

16.2.1 缩容 Task 节点失败

用户问题

客户在MRS 2.x集群详情界面执行调整集群，将Task节点调整成0个，最终缩容失败。

问题现象

客户在MRS集群详情页面调整集群Task节点，最终缩容失败，提示“This operation is not allowed because the number of instances of NodeManager will be less than the minimum configuration after scale-in, which may cause data loss.”

原因分析

客户将Core节点的NodeManager服务停止了，导致在检查Task节点退服过程中发现Task如果全部退订，则将没有NodeManager，则Yarn服务就不可用，而MRS判断剩余的NodeManger必须大于等于1才能退服Task节点。

处理步骤

步骤1 勾选Core节点的NodeManager实例，选择“更多 > 启动实例”。

步骤2 在集群列表页面缩容Task节点。

1. 单击集群名称进入集群详情页面，选择“节点管理”。
2. 在Task节点组所在行的“操作”列单击“缩容”。
3. 单击“确定”并在弹出框选择“是”。

步骤3 等缩容成功后，若不想用Core节点的NodeManager再将其停止。

----结束

建议与总结

Core节点的NodeManager通常不会将其停止，客户不要随意变更集群部署结构。

16.2.2 MRS 集群添加新磁盘

用户问题

MRS HBase服务不可用。

问题现象

用户主机的磁盘占用率过高导致服务故障。

原因分析

Core节点的磁盘容量不足导致无法提供正常服务。

处理步骤

步骤1 购买云硬盘。

步骤2 挂载云硬盘。

- 若挂载云硬盘完成，请执行**步骤6**。
- 若在云硬盘控制台执行 "挂载" 操作时无法选定云服务器，请执行**步骤3**。

步骤3 登录弹性云服务器控制台，单击待扩容（挂载新磁盘）的弹性云服务器名称。

步骤4 在“云硬盘”页签，单击“挂载磁盘”。

步骤5 选择待挂载的新磁盘并单击“确定”完成磁盘挂载。

步骤6 初始化Linux数据盘。

说明

- 挂载点目录根据节点DataNode已有的实例编号递增，例如：使用df –h命令查到当前已有的编号为/srv/BigData/hadoop/data1，则新增挂载点为/srv/BigData/hadoop/data2。初始化Linux数据盘新建挂载点时，将新建挂载点命名为/srv/BigData/hadoop/data2，并将新建分区挂载到该挂载点下。例如
  ```
  mkdir /srv/BigData/hadoop/data2
  mount /dev/xvdb1 /srv/BigData/hadoop/data2
  ```

/srv/BigData/hadoop/data2路径说明：本章节后续提到/srv/BigData/hadoop/data2路径均请按照以下场景自行修改。

- 3.X版本目录为：/srv/BigData/data2
- 3.X之前版本目录为：/srv/BigData/hadoop/data2

步骤7 执行以下命令为新磁盘增加omm用户权限。

chown omm:wheel *新增挂载点*

例如：**chown omm:wheel /srv/BigData/hadoop/data2**

步骤8 执行chmod 701命令为新增的挂载点目录添加执行权限。

chmod 701 *新增挂载点*

例如：**chmod 701 /srv/BigData/hadoop/data2**

说明

chmod 701命令中701仅为示例，请以已有数据盘data1的数值为准。

步骤9 登录Manager，扩容DataNode实例和NodeManager实例的数据磁盘。

步骤10 修改当前节点DataNode实例配置。

MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > HDFS > 实例 > 扩容的DataNode节点 > 实例配置”，“参数类别”选择 "全部配置"。

FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > 扩容的DataNode节点 > 实例配置”，选择 "全部配置"。

- 方式一：手动修改当前节点DataNode实例配置。
  - 在 "搜索" 中输入"dfs.datanode.fsdataset.volume.choosing.policy"，将参数值改为 "org.apache.hadoop.hdfs.server.datanode.fsdataset.AvailableSpaceVolumeChoosingPolicy"。
  - 在 "搜索" 中输入 "dfs.datanode.data.dir"，将参数值改为"/srv/BigData/hadoop/data1/dn,/srv/BigData/hadoop/data2/dn"

  若此两个参数有修改，则单击 "保存配置"，并勾选 "重启角色实例" ，重启DataNode实例。
- 方式二：自动同步当前节点DataNode实例配置。

a. 单击“同步配置”为HDFS服务启用新的配置参数。
b. 完成同步配置后，请重启实例以使配置生效。

说明

- 如果确认当前未使用HDFS，并且希望较快完成重启，可以选择直接"重启角色实例" 。
- 如果有任务在使用HDFS，为了防止数据异常或者任务失败，必须选择滚动重启。

步骤11 修改当前节点Yarn NodeManager的实例配置。

MRS Manager界面操作入口：登录MRS Manager，依次选择 "服务管理 > Yarn > 实例 > 扩容节点的NodeManager > 实例配置"，"参数类别" 选择 "全部配置"。

FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”单击扩容节点的NodeManager，选择“实例配置 > 全部配置”。

- 方式一：手动修改当前节点Yarn NodeManager的实例配置。
  - 在“搜索”中输入 "yarn.nodemanager.local-dirs"，将参数值修改为："/srv/BigData/hadoop/data1/nm/localdir,/srv/BigData/hadoop/data2/nm/localdir" 。
  - 在“搜索”中输入 "yarn.nodemanager.log-dirs"，将参数值修改为："/srv/BigData/hadoop/data1/nm/containerlogs,/srv/BigData/hadoop/data2/nm/containerlogs"。

    若此两个参数有修改，则保存配置，并勾选 "重启角色实例" ，重启NodeManager实例。
- 方式二：自动同步当前节点Yarn NodeManager的实例配置。
  a. 单击“同步配置”为Yarn服务启用新的配置参数。
  b. 完成同步配置后，请重启实例以使配置生效。

说明

- 如果确认当前未使用Yarn，并且希望较快完成重启，可以选择直接"重启角色实例" 。
- 如果有任务在使用Yarn，为了防止数据异常或者任务失败，必须选择滚动重启。

步骤12 查看扩容是否成功。

MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > HDFS > 实例 > 扩容的DataNode节点 "。

FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，单击扩容的DataNode节点。

在图表区域，查看实时监控项 "DataNode存储" 中 配置的总磁盘容量是否提升，若图表区域没有监控项 "DataNode存储" ，请单击“定制”增加该监控项。

- 若配置的总磁盘容量已提升，则扩容完成。
- 若配置的总磁盘容量未提升，请联系技术支持处理。

步骤13 （可选）扩容Kafka实例的数据盘。

修改当前节点Kafka实例配置。

1. 进入Kafka扩容的Broker节点参数配置界面。

MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka > 实例 > 扩容的Broker节点 > 实例配置"，"参数类别" 选择 "全部配置"。

FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例> 扩容的Broker节点 > 实例配置”， 选择 “全部配置”。

2. 在 "搜索" 中输入"log.dirs"，加入新增磁盘信息，中间用英文 “,” 分割。

   例如原始只有一块Kafka数据盘，新增一块，则将"/srv/BigData/kafka/data1/kafka-logs" 改为 "/srv/BigData/kafka/data1/kafka-logs,/srv/BigData/kafka/data2/kafka-logs"。

3. 保存配置，并勾选 "重启角色实例" 后根据提示重启实例。
4. 查看扩容是否成功。

   MRS Manager界面操作入口：登录MRS Manager，依次选择 "服务管理 > Kafka > 实例 > 扩容的Broker节点 "。

   FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例> 扩容的Broker节点 ”。

   查看实时监控项 "Broker磁盘容量大小" 中配置的总磁盘容量是否提升。

----结束

须知

集群的节点扩容磁盘之后，若再扩容集群节点时需要在新扩容的节点上参考该页面处理步骤执行添加磁盘的操作，否则会有数据丢失的风险 。

建议与总结

- 当磁盘的使用率超过85%时，建议用户进行磁盘扩容，并将新购买的磁盘挂载到弹性云服务器上与集群进行关联。
- 具体挂载步骤、配置参数请根据实际情况进行。

16.2.3 MRS 集群更换磁盘（适用于 2.x 及之前）

用户问题

磁盘无法访问。

问题现象

客户创建本地盘系列MRS集群，其中1个Core节点的磁盘存在硬件损坏，导致读取文件失败。

原因分析

磁盘硬件故障。

处理步骤

> **说明**
>
> 该指导适用于MRS 3.x之前版本分析集群，如需为流式集群或混合集群更换磁盘，请联系技术支持处理。

步骤1 登录。

步骤2 选择“主机管理”并单击需要退服主机的“主机名称”，在“角色”列表中单击RegionServer，选择“更多 > 退服”。

步骤3 选择“主机管理”并单击需要退服主机的“主机名称”，在“角色”列表中单击DataNode，选择“更多 > 退服”。

步骤4 选择“主机管理”并单击需要退服主机的“主机名称”，在“角色”列表中单击NodeManager，选择“更多 > 退服”。

> **说明**
>
> 该主机下若还有其他实例，请参考该步骤方式进行退服。

步骤5 执行**vim /etc/fstab**命令编辑注释旧磁盘的挂载点。

图 16-1 注释旧磁盘的挂载点

```
[root@node-ana-coregexX0001 ~]# vim /etc/fstab

devpts   /dev/pts          devpts  mode=0620,gid=5 0 0
proc     /proc             proc    defaults        0 0
sysfs    /sys              sysfs   noauto          0 0
debugfs  /sys/kernel/debug debugfs noauto          0 0
tmpfs    /run              tmpfs   noauto          0 0
/dev/disk/by-label/ROOT / ext4 defaults,noatime 1 1
#UUID=0f871b41-61e0-4f7f-af54-a03a1bfb3753 /srv/BigData/hadoop/data1 ext4 defaults,noatime,nodiratime 1 0
~
~
```

步骤6 迁移旧磁盘上（例如：/srv/BigData/hadoop/data1/）的用户自有数据。

步骤7 登录MRS管理控制台。

步骤8 在集群详情页面，选择“节点管理”。

步骤9 单击待更换磁盘的“节点名称”进入弹性云服务器管理控制台，单击“关机”。

步骤10 联系支持人员在后台更换磁盘。

步骤11 在弹性云服务器管理控制台，单击“开机”，将已更换磁盘的节点开机。

步骤12 执行**fdisk -l**命令，查看新增磁盘。

步骤13 使用**cat /etc/fstab**获取盘符。

图 16-2 获取盘符

```
[omm@node-master1dGom ~]$ cat /etc/fstab

#
# /etc/fstab
# Created by anaconda on Wed Feb 27 06:58:49 2019
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
UUID=b13ee9c8-0ef0-4159-9b90-fc47bde0d464 /                       ext4    defaults,noatime        1 1
UUID=029408e0-71a6-4f73-b817-42d7049b7595 /srv/BigData ext4 defaults,noatime,nodiratime 1 0
UUID=f9cb8844-dabf-4a69-aff4-587de2fc4d7c /srv/BigData1 ext4 defaults,noatime,nodiratime 1 0
UUID=876e73be-1f80-4466-92b7-01d7c68bbb1b /srv/BigData2 ext4 defaults,noatime,nodiratime 1 0
UUID=0d5fce7f-afd0-420a-b1bb-e5500a1851cd /srv/BigData3 ext4 defaults,noatime,nodiratime 1 0
```

步骤14 使用对应的盘符对新磁盘进行格式化。

例如：**mkfs.ext4 /dev/sdh**

步骤15 执行如下命令挂载新磁盘。

mount *新磁盘 挂载点*

例如：**mount /dev/sdh /srv/BigData/hadoop/data1**

步骤16 执行如下命令为新磁盘增加omm用户权限。

chown omm:wheel *挂载点*

例如：**chown -R omm:wheel /srv/BigData/hadoop/data1**

步骤17 在fstab文件中新增新磁盘UUID信息。

1. 使用**blkid**命令查看新磁盘的UUID。

```
[root@node-ana-coreKpoT0003 ~]# blkid
/dev/vda1: LABEL="ROOT" UUID="2aa97872-11ec-422e-9513-0f28b925ad5e" TYPE="ext4"
/dev/vdb: UUID="e5f652c3-f9af-427f-89da-f2545618688d" TYPE="ext4"
[root@node-ana-coreKpoT0003 ~]#
```

2. 打开“/etc/fstab”文件，新增如下信息：
   UUID=*新盘UUID* /srv/BigData/hadoop/data1 ext4 defaults,noatime,nodiratime 1 0

步骤18 （可选）执行如下命令新建日志目录。

mkdir -p /srv/BigData/Bigdata

chown omm:ficommon /srv/BigData/Bigdata

chmod 770 /srv/BigData/Bigdata

说明

执行如下命令确认Bigdata日志软链接目录是否已存在，若存在则忽略该步骤。

ll /var/log

步骤19 登录。

步骤20 选择“主机管理”并单击需要入服主机的“主机名称”，在“角色”列表中单击RegionServer，选择“更多 > 入服”。

步骤21 选择“主机管理”并单击需要入服主机的“主机名称”，在“角色”列表中单击DataNode，选择“更多 > 入服”。

步骤22 选择“主机管理”并单击需要入服主机的“主机名称”，在“角色”列表中单击NodeManager，选择“更多 > 入服”。

说明

该主机下若还有其他实例，请参考该步骤方式进行入服。

步骤23 选择“服务管理 > HDFS”，在“服务状态”页签的“HDFS概述”模块查看“丢失块数”是否为“0”。

- “丢失块数”是为“0”，则操作完成。
- “丢失块数”不为“0”，请联系支持人员进行处理。

----结束

16.2.4 MRS 集群更换磁盘（适用于 3.x）

用户问题

磁盘无法访问。

问题现象

客户创建本地盘系列MRS集群，其中1个Core节点的磁盘存在硬件损坏，导致读取文件失败。

原因分析

磁盘硬件故障。

处理步骤

说明

该指导适用于本地盘系列（d/i/ir/ki系列）MRS集群，针对Core、Task类型节点的磁盘存在硬件故障。

Kafka组件不支持更换磁盘，如果存储Kafka数据的节点故障，请联系技术支持处理。

步骤1 登录。

步骤2 选择“主机”并单击故障主机的“主机名称”，在“实例”列表中单击DataNode，选择“更多 > 退服”。

说明

- 该主机下若存在DataNode、NodeManager、RegionServer和ClickHouseServer实例，请参考该步骤进行退服操作；
- MRS 3.1.2版本之后支持退服ClickHouseServer角色实例。

步骤3 选择“主机”并勾选故障主机“主机名称”前的复选框，选择“更多 > 停止所有实例”。

步骤4 执行**vim /etc/fstab**命令编辑注释旧磁盘的挂载点。

图 16-3 注释旧磁盘的挂载点

```
[root@node-ana-coreXZYb0001 ~]# vim /etc/fstab


#
# /etc/fstab
# Created by anaconda on Sat Feb 27 07:10:42 2021
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
UUID=c89eca08-5da4-43de-add0-4bb58e820d78 /                       ext4    defaults,errors=panic,noatime        1 1
UUID=4b16f96b-6d16-4dbe-9517-9f63423f9f6e /tmp  ext4   defaults,noatime,nodiratime,errors=panic 1 0
UUID=e539a0fd-a639-41dc-aa88-5fdc0e4bb7b3 /var ext4    defaults,noatime,nodiratime,errors=panic 1 0
UUID=51ba7a26-67de-4762-8bea-85fc004065c2 /srv/BigData ext4   defaults,noatime,nodiratime 1 0
UUID=03ba5f78-d188-4e6b-b640-1915b858183a /var/log  ext4   defaults,noatime,nodiratime,errors=panic 1 0
#UUID=91c84554-22eb-4130-a7a1-5ceaf03c8c06 /srv/BigData/data1 ext4 defaults,noatime,nodiratime,nodev 1 0
```

步骤5 如果旧磁盘仍可访问，迁移旧磁盘上（例如：/srv/BigData/data1/）的用户自有数据。

cp -r *旧磁盘挂载点 临时数据保存目录*

例如：**cp -r /srv/BigData/data1 /tmp/**

步骤6 登录MRS管理控制台。

步骤7 在集群详情页面，选择“节点管理”。

步骤8 单击待更换磁盘的“节点名称”进入弹性云服务器管理控制台，单击“关机”。

步骤9 联系支持人员在后台更换磁盘。

步骤10 在弹性云服务器管理控制台，单击“开机”，将已更换磁盘的节点开机。

步骤11 初始化Linux数据盘。

步骤12 执行**lsblk**命令，查看新增磁盘分区信息。

图 16-4 查看新增磁盘（分区）

```
[root@ecs-fcq ~]# lsblk
NAME   MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
sda      8:0    0  1.7T  0 disk
sdb      8:16   0  1.7T  0 disk
sdc      8:32   0  1.7T  0 disk
└─sdc1   8:33   0  1.7T  0 part
sdd      8:48   0  1.7T  0 disk
└─sdd1   8:49   0  1.7T  0 part
```

步骤13 使用**df -TH**获取文件系统类型。

图 16-5 获取文件系统类型

```
[root@node-ana-coreWQaI0001 ~]# df -TH
Filesystem     Type      Size  Used Avail Use% Mounted on
/dev/vda1      ext4      233G   44G  179G  20% /
devtmpfs       devtmpfs   34G     0   34G   0% /dev
tmpfs          tmpfs      34G     0   34G   0% /dev/shm
tmpfs          tmpfs      34G  9.3M   34G   1% /run
tmpfs          tmpfs      34G     0   34G   0% /sys/fs/cgroup
/dev/vda5      ext4       11G   40M   10G   1% /tmp
/dev/vda7      ext4       64G  152M   60G   1% /srv/BigData
/dev/vda6      ext4       11G  1.2G  8.9G  12% /var
/dev/vda8      ext4      190G  211M  180G   1% /var/log
/dev/sdc1      ext4      1.8T  1.4G  1.8T   1% /srv/BigData/data2
tmpfs          tmpfs     6.8G     0  6.8G   0% /run/user/2000
tmpfs          tmpfs     6.8G     0  6.8G   0% /run/user/0
[root@node-ana-coreWQaI0001 ~]#
```

步骤14 使用对应的文件系统类型对新磁盘（分区）进行格式化。

例如：**mkfs.ext4 /dev/sdd1**

步骤15 执行如下命令挂载新磁盘。

mount *新磁盘 挂载点*

例如：**mount /dev/sdd1 /srv/BigData/data1**

说明

如果挂载不上，请执行如下命令重载配置后重新挂载。

systemctl daemon-reload

步骤16 执行如下命令为新磁盘增加omm用户权限。

chown omm:wheel *挂载点*

例如：**chown -R omm:wheel /srv/BigData/data1**

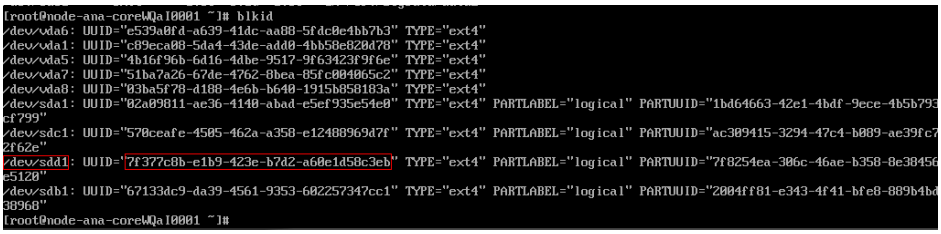
步骤17 将旧磁盘上（例如：/srv/BigData/data1/）的用户自有数据迁移到新磁盘上。

cp -r*临时数据保存目录 新磁盘挂载点*

例如：**cp -r /tmp/data1/* /srv/BigData/data1/**

步骤18 在fstab文件中新增新磁盘UUID信息。

1. 使用**blkid**命令查看新磁盘的UUID。



2. 打开“/etc/fstab”文件，新增如下信息：
   UUID=*新盘UUID* /srv/BigData/data1 ext4 defaults,noatime,nodiratime,nodev 1 0

步骤19 登录。

步骤20 选择“主机”并单击需要入服主机的“主机名称”，在“实例”列表中单击DataNode，选择“更多 > 入服”。

说明

- 该主机下若存在DataNode、NodeManager、RegionServer和ClickHouseServer实例，请参考该步骤进行入服操作；
- MRS 3.1.2版本之后支持入服ClickHouseServer角色实例。

步骤21 选择“主机”，并勾选故障主机“主机名称”前的复选框，选择“更多 > 启动所有实例”。

步骤22 选择“集群 > HDFS”，在“概览”页签的“基本信息”模块查看“丢失块数”是否为“0”。

- “丢失块数”是为“0”，则操作完成。
- “丢失块数”不为“0”，请联系支持人员进行处理。

----结束

16.2.5 MRS 备份失败

用户问题

MRS备份总是失败。

问题现象

MRS备份总是失败。

原因分析

备份目录软链接到系统盘，系统盘满了之后备份便会失败。

处理步骤

步骤1 检查备份目录是否软链接到系统盘。

1. 以root用户登录集群主备Master节点。
2. 执行**df -h**命令查看磁盘情况，检查系统盘的存储情况。
3. 执行 **ll /srv/BigData/LocalBackup**命令， 查看备份目录是否软连接到**/opt/Bigdata/LocalBackup**。

   检查备份文件是否软链接到系统盘且系统盘空间是否足够。如果软链接到系统盘且系统盘空间不足，请执行**步骤2**。如果否，说明不是由于系统盘空间不足导致，请联系技术服务。

步骤2 将历史备份数据移到数据盘的新目录中。

1. 以root用户登录Master节点。
2. 执行**su - omm**命令，切换到omm用户。
3. 执行**rm -rf /srv/BigData/LocalBackup**命令，删除备份目录软连接。
4. 执行**mkdir -p /srv/BigData/LocalBackup**命令，创建备份目录。
5. 执行**mv /opt/Bigdata/LocalBackup/* /srv/BigData/LocalBackup/**命令，将历史备份数据移到新目录。

----结束

16.2.6 Core 节点出现 df 显示的容量和 du 显示的容量不一致

用户问题

Core节点出现df显示的容量和du显示的容量不一致

问题现象

Core节点出现df显示的容量和du显示的容量不一致：

分别使用命令**df -h** 和命令**du -sh /srv/BigData/hadoop/data1/**查询得到的/srv/BigData/hadoop/data1/目录磁盘占用量相差较大（大于10G）。

原因分析

使用命令**lsof |grep deleted**可以查询到此目录下有大量log文件处于deleted状态。

出现此问题的一种情况是长时间运行某些spark任务，任务中的一些container一直运行，并且持续产生日志；spark的executor在打印日志的时候使用了log4j的日志滚动功能，将日志输出到stdout文件下；而container同时也会监控这个文件，导致此文件被两个进程同时监控。当其中一个进程按照配置滚动的时候，删除了最早的日志文件，但是另一个进程依旧占用此文件句柄。从而产生了deleted状态的文件。

处理步骤

将spark的executor日志输出目录修改成其他名称

1. 打开日志配置文件，默认在<客户端地址>/Spark/spark/conf/log4j-executor.properties。
2. 将日志输出文件改名，例如：

   log4j.appender.sparklog.File = ${spark.yarn.app.container.log.dir}/stdout改为：log4j.appender.sparklog.File = ${spark.yarn.app.container.log.dir}/stdout.log
3. 保存退出
4. 重新提交任务。

16.2.7 如何解除关联子网

操作场景

您可根据自身网络需求，解除网络ACL与子网关联。

操作步骤

步骤1 登录管理控制台。

步骤2 在系统首页，单击“网络 > 虚拟私有云”。

步骤3 在左侧导航栏单击“网络ACL”。

步骤4 在右侧在“网络ACL”列表区域，选择网络ACL的名称列，单击您需要修改的“网络ACL名称”进入网络ACL详情页面。

步骤5 在详情页面，单击“关联子网”页签。

步骤6 在“关联子网”页签详情区域，选择对应子网的“操作”列，单击“取消关联”。

步骤7 单击“确认”。

----结束

16.2.8 修改 hostname，导致 MRS 状态异常

用户问题

修改hostname后，MRS状态异常怎么处理?

问题现象

修改hostname，导致MRS状态异常。

原因分析

修改hostname导致兼容性问题和故障。

处理步骤

步骤1 以root用户登录集群的任意节点。

步骤2 在集群节点中执行**cat /etc/hosts**命令，查看各个节点的hostname值，根据此值来配置newhostname变量值 。

步骤3 在hostname被修改的节点上执行**sudo hostnamectl set-hostname ${newhostname}**命令，恢复正确的hostname。

说明

${newhostname}：表示新的hostname取值。

步骤4 修改完成后，重新登录修改hostname的节点，查看修改的hostname是否生效。

----结束

16.2.9 如何定位进程被 kill

问题背景与现象

在某环境出现DataNode异常重启，且确认此时未从页面做重启DataNode的操作，需要定位是什么进程kill DataNode服务端进程。

原因分析

常见的进程被异常终止有2种原因：

- **Java进程OOM被Kill**

  一般Java进程都会配置OOM Killer，当检测到OOM会自动Kill，OOM日志通常被打印到out日志中，此时可以看运行日志（如DataNode的日志路径为 /var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-主机名.log），看是否有 OutOfMemory 内存溢出的打印。

- **被其他进程kill，或者人为kill。**

  排查DataNode运行日志（/var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-主机名.log），是先收到“RECEIVED SIGNAL 15”再健康检查失败。即如下示例中DataNode先于 11:04:48被kill，然后过2分钟，于 11:06:52启动。

  ```
  2018-12-06 11:04：48,433 | ERROR | SIGTERM handler | RECEIVED SIGNAL 15: SIGTERM | LogAdapter.java:69
  2018-12-06 11:04:48,436 | INFO  | Thread-1 | SHUTDOWN_MSG:
  /************************************************************
  SHUTDOWN_MSG: Shutting down DataNode at 192-168-235-85/192.168.235.85
  ************************************************************/ | LogAdapter.java:45
  2018-12-06 11:06:52,744 | INFO  | main | STARTUP_MSG:
  ```

  以上日志说明，DataNode先被其他进程关闭，然后健康检查失败，2分钟后，被NodeAgent启动DataNode进程。

处理步骤

打开操作系统审计日志，给审计日志增加记录kill命令的规则，即可定位是何进程发送的kill命令。

操作影响

- 打印审计日志，会消耗一定操作系统性能，经过分析仅影响不到1%。
- 打印审计日志，会占用一定磁盘空间。该日志打印量不大，MB级别，且默认配置有老化机制和检测磁盘剩余空间机制，不会占满磁盘。

定位方法

在DataNode进程可能发生重启的所有节点，分别执行以下操作。

步骤1 以**root**用户登录节点，执行**service auditd status**命令，确认该服务状态。

```
Checking for service auditd  running
```

如果该服务未启动，执行**service auditd restart**命令重启该服务（无影响，耗时不到1秒）

```
Shutting down auditd done
Starting auditd done
```

步骤2 审计日志临时增加kill命令审计规则。

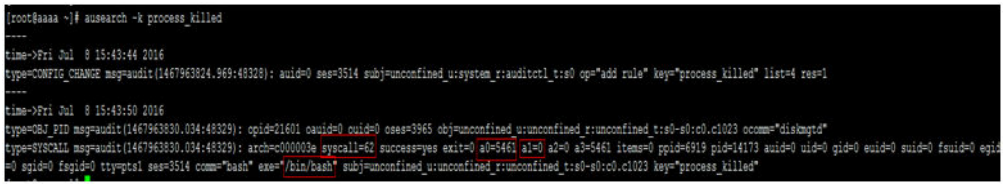
增加规则：

auditctl -a exit,always -F arch=b64 -S kill -S tkill -S tgkill -F a1!=0 -k process_killed

查看规则：

auditctl -l

步骤3 当进程有异常被kill后，使用**ausearch -k process_killed**命令，可以查询kill历史。



说明

a0是被kill进程的pid（16进制），a1是kill命令的信号量。

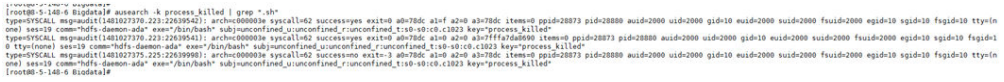
----结束

验证方法

步骤1 从MRS页面重启该节点一个实例，如DataNode。

步骤2 执行**ausearch -k process_killed**命令，确认是否有日志打印。

例如以下命令ausearch -k process_killed |grep “.sh”，可以看到是hdfs-daemon-ada* 脚本，关闭的DataNode进程。



----结束

停止审计kill命令方法

步骤1 执行**service auditd restart**命令，即会清理临时增加的kill审计日志。

步骤2 执行**auditctl -l**命令，如果没有kill相关信息，即说明已清理该规则。

----结束

16.2.10 MRS 集群使用 pip3 安装 python 包提示网络不可达

用户问题

使用pip3安装python包报错网络不可达。

问题现象

执行pip3 install 安装python包报错网络不可达。具体如下图所示：

原因分析

客户未给Master节点绑定弹性公网IP，造成报错的发生。

处理步骤

步骤1 登录MRS服务管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中当前安装出问题的集群并单击集群名称，进入集群基本信息页面。

步骤3 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。

步骤4 选择“弹性公网IP”页签，单击“绑定弹性公网IP”为弹性云服务器绑定一个弹性公网IP。

步骤5 登录Master节点执行pip3 install安装python包。

----结束

16.2.11 MRS 集群客户端无法下载

用户问题

在本地的Master主机上想给另外一台远端主机下载一个MRS集群客户端进行使用，但是一直提示网络或者参数有问题

问题现象

在本地的Master主机上想给另外一台远端主机下载一个MRS集群客户端进行使用，但是一直提示网络或者参数有问题

原因分析

- 可能是两台主机处于不同VPC网络中
- 密码填写错误
- 远端主机开启防火墙

处理步骤

- 两台主机处于不同VPC网络中
  放开远端主机的22端口
- 密码填写错误
  请检查密码是否正确，密码中不能有特殊符号。
- 远端主机开启防火墙
  使用规避方案，先将这个MRS集群客户端下载到“服务器端”主机，然后通过linux提供的scp命令远程发送到远端主机。

16.2.12 扩容失败

用户问题

Console界面正常，MRS集群扩容失败

问题现象

Console界面正常，查看MRS Manager界面也无警告无错误，但MRS集群无法扩容报“集群存在非运行状态节点，请稍后重试”的错误。

原因分析

MRS集群的扩缩容要建立在集群处于正常运行的基础上，所以首先要检查集群是否处于正常与否，现在报的是集群存在非运行状态节点，而console界面和MRS Manager界面都是正常的，所以可能原因就是数据库中集群状态不正常或未刷新导致集群相关节点处于非正常运行状态导致的。

处理步骤

步骤1 登录MRS控制台，单击集群名称进入集群详情页面查看集群状态，确保集群状态为“运行中”。

步骤2 单击“节点管理”，查看所有节点的状态，确保所有节点的状态为“运行中”。

步骤3 登录集群的podMaster节点跳转到MRS的deployer节点，查看api-gateway.log的日志。

1. 用**kubectl get pod -n mrs**命令查看MRS 对应的**deployer**节点的**pod**。
2. 用**kubectl exec -ti ${deployer节点的pod} -n mrs /bin/bash**命令登录相应的pod，如执行**kubectl exec -ti mrsdeployer-78bc8c76cf-mn9ss -n mrs /bin/bash**命令进入MRS的deployer容器。
3. 在/opt/cloud/logs/apigateway目录下查看最新的api-gateway.log日志，检索里面的关键信息（如：ERROR，scaling，clusterScaling，HostState，state-check，集群ID等)查看报错类型。
4. 根据报错提示信息进行相应处理，然后再次执行扩容操作。
   - 扩容成功，则处理完成。
   - 扩容失败，则执行**步骤4**。

步骤4 用**/opt/cloud/mysql -u${用户名} -P${端口} -h${地址} -p${密码}**登录数据库。

步骤5 执行**select cluster_state from cluster_detail where cluster_id="集群ID";**查看cluster_state。

- cluster_state为2，则集群状态正常，执行**步骤6**。
- cluster_state不为2，说明集群状态在数据库中异常，可用**update cluster_detail set cluster_state=2 where cluster_id="集群ID";**刷新集群状态，并查看cluster_state。
  - cluster_state为2，则集群状态正常，执行**步骤6**
  - cluster_state不为2，则请联系技术工程师处理。

步骤6 执行**select host_status from host where cluster_di="clusterID";**命令查询集群主机状态。

- 如果主机状态为started，则处理完成。
- 如果主机状态不为started，则可执行**update host set host_status='started' where cluster_id="集群ID";**命令更新主机状态到数据库。
  - 如果主机状态为started，则处理完成。
  - 如果主机状态不为started，则请联系技术工程师处理。

----结束

16.2.13 MRS 通过 beeline 执行插入命令的时候出错

用户问题

MRS通过beeline执行插入命令的时出错

问题现象

在hive的beeline中执行**insert into**插入语句的时候会报以下的错误:

```
Mapping run in Tez on Hive transactional table fails when data volume is high with error: "org.apache.hadoop.hive.ql.lockmgr.LockException Reason: Transaction... already aborted, Hive SQL state [42000]."
```

原因分析

对于Join操作，由于集群配置不理想和Tez资源设置不合理导致该问题。

处理步骤

可以在beeline上设置配置参数进行解决。

步骤1 设置以下属性以优化性能（建议在集群级别进行更改）

- 设置hive.auto.convert.sortmerge.join = true
- 设置hive.optimize.bucketmapjoin = true
- 设置hive.optimize.bucketmapjoin.sortedmerge = true

步骤2 更改以下内容以调整Tez的资源 。

- 设置hive.tez.container.size = {与YARN容器相同的大小}

- 将hive.tez.container.size设置为与YARN容器大小yarn.scheduler.minimum-allocation-mb相同或更小的值（如设置为二分之一或四分之一的值），但不要超过yarn.scheduler.maximum-allocation-mb 。

----结束

16.2.14 MRS 集群如何进行 Euleros 系统漏洞升级?

用户问题

Euleros系统底层存在漏洞，MRS集群如何进行漏洞升级?

问题现象

在使用绿盟软件测试集群，发现有Euleros系统底层存在漏洞，漏洞报告如下：

IP地址	操作系统	漏洞风险（个）				主机风险值
		高	中	低	合计	
[redacted]	Linux	7	13	21	41	10.0
	Linux	7	11	17	35	10.0
	Linux	6	11	9	26	8.9
	Linux	6	10	8	24	8.9
		6	11	9	26	8.9
	Linux	6	14	23	43	8.9
		6	13	10	29	8.9
		6	13	22	41	8.9
		6	10	9	25	8.9
	Linux	6	9	8	23	8.8
	Linux	6	9	7	22	8.8
		6	9	7	22	8.8
		6	9	8	23	8.8
	Linux	6	9	8	23	8.8
	Windows Server 2008 R2 Sta	1	138	28	167	8.3

原因分析

在使用绿盟软件测试集群，发现有Euleros系统底层存在漏洞，MRS服务部署在Euleros系统中，因此需要进行漏洞升级。

处理步骤

说明

修复漏洞前请确认是否开启了企业主机安全（Host Security Service，简称HSS）服务，如果已开启，需要先暂时关闭HSS服务对MRS集群的监测，漏洞修复完成后重新开启HSS服务。

步骤1 登录MRS服务控制台。

步骤2 单击集群名称进入集群详情页面，并选择“节点管理”。

步骤3 在Core节点组中勾选任意一个Core节点，单击“节点操作 > 停止所有角色”。

步骤4 通过远登录Core节点后台，配置yum源。

步骤5 使用**uname –r**或**rpm –qa |grep kernel**命令，查询并记录当前节点内核版本。

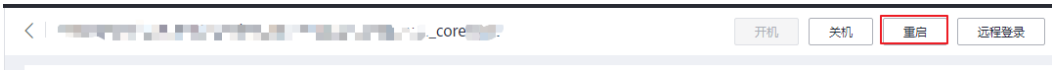
步骤6 执行**yum update –y --skip-broken --setopt=protected_multilib=false**命令更新补丁。

步骤7 完成更新后查询内核版本，并执行**rpm -e *旧内核版本*** 命令删除旧内核版本。

步骤8 在集群详情页，选择“节点管理”。

步骤9 在Core节点组中单击已更新补丁的Core名称，进入弹性云服务管理控制台。

步骤10 在页面右上角单击“重启”，重启Core节点。



步骤11 重启完成后，在集群详情页的“节点管理”的Core节点组中勾选Core节点，单击“节点操作 > 启动所有角色”。

步骤12 重复**步骤1**~**步骤11**的操作，升级其他Core节点。

步骤13 所有Core节点升级完成后，参考**步骤1**~**步骤11**的操作先升级备Master节点，再升级主Master节点。

----结束

16.2.15 使用 CDM 迁移数据至 HDFS

用户问题

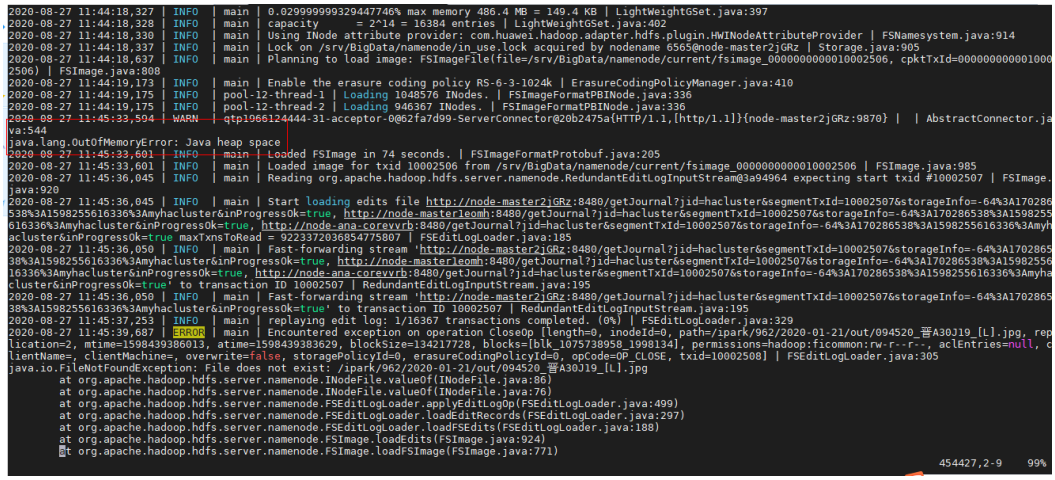
使用CDM从旧的集群迁移数据至新集群的HDFS过程失败。

问题现象

使用CDM从源HDFS导入目的端HDFS，发现目的端MRS集群故障，NameNode无法启动。

查看日志发现在启动过程中存在Java heap space报错，需要修改NN的JVM参数。

图 16-6 故障日志



原因分析

客户在使用CDM迁移数据的过程中，HDFS的数据量过大，导致在合并元数据时发生堆栈异常。

处理步骤

步骤1 搜索并修改“HDFS->NameNode” 中的“GC_OPTS”参数，将其中的“-Xms512M -Xmx512M”两个参数的值根据实际情况调整为较大的值。

步骤2 保存配置并重启受影响的服务或实例。

----结束

16.2.16 MRS 集群频繁产生告警

用户问题

集群频繁发出Manager主备节点间心跳中断，DBService主备节点间心跳中断，节点故障等告警，偶尔会造成hive不可用。

问题现象

集群频繁发出Manager主备节点间心跳中断，DBService主备节点间心跳中断，节点故障等告警，偶尔会造成hive不可用，影响客户业务。

原因分析

1. 在出现告警时间点发现虚拟机发生了重启，告警发生的原因是因虚拟机重启导致的。

```
[omm@node-master1yqIY nodeagent]$ last
omm      pts/0        100.125.0.70     Thu Sep 24 10:33   still logged in
omm      pts/1        100.125.0.70     Thu Sep 24 09:26 - 09:47  (00:20)
omm      pts/0        100.125.0.70     Thu Sep 24 09:22 - 10:21  (00:59)
omm      pts/1        100.125.0.70     Wed Sep 23 17:32 - 17:37  (00:05)
root     pts/0        10.203.216.102   Wed Sep 23 17:13 - 18:35  (01:21)
omm      pts/0        100.125.0.70     Wed Sep 23 16:55 - 16:56  (00:00)
omm      pts/0        100.125.0.70     Wed Sep 23 16:20 - 16:25  (00:05)
reboot   system boot  4.19.36-vhulk190 Wed Sep 23 16:06   still running
root     pts/1        10.203.216.102   Tue Sep 22 19:13 - 19:48  (00:34)
omm      pts/0        100.125.0.70     Tue Sep 22 19:08 - 20:03  (00:54)
root     pts/0        10.203.216.102   Tue Sep 22 17:03 - 17:52  (00:48)
omm      pts/1        100.125.0.70     Tue Sep 22 15:55 - 16:00  (00:05)
```

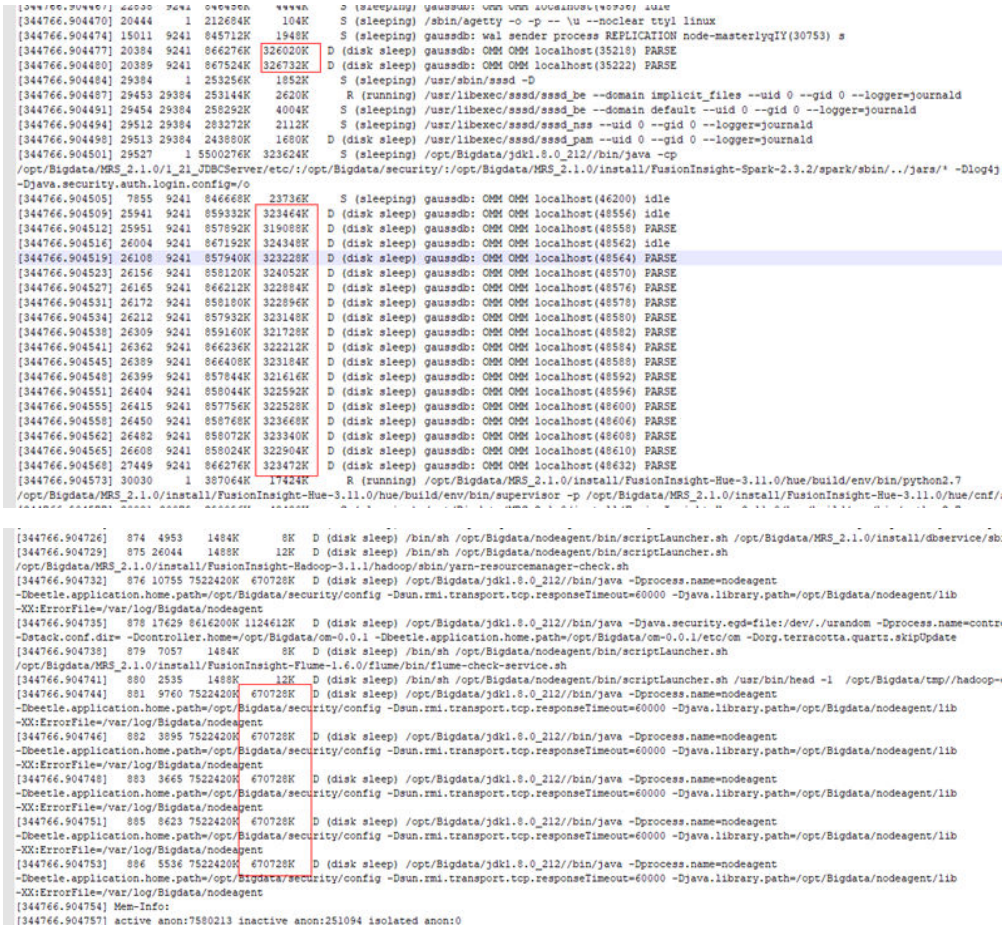
```
[omm@node-master2WbYp ~]$ last
omm      pts/0        10.80.0.56       Thu Sep 24 11:00   still logged in
omm      pts/0        10.80.0.56       Thu Sep 24 09:24 - 10:21  (00:56)
omm      pts/0        10.80.0.56       Wed Sep 23 17:32 - 17:37  (00:05)
omm      pts/0        10.80.0.56       Tue Sep 22 19:15 - 19:15  (00:00)
omm      pts/0        10.80.0.56       Tue Sep 22 15:57 - 16:21  (00:23)
omm      pts/0        10.80.0.56       Tue Sep 22 15:23 - 15:35  (00:12)
omm      pts/0        10.80.0.56       Tue Sep 22 15:07 - 15:12  (00:05)
omm      pts/0        10.80.0.56       Tue Sep 22 14:21 - 14:26  (00:05)
omm      pts/0        10.80.0.56       Mon Sep 21 10:57 - 11:06  (00:09)
omm      pts/0        10.80.0.56       Mon Sep 21 10:42 - 10:56  (00:14)
omm      pts/0        10.80.0.56       Thu Sep 17 16:05 - 16:15  (00:10)
omm      pts/0        10.80.0.56       Wed Sep 16 20:52 - 20:58  (00:06)
reboot   system boot  4.19.36-vhulk190 Wed Sep 16 18:05   still running
omm      pts/0        10.80.0.56       Wed Sep 16 15:43 - 16:10  (00:26)
omm      pts/0        10.80.0.56       Wed Sep 16 14:35 - 14:53  (00:17)
omm      pts/0        10.80.0.56       Wed Sep 16 14:33 - 14:33  (00:00)
omm      pts/0        10.80.0.56       Wed Sep 16 14:11 - 14:29  (00:17)
omm      pts/0        10.80.0.56       Wed Sep 16 14:02 - 14:09  (00:06)
omm      pts/0        10.80.0.56       Wed Sep 16 11:56 - 12:04  (00:08)
omm      pts/0        10.80.0.56       Wed Sep 16 11:26 - 11:31  (00:04)
omm      pts/0        10.80.0.56       Wed Sep 16 11:09 - 11:24  (00:15)
root     pts/0        10.203.230.193   Mon Sep 14 15:54 - 16:30  (00:35)
root     pts/0        10.203.172.29    Fri Sep 11 17:15 - 17:45  (00:30)
root     pts/0        10.203.172.29    Fri Sep 11 16:53 - 17:12  (00:19)
root     tty1                          Fri Sep 11 16:23 - 17:25  (01:01)
reboot   system boot  4.19.36-vhulk190 Fri Sep 11 10:07   still running
reboot   system boot  4.19.36-vhulk190 Thu Aug 27 16:41   still running
root     tty1                          Thu Aug 20 09:46 - 10:17  (00:30)
reboot   system boot  4.19.36-vhulk190 Wed Aug 19 17:48   still running
reboot   system boot  4.19.36-vhulk190 Wed Aug 19 17:46   still running
```

2. 经OS定位虚拟机发生重启的原因是节点没有可用的内存，系统发生内存溢出触发了oom-killer，当进程处于被调用的状态会使进程处于disk sleep状态，最终导致虚拟机发生重启。

```
mem info:
[344766.903734] MemTotal:       32397404 kB
MemFree:          160404 kB
MemAvailable:      31668 kB
Buffers:            2172 kB
Cached:          2768904 kB
SwapCached:            0 kB
Active:         30328872 kB
Inactive:        1035844 kB
Active(anon):   30320852 kB
Inactive(anon):  1004376 kB
Active(file):       8020 kB
Inactive(file):    31468 kB
Unevictable:           0 kB
Mlocked:               0 kB
[344766.903738] SwapTotal:             0 kB
SwapFree:              0 kB
```

总内存

用户态使用



3. 查看占用的内存进程，发现占用内存都是正常的业务进程。

结论：虚拟机内存不能满足服务需求。

处理步骤

- 建议扩大节点内存。
- 建议关闭不需要的服务来规避该问题。

16.2.17 PMS 进程占用内存高问题处理

用户问题

主Master节点内存使用率高如何处理?

问题现象

主Master节点内存使用率高，且用**top -c**命令查询得内存占用量高的是如下idle的进程。

```
12180 ommdba    20   0 1395492 1.180g 1.082g S   0.0  3.8  23:14.29 gaussdb: OMM OMM localhost(60598) idle
14828 ommdba    20   0 1395904 1.180g 1.081g S   0.0  3.8  23:17.08 gaussdb: OMM OMM localhost(60698) idle
15016 ommdba    20   0 1395840 1.180g 1.081g S   0.0  3.8  23:11.19 gaussdb: OMM OMM localhost(60824) idle
14943 ommdba    20   0 1395900 1.180g 1.081g S   0.0  3.8  23:14.76 gaussdb: OMM OMM localhost(60764) idle
14908 ommdba    20   0 1395840 1.180g 1.081g S   0.0  3.8  23:15.18 gaussdb: OMM OMM localhost(60738) idle
14953 ommdba    20   0 1395824 1.180g 1.081g S   0.0  3.8  23:15.96 gaussdb: OMM OMM localhost(60770) idle
14995 ommdba    20   0 1395560 1.180g 1.081g S   0.0  3.8  23:13.28 gaussdb: OMM OMM localhost(60812) idle
15062 ommdba    20   0 1395820 1.180g 1.081g S   0.0  3.8  23:16.12 gaussdb: OMM OMM localhost(60868) idle
15064 ommdba    20   0 1395512 1.180g 1.081g S   0.0  3.8  23:13.33 gaussdb: OMM OMM localhost(60870) idle
14973 ommdba    20   0 1395528 1.180g 1.081g S   0.0  3.8  23:12.74 gaussdb: OMM OMM localhost(60790) idle
14835 ommdba    20   0 1395536 1.180g 1.081g S   0.0  3.8  23:17.39 gaussdb: OMM OMM localhost(60704) idle
14822 ommdba    20   0 1395524 1.180g 1.081g S   0.0  3.8  23:13.80 gaussdb: OMM OMM localhost(60692) idle
14991 ommdba    20   0 1395808 1.180g 1.081g S   0.0  3.8  23:17.96 gaussdb: OMM OMM localhost(60808) idle
14975 ommdba    20   0 1395812 1.180g 1.081g S   0.0  3.8  23:12.57 gaussdb: OMM OMM localhost(60792) idle
15038 ommdba    20   0 1395520 1.180g 1.081g S   0.0  3.8  23:12.75 gaussdb: OMM OMM localhost(60846) idle
14919 ommdba    20   0 1395540 1.180g 1.081g S   0.0  3.8  23:11.58 gaussdb: OMM OMM localhost(60744) idle
14832 ommdba    20   0 1395476 1.180g 1.081g S   0.0  3.8  23:13.11 gaussdb: OMM OMM localhost(60702) idle
14989 ommdba    20   0 1395500 1.180g 1.081g S   0.0  3.8  23:15.63 gaussdb: OMM OMM localhost(60806) idle
14979 ommdba    20   0 1395448 1.180g 1.081g S   0.0  3.8  23:13.17 gaussdb: OMM OMM localhost(60796) idle
15047 ommdba    20   0 1395512 1.180g 1.081g S   0.0  3.8  23:12.10 gaussdb: OMM OMM localhost(60854) idle
14977 ommdba    20   0 1395496 1.180g 1.081g S   0.0  3.8  23:16.90 gaussdb: OMM OMM localhost(60794) idle
15028 ommdba    20   0 1395800 1.180g 1.081g S   0.0  3.8  23:09.35 gaussdb: OMM OMM localhost(60836) idle
```

原因分析

- PostgreSQL缓存：除了常见的执行计划缓存、数据缓存，PostgreSQL为了提高生成执行计划的效率，还提供了catalog，relation等缓存机制。长连接场景下这些缓存中的某些缓存是不会主动释放的，因此可能导致长连接占用大量的内存不释放。
- PMS是MRS的监控进程，此进程会经常创建表分区或者新表，由于PostgreSQL会缓存当前会话访问过的对象的元数据，且PMS的数据库连接池连接会长时间存在，所以连接占用的内存会逐渐上升。

处理步骤

步骤1 以**root**用户登录主Master节点。

步骤2 执行如下命令查询PMS进程号。

ps -ef | grep =pmsd |grep -v grep

步骤3 执行如下命令关闭PMS进程，其中**PID**为步骤2中获取的PMS进程号。

kill -9 PID

步骤4 等待PMS进程自动启动。

PMS启动需要2-3分钟。PMS是监控进程，重启不影响大数据业务。

----结束

16.2.18 Knox 进程占用内存高

用户问题

knox进程占用内存高

问题现象

主Master节点内存使用率高，用**top -c**命令查看到占用内存较高的进程中有knox进程，且此进程占用内存超过4G。

原因分析

knox进程没有单独配置内存，进程会自动根据系统内存大小按照比例划分可用内存，导致knox占用内存大。

处理步骤

步骤1 以root用户分别登录Master节点。

步骤2 打开文件“/opt/knox/bin/gateway.sh”，查找APP_MEM_OPTS，并设置该参数的值为：“-Xms3072m -Xmx4096m”。

步骤3 登录Manager页面，单击“主机管理”，找到主Master节点的IP（即主机名称前带有实心五角星的节点），并登录该节点后台。

步骤4 执行如下命令重启进程。

su - omm

sh /opt/knox/bin/restart-knox.sh

----结束

16.2.19 安全集群外节点安装客户端访问 HBase 很慢

用户问题

安全集群外节点安装了集群的客户端，并使用客户端命令hbase shell访问hbase，发现访问hbase非常慢。

问题现象

客户创建了安全集群，在集群外节点安装了集群的客户端，并使用客户端命令hbase shell访问hbase，发现访问hbase非常慢。

原因分析

安全集群需要进行Kerberos认证，需要在客户端节点的hosts中配置信息，访问速度才不会收到影响。例如，hosts配置信息为：

```
1.1.1.1 hadoop.782670e3_1364_47e2_8c70_1b61bb80479c.com
1.1.1.1 hadoop.hadoop.com
1.1.1.1 hacluster
1.1.1.1 haclusterX
1.1.1.1 haclusterX1
1.1.1.1 haclusterX2
1.1.1.1 haclusterX3
1.1.1.1 haclusterX4
1.1.1.1 ClusterX
1.1.1.1 manager
ip1 hostname1
ip2 hostname2
ip3 hostname3
ip4 hostname4
```

处理步骤

将集群节点上的hosts文件内容复制到安装客户端节点的hosts文件中。

16.2.20 作业无法提交如何定位?

问题背景与现象

客户通过DGC或者在MRS Console无法提交作业。

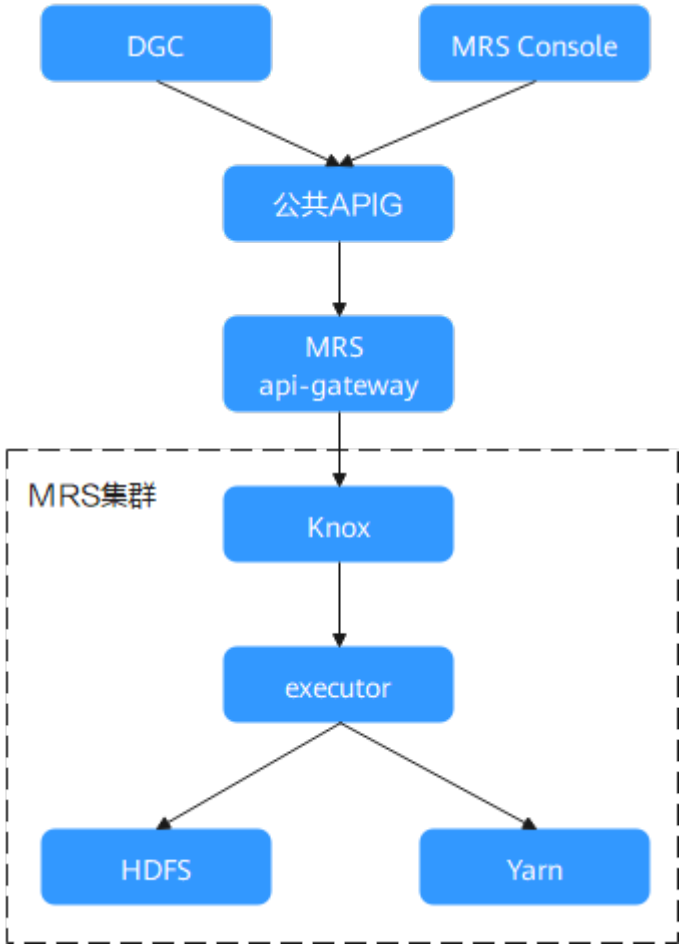
问题影响

作业无法提交，业务中断。

作业流程简介

1. 所有请求会先经过APIG网关，受到APIG配置的流控限制。
2. APIG将请求转发到MRS管控面的api-gateway中。
3. MRS管控面API节点轮询主备oms的Knox，确认主oms的Knox。
4. MRS管控面API提交任务到主oms的Knox。
5. Knox转发请求到本节点的executor进程。
6. executor进程提交任务到Yarn。

图 16-7 作业流程



处理步骤

前期准备：

- 确定作业是通过DGC或在MRS Console提交。
- 准备如表16-1信息。

表 16-1 修复前准备事项

序号	项目	操作方式
1	集群帐号信息	申请集群**admin**账户的密码。
2	节点帐号信息	申请集群内节点的**omm**、**root**用户密码。
3	SSH远程登录工具	准备PuTTY或SecureCRT等工具。
4	客户端	已提前安装好客户端。

步骤1 确认异常来源。

查看作业日志中收到的错误码，确认错误码是属于APIG还是MRS。

- 若是公共APIG的错误码（APIG的错误码是APIGW开头），联系公共APIG维护人员。
- 若是MRS侧错误，继续下一步。

步骤2 排查服务和进程运行状态等基本情况。

1. 登录Manager界面确认是否有服务故障，如果有作业相关服务故障或者底层基础服务故障，需要解决故障。
2. 查看是否有严重告警。
3. 登录主Master节点。
4. 执行如下命令查看oms状态是否正常，主oms节点executor和knox进程是否正常。knox是双主模式，executor是单主模式。

   /opt/Bigdata/om-0.0.1/sbin/status-oms.sh
5. 以omm用户执行**jmap -heap PID**检查knox和executor进程内存使用情况，如果多次执行查看到老生代内存使用率为99.9%说明有内存溢出。

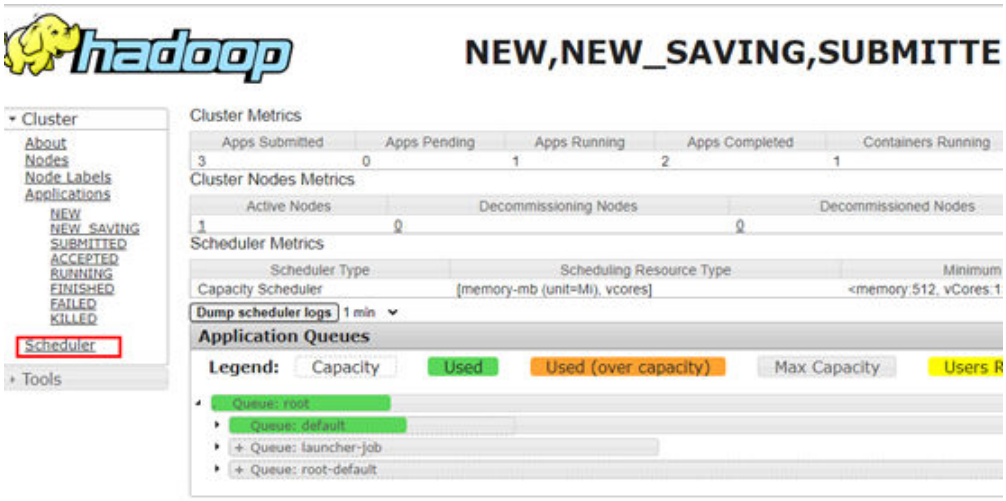
   查询executor进程PID：netstat -anp | grep 8181 | grep LISTEN

   查询knox进程PID：ps -ef|grep knox | grep -v grep

   如果内存溢出，需要现在执行**jmap -dump:format=b,file=/home/omm/temp.bin PID**，导出内存信息后重启进程进行恢复。
6. 查看Yarn的原生界面，确认队列资源情况，以及任务是否提交到了yarn上。

   Yarn的原生界面：在集群详情页选择“组件管理 > Yarn > ResourceManager WebUI > ResourceManager (主)”

图 16-8 Yarn 界面队列资源情况



步骤3 排查任务提交失败点。

1. 登录MRS控制台，单击集群名称进入集群详情页面。
2. 选择“作业管理”页签，单击作业所在行“操作”列的“查看日志”。

图 16-9 作业日志

3. 若没有日志或者日志信息不详细，则在“作业名称/ID”列复制作业ID。
4. 在主oms节点执行如下命令确认任务请求是否下发到了knox，如果请求没有到knox则可能是knox出了问题，需要尝试重启knox进行恢复。

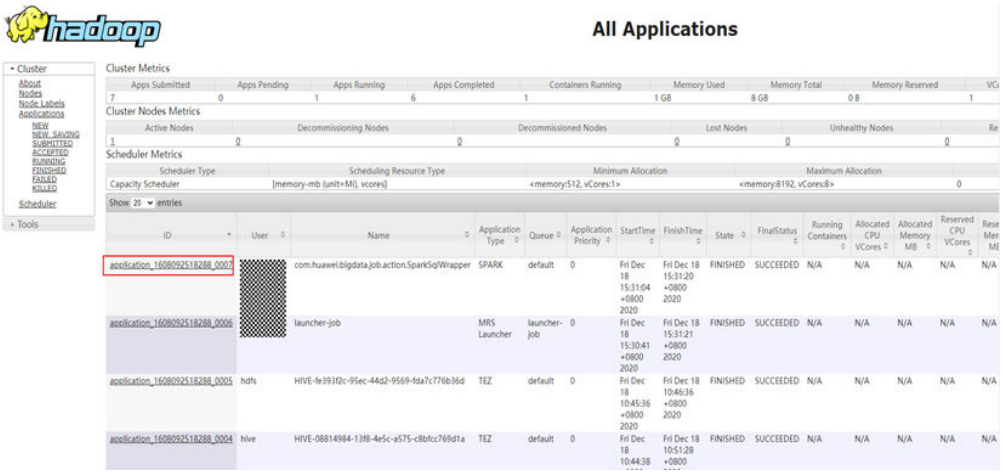
   grep "mrsjob" /var/log/Bigdata/knox/logs/gateway-audit.log | tail -10
5. 进入executor的日志中搜索作业ID，查看报错信息。

   日志路径：/var/log/Bigdata/executor/logs/exe.log
6. 修改“/opt/executor/webapps/executor/WEB-INF/classes/log4j.properties”文件开启executor的debug日志，提交测试任务，查看executor的日志并确认作业提交过程中的报错。

   日志路径：/var/log/Bigdata/executor/logs/exe.log
7. 如果当前任务在exeutor中出错，执行如下命令打印executor的jstack信息，确认线程当前执行状态。

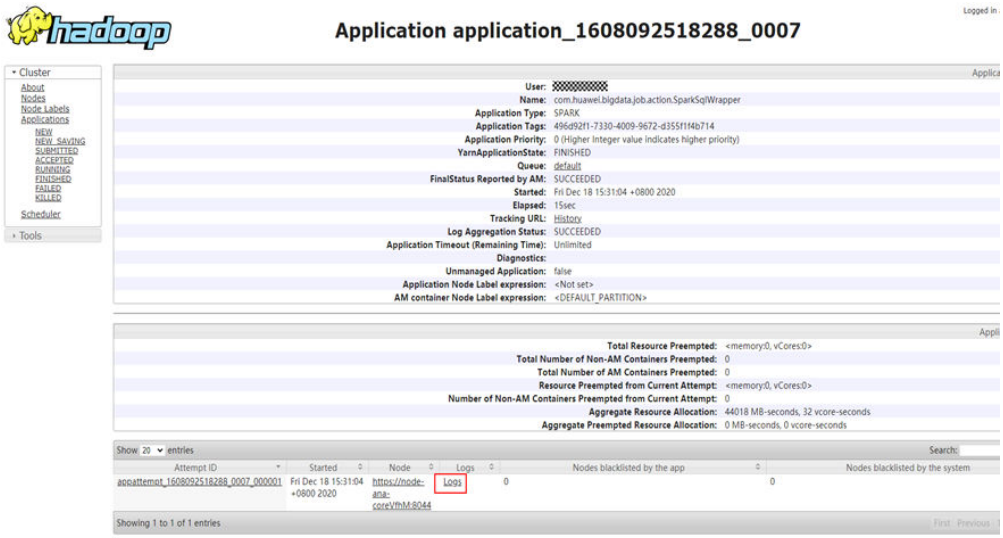
   jstack PID > xxx.log
8. 在集群详情页面选择“作业管理”页签，单击作业所在行“操作”列的“查看详情”，获取“实际作业编号”applicationID。
9. 在集群详情页选择“组件管理 > Yarn > ResourceManager WebUI > ResourceManager (主)”进去Yarn的原生界面，单击applicationID。

图 16-10 Yarn 的 Applications



10. 在任务详情页面查看日志。

图 16-11 任务日志



----结束

16.2.21 HBase 日志文件过大导致 OS 盘空间不足

用户问题

OS盘/var/log分区空间不足。

问题现象

“/var/log/Bigdata/hbase/*/hbase-omm-*.out”日志文件过大，造成OS盘/var/log分区空间不足。

原因分析

在HBase长时间运行场景下，操作系统会把JVM创建的“/tmp/.java_pid*”文件定期清理。因为HBase的内存监控使用了JVM的jinfo命令，而jinfo依赖“/tmp/.java_pid*”文

件，当该文件不存在时，jinfo会执行**kill -3**将jstack信息打印到.out日志文件里，从而导致.out日志文件过大。

处理步骤

在每个HBase实例的节点上部署定期清理.out日志文件的定时任务。后台登录HBase的实例节点，在crontab -e中添加每天0点清理.out日志的定时任务。

crontab -e

00 00 * * * for file in `ls /var/log/Bigdata/hbase/*/hbase-omm-*.out`; do echo "" > $file; done

说明

如果.out大文件出现比较频繁，可以每天清理多次或者调整操作系统的自动清理策略。

16.2.22 Manager 页面新建的租户删除失败

问题现象

在FusionInsight Manager的“租户资源”页面添加租户后，删除租户时，报“删除租户角色失败”。

原因分析

在创建租户时会生成对应的角色，执行删除租户操作时会首先删除对应的角色。此时如果支持权限配置的组件状态异常，则会导致删除这个角色对应的资源权限失败。

处理步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，在“配置资源权限”中单击集群名称，确认可配置资源权限的组件。

步骤3 选择“集群 > 服务”，查看可配置资源权限的组件的运行状态是否都为“良好”。

步骤4 如果不为“良好”，请启动或者修复组件，直至状态为“良好”。

步骤5 再次执行删除租户操作。

----结束

16.3 使用 Alluxo

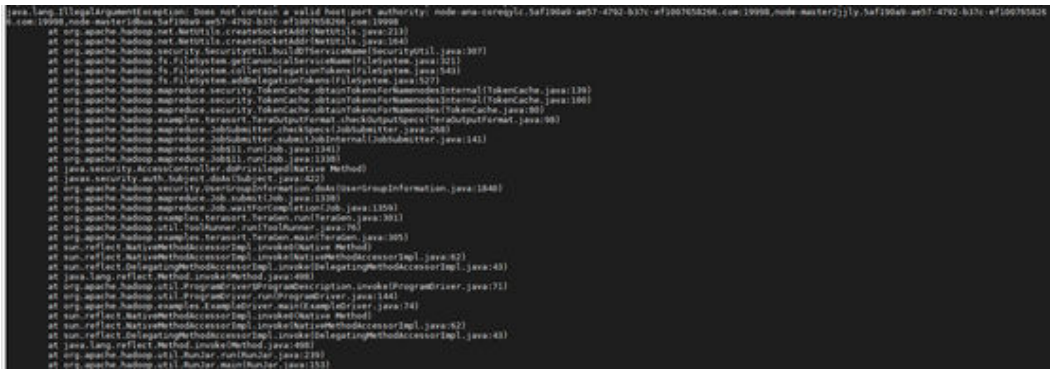
16.3.1 Alluxo 在 HA 模式下出现 Does not contain a valid host:port authority 报错

用户问题

安全集群Alluxo在HA模式下出现Does not contain a valid host:port authority的报错，如何处理?

问题现象

安全集群中，Alluixo在HA模式下出现Does not contain a valid host:port authority的报错。



原因分析

org.apache.hadoop.security.SecurityUtil.buildDTServiceName不支持在uri中填写多个alluxiomaster的地址。

处理步骤

使用alluxio:///或者alluxio://<主AlluxioMaster的ip或hostname>:19998/进行访问。

16.4 使用 ClickHouse

16.4.1 ZooKeeper 上数据错乱导致 ClickHouse 启动失败问题

问题现象

ClickHouse集群中某实例节点启动失败，该实例节点启动日志中有如下类似报错信息：

```
2021.03.15 21:01:19.816593 [ 11111 ] {} <Error> Application: DB::Exception:
The local set of parts of table DEFAULT.lineorder doesn't look like the set ofdoesn't look like the set of
parts in ZooKeeper: 59.99 million rows of 59.99 million total rows in
filesystem are suspicious. There are 30 unexpected parts with 59986052 rows
(14 of them is not just-written with 59986052 rows), 0 missing parts (with 0
blocks).: Cannot attach table `DEFAULT`.`lineorder` from metadata file
...
: while loading database
```

原因分析

使用ClickHouse过程中，ClickHouse实例异常场景下，重复创建集群ReplicatedMergeTree引擎表，后续又进行删除表等操作导致ZooKeeper上的数据异常，致使ClickHouse启动失败。

解决办法

步骤1 备份问题节点数据库下所有表数据到其他目录。

- 备份表数据：
  cd /srv/BigData/data*1*/**clickhouse/data/***数据库名*
  mv *表名 待备份的目录*/**data***1*

  说明

  如果存在多磁盘的场景，需要对**data1**到**dataN**的磁盘数据都执行相同的备份操作。
- 备份元数据信息：
  cd /srv/BigData/data1/**clickhouse_path/metadata**
  mv *表名*.**sql** *待备份的目录*

例如，下面是备份default数据库下的表lineorder数据到**/home/backup**目录下。

cd /srv/BigData/data1/clickhouse/data/default

mv lineorder /home/backup/data1

cd /srv/BigData/data1/clickhouse_path/metadata

mv lineorder.sql /home/backup

步骤2 登录MRS Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，选择对应的实例节点，单击“启动实例”，完成实例启动。

步骤3 实例启动成功后，使用ClickHouse客户端登录问题节点。

clickhouse client --host *clickhouse实例IP* **--user** *用户名* **--password** *密码*

步骤4 执行以下命令获取当前表所在的ZooKeeper路径**zookeeper_path**和对应节点所在的副本编号**replica_num**。

SELECT zookeeper_path FROM system.replicas WHERE database = '*数据库名***' AND table = '***表名***';**

SELECT replica_num,host_name FROM system.clusters;

步骤5 执行以下命令连接ZooKeeper命令行界面。

zkCli.sh -server *ZooKeeper所在节点的IP*:**2181**

步骤6 找到对应故障节点表数据对应的ZooKeeper路径。

ls *zookeeper_path*/**replicas/***replica_num*

说明

zookeeper_path为**步骤4**中查询到的**zookeeper_path**值。
replica_num为**步骤4**中节点主机对应的副本编号**replica_num**的值。

步骤7 执行以下命令，删除ZooKeeper上的副本数据。

deleteall *zookeeper_path*/**replicas/***replica_num*

步骤8 使用ClickHouse客户端登录问题节点，重新执行create创建集群ReplicatedMergeTree引擎表。

clickhouse client --host *clickhouse实例IP* **--multiline --user** *用户名* **--password** *密码*

CREATE TABLE *数据库名.表名* **ON CLUSTER** *集群名*

...

ENGINE = ReplicatedMergeTree ...

其他副本节点有如下提示表已经存在的报错信息，属于正常现象，可以忽略。

```
Received exception from server (version 20.8.7):
Code: 57. DB::Exception: Received from x.x.x.x:9000. DB::Exception:
There was an error on [x.x.x.x:9000]: Code: 57, e.displayText() =
DB::Exception: Table DEFAULT.lineorder already exists. (version 20.8.11.17
(official build)).
```

建表成功后问题节点上表数据会自动进行同步，数据恢复完成。

----结束

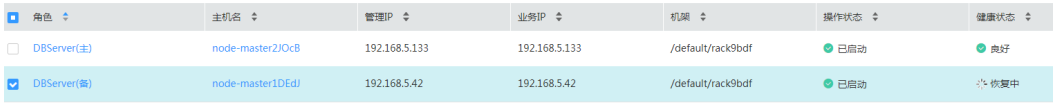
16.5 使用 DBservice

16.5.1 DBServer 实例状态异常

问题背景与现象

DBServer实例状态一直是concerning。

图 16-12 DBServer 实例状态



原因分析

数据目录下文件或目录的权限不对，GaussDB要求文件权限至少是600，目录权限至少为700。

图 16-13 目录权限列表

```
omm@192-168-234-176:/srv/BigData/dbdata_service> ll
total 4
drwx------ 19 omm wheel 4096 Dec 14 10:15 data
```

图 16-14 文件权限列表

```
omm@192-168-234-176:/srv/BigData/dbdata_service/data> ll
total 128
drwx------ 6 omm wheel  4096 Dec  9 15:47 base
-rw------- 1 omm wheel   922 Dec  9 15:34 dblink.conf
-rw------- 1 omm wheel    16 Dec 14 10:15 gaussdb.state
drwx------ 2 omm wheel  4096 Dec 14 10:17 global
drwx------ 2 omm wheel  4096 Dec 11 00:00 pg_audit
drwx------ 2 omm wheel  4096 Dec 14 10:15 pg_blackbox
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_clog
drwx------ 2 omm wheel  4096 Dec 14 10:15 pg_confile_backup
-rw------- 1 omm wheel  1024 Dec  9 15:34 pg_ctl.lock
-rw------- 1 omm wheel  4245 Dec  9 15:47 pg_hba.conf
-rw------- 1 omm wheel  1024 Dec  9 15:47 pg_hba.conf.lock
-rw------- 1 omm wheel  1636 Dec  9 15:34 pg_ident.conf
drwx------ 2 omm wheel  4096 Dec  9 15:38 pg_log
drwx------ 4 omm wheel  4096 Dec  9 15:34 pg_multixact
drwx------ 2 omm wheel  4096 Dec 14 10:15 pg_notify
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_serial
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_snapshots
drwx------ 2 omm wheel  4096 Dec 14 11:56 pg_stat_tmp
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_subtrans
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_tblspc
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_twophase
-rw------- 1 omm wheel     4 Dec  9 15:34 PG_VERSION
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_wallet
drwx------ 3 omm wheel  4096 Dec  9 15:39 pg_xlog
-rw------- 1 omm wheel 13309 Dec 14 10:15 postgresql.conf
-rw------- 1 omm wheel  1024 Dec  9 15:34 postgresql.conf.lock
-rw------- 1 omm wheel   105 Dec 14 10:15 postmaster.opts
-rw------- 1 omm wheel    96 Dec 14 10:15 postmaster.pid
```

解决办法

步骤1 按照图16-13和图16-14的权限列表，修改相应文件和目录的权限。

步骤2 重启相应的DBServer实例。

----结束

16.5.2 DBServer 实例一直处于 Restoring 状态

问题背景与现象

DBServer实例状态一直是Restoring状态，重启之后仍然不恢复。

原因分析

1. DBService组件会对“${BIGDATA_HOME}/MRS_*XXX*/install/dbservice/ha/module/harm/plugin/script/gsDB/.startGS.fail”这个文件监控。其中*XXX*是产品版本号。
2. 如果这个文件中的值大于3就会启动失败，NodeAgent会一直尝试重启该实例，此时仍会失败而且这个值每启动失败一次就会加1。

解决办法

步骤1 登录Manager管理界面。

步骤2 停止该DBServer实例。

步骤3 使用**omm**用户登录到DBServer实例异常的节点。

步骤4 修改“${BIGDATA_HOME}/MRS_*XXX*/install/dbservice/ha/module/harm/plugin/script/gsDB/.startGS.fail”配置文件中的值为0。其中*XXX*是产品版本号。

步骤5 启动该DBServer实例。

----结束

16.5.3 默认端口 20050 或 20051 被占用

问题背景与现象

执行DBService服务重启操作时，DBService服务启动失败，打印的错误日志中出现20050或20051端口被占用等信息。

原因分析

1. 由于DBService使用的默认端口20050或20051被其他进程占用。
2. DBService进程没有停止成功，使用的端口未释放。

解决办法

该解决办法以20051端口被占用为例，20050端口被占用的解决办法与该办法类似。

步骤1 以**root**用户登录DBService安装报错的节点主机，执行命令：**netstat -nap | grep 20051**查看占用20051端口的进程。

步骤2 使用**kill**命令强制终止使用20051端口的进程。

步骤3 约2分钟后，再次执行命令：**netstat -nap | grep 20051**，查看是否还有进程占用该端口。

步骤4 确认占用该端口进程所属的服务，并修改为其他端口。

步骤5 分别在“/tmp”和“/var/run/MRS-DBService/”目录下执行**find . -name "*20051*"**命令，将搜索到的文件全部删除。

步骤6 登录Manager，重启DBService服务。

----结束

16.5.4 /tmp 目录权限不对导致 DBserver 实例状态一直处于 Restoring

问题背景与现象

DBServer实例状态一直是Restoring状态，重启之后仍然不恢复。

原因分析

1. 查看“/var/log/Bigdata/dbservice/healthCheck/dbservice_processCheck.log”，可以看到gaussdb异常。

图 16-15 gaussdb 异常

```
[2019-07-22 10:57:00] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:00] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:10] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:10] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
Normal.
[2019-07-22 10:57:10] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:10] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:20] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:20] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
Normal.
[2019-07-22 10:57:20] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:20] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:30] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:31] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
Normal.
[2019-07-22 10:57:31] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:31] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:41] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:41] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
```

2. 检查发现“/tmp”权限不对。

图 16-16 /tmp 权限

```
[root@node-master1DEdJ DB]# ll / -rlth
total 76K
drwxr-xr-x.   2 root root 4.0K Dec 12  2016 mnt
drwxr-xr-x.   2 root root 4.0K Dec 12  2016 media
drwxr-xr-x.  13 root root 4.0K Jul 15 16:25 usr
-rwxr-xr-x.   1 root root 3.8K Jul 15 16:25 README
-rwxr-xr-x.   1 root root    0 Jul 15 16:25 OTC_EulerOS_2.x86_64-0.9.1-20170904-0513
lrwxrwxrwx.   1 root root    8 Jul 15 16:26 sbin -> usr/sbin
lrwxrwxrwx.   1 root root    9 Jul 15 16:26 lib64 -> usr/lib64
lrwxrwxrwx.   1 root root    7 Jul 15 16:26 lib -> usr/lib
lrwxrwxrwx.   1 root root    7 Jul 15 16:26 bin -> usr/bin
drwxr-xr-x.   3 root root 4.0K Jul 15 16:29 srv
drwxr-xr-x.   7 root root 4.0K Jul 15 16:39 CloudResetPwdUpdateAgent
drwxr-xr-x.   7 root root 4.0K Jul 15 16:39 CloudrResetPwdAgent
drwx------.   2 root root  16K Jul 15 16:46 lost+found
dr-xr-xr-x. 236 root root    0 Jul 19 17:36 proc
dr-xr-xr-x.   4 root root 4.0K Jul 19 17:37 boot
dr-xr-xr-x.  13 root root    0 Jul 19 17:37 sys
drwxr-xr-x.  19 root root 4.0K Jul 19 17:37 var
drwxr-xr-x.  19 root root 3.0K Jul 19 17:37 dev
drwxr-xr-x.   2 root root 4.0K Jul 19 17:38 tmpdir
drwxr-xr-x.   7 root root 4.0K Jul 19 17:38 opt
-rw-------.   1 root root    0 Jul 19 17:39 install_os_optimization.log
drwxr-xr-x.   6 root root 4.0K Jul 19 17:54 home
drwxr-xr-x.  86 root root 4.0K Jul 19 17:54 etc
drwxr-xr-x.  30 root root  960 Jul 22 10:49 run
drwx------.  23 root root 4.0K Jul 22 11:42 tmp
drwx------.   5 root root 4.0K Jul 22 11:50 root
```

解决办法

步骤1 修改/tmp的权限。

chmod 1777 /tmp

步骤2 等待实例状态恢复。

----结束

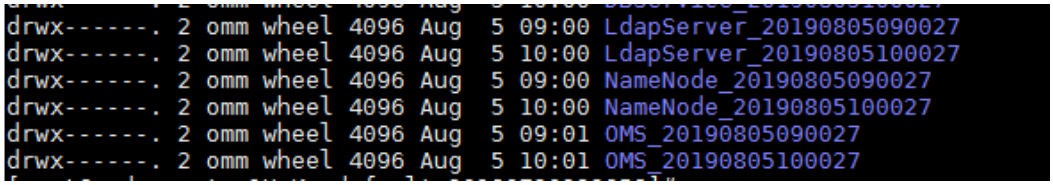
16.5.5 DBService 备份失败

问题背景与现象

ls /srv/BigData/LocalBackup/default_20190720222358/ -rlth

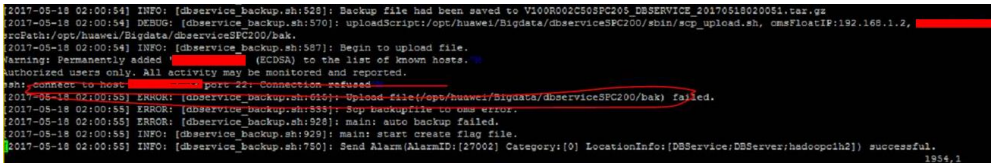
查看备份文件路径中没有DBService的备份文件。

图 16-17 查看备份文件



原因分析

- 查看DBService的备份日志/var/log/Bigdata/dbservice/scriptlog/backup.log，其实备份已经成功，只是上传至OMS节点时失败。



- 失败原因是由于ssh不通。

解决办法

步骤1 网络问题，联系网络工程师处理。

步骤2 网络问题解决之后重新备份即可。

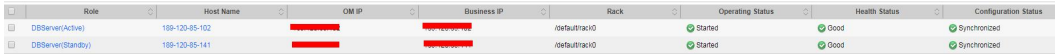
----结束

16.5.6 DBService 状态正常，组件无法连接 DBService

问题背景与现象

上层组件连接DBService失败，检查DBService组件状态正常，两个实例状态也正常。

图 16-18 DBService 状态



原因分析

1. 上层组件是通过dbservice.floatip连接的DBService。
2. 在DBServer所在节点执行命令netstat -anp | grep 20051发现，DBService的Gauss进程在启动时并未绑定floatip，只监听了127.0.0.1的本地ip。

解决办法

步骤1 重新启动DBService服务。

步骤2 启动完成之后在主DBServer节点执行**netstat -anp | grep 20051**命令检查是否绑定了dbservice.floatip。

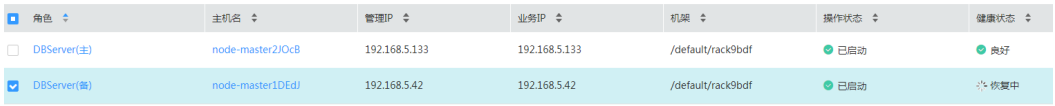
----结束

16.5.7 DBServer 启动失败

问题背景与现象

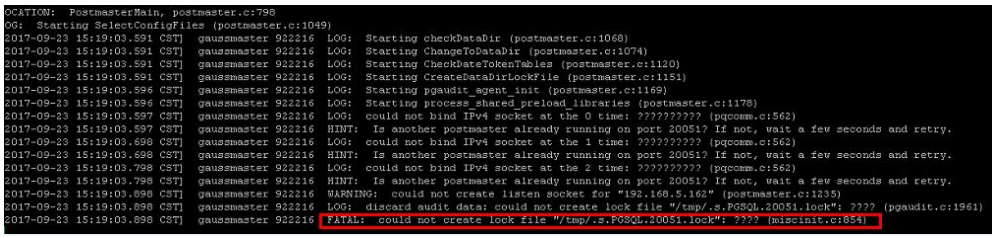
DBService组件启动失败，重启还是失败，实例状态一直为正在恢复状态。

图 16-19 DBService 的状态

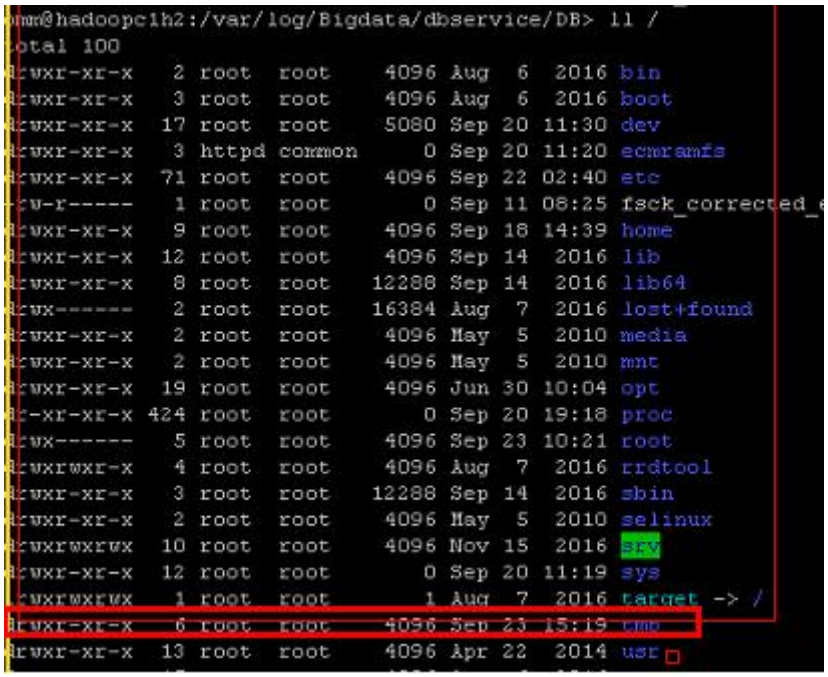


原因分析

1. 查看DBService的日志/var/log/Bigdata/dbservice/DB/gs_ctl-current.log，报如下错误。



2. 检查发现/tmp权限不正确，正确的权限应该为777。



解决办法

步骤1 修改/tmp权限为777。

步骤2 重新启动DBService组件。

----结束

16.5.8 浮动 IP 不通导致 DBService 备份失败

问题背景与现象

在默认备份default中DBService备份失败，其他备份（NameNode、LdapServer、OMS备份）成功。

原因分析

1. 查看DBService的备份页面错误信息，有如下错误信息提示：

```
Clear temporary files at backup checkpoint DBService_test_DBService_DBService_20180326155921 that fialed last time.
Temporyary files at backup checkpoint DBService_test_DBService_DBService20180326155921 that failed last time are cleared successfully.
Start executing the backup task.
The backup of configuration DBService is started.
Check the backup available disk space.
Backup initialization succeeded for configuration DBService.
Clear temporary files at backup checkpoint DBService_test_DBService_DBService_20180326155921 that failed last time.
Temporary files at backup checkpoint DBService_test_DBService_DBService_20180326155921 that failed last time are cleared successfully.
Checkpoint DBService_test_DBService_DBService_20180326162235 is verified successfully before backup.
Temporary files are cleared successfully before backup checkpoint DBService_test_DBService_DBService_20180326162235.
Prestart backup succeeded for checkpoint DBService_test_DBService_DBService_20180326162235.
The snapshot is created successfully for checkpoint DBService_test_DBService_DBService_20180326162235 before backup.
Backup is being performed for checkpoint DBService_test_DBService_DBService_20180326162235.
Backup execution failed. Task ID: 2
Detail: DBService backup task failed, please view details in logs.
Temporary files are cleared successfully after backup checkpoint DBService_test_DBService_DBService_20180326162235.
checkpoint DBService_test_DBService_DBService_20180326162235 is deleted successfully after backup failure.
Failed to backup configuration DBService.
```

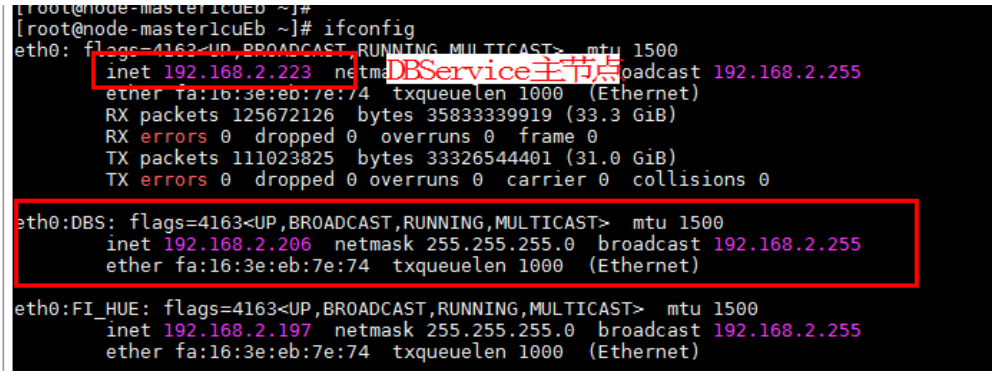
2. 查看/var/log/Bigdata/dbservice/scriptlog/backup.log文件，发现日志停止打印，并没有备份相关信息。
3. 查看主OMS节点 /var/log/Bigdata/controller/backupplugin.log日志发现如下错误信息：

```
result error is ssh:connect to host 172.16.4.200 port 22 : Connection refused (172.16.4.200是DBService的浮动IP)
DBService backup failed.
```

```
2018-03-27 07:00:35,758 INFO  [pool-1-thread-5] Create adapter from com.huawei.bigdata.om.backup.MetadataPluginAdapter success.
com.huawei.bigdata.om.backup.plugin.AbstractBackupRecoveryPlugin.initializePluginAdapter(AbstractBackupRecoveryPlugin.java:92)
2018-03-27 07:00:35,759 INFO  [pool-1-thread-5] floatIp is 172.16.4.200. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.getFloatIp(BackupRecoveryPlugin.java:233)
2018-03-27 07:00:35,759 INFO  [pool-1-thread-5] cmd is ssh 172.16.4.200 /opt/huawei/Bigdata/FusionInsight_V100R002C60U20/dbservice/sbin/dbservice_backup.sh -b -d
/srv/BigData/LocalBackup/default_20180326213206/DBService_20180327070010. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.startBackup(BackupRecoveryPlugin.java:166)
2018-03-27 07:00:35,759 INFO  [pool-1-thread-5] create task taskId is 6. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.startBackup(BackupRecoveryPlugin.java:169)
2018-03-27 07:00:35,760 INFO  [pool-1-thread-5] startBackup result OperateResult(errorCode:RUNNING, result:6, detailInfo:, packageName:null).
com.huawei.bigdata.om.backup.BackupPluginContainerHandler.startBackup(BackupPluginContainerHandler.java:246)
2018-03-27 07:00:35,760 INFO  [Thread-132] Executing the command with arguments and env, timeout: 900000
com.huawei.bigdata.om.controller.api.extern.monitor.script.LinuxScriptExecutionHandler.logMessage(LinuxScriptExecutionHandler.java:64)
2018-03-27 07:00:35,863 INFO  [Thread-132] Execute command : /opt/huawei/Bigdata/om-0.0.1/sbin/scriptLauncher.sh ssh 172.16.4.200
/opt/huawei/Bigdata/FusionInsight_V100R002C60U20/dbservice/sbin/dbservice_backup.sh -b -d /srv/BigData/LocalBackup/default_20180326213206/DBService_20180327070010.
com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:48)
2018-03-27 07:00:35,863 INFO  [Thread-132] result status is 255. com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:49)
2018-03-27 07:00:35,863 INFO  [Thread-132] result output is . com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:50)
2018-03-27 07:00:35,863 INFO  [Thread-132] result erro is ssh: connect to host 172.16.4.200 port 22: Connection refused
. com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:51)
2018-03-27 07:00:35,863 ERROR [Thread-132] DBService backup failed. com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:64)
2018-03-27 07:00:40,868 INFO  [pool-1-thread-5] query backup taskId is 6. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.getBackupProgress(BackupRecoveryPlugin.java:247)
```

解决办法

步骤1 登录DBService主节点（绑定有DBService浮动IP的master节点）。



步骤2 检查 /etc/ssh/sshd_config文件中ListenAddress配置项，添加DBService浮动IP到ListenAdress或者注释掉ListenAdress配置项。

步骤3 执行如下命令重启sshd服务。

service sshd restart

步骤4 观察下次备份DBServie是否备份成功。

----结束

16.5.9 DBService 配置文件丢失导致启动失败

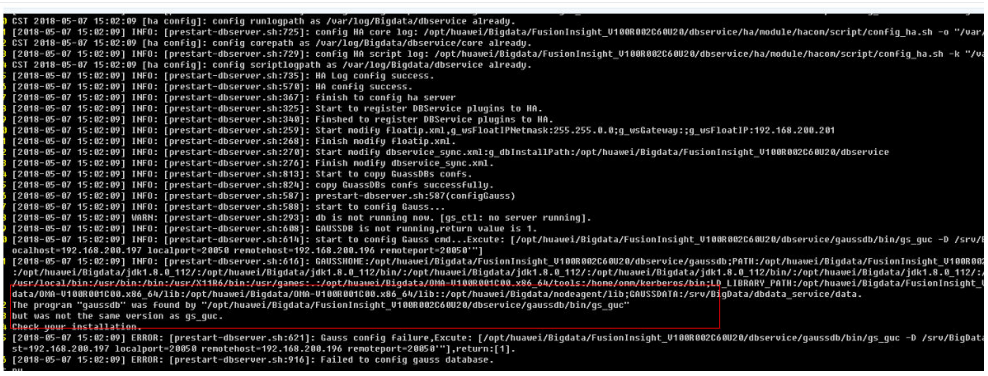
问题背景与现象

节点异常下电，重启备DBService失败。

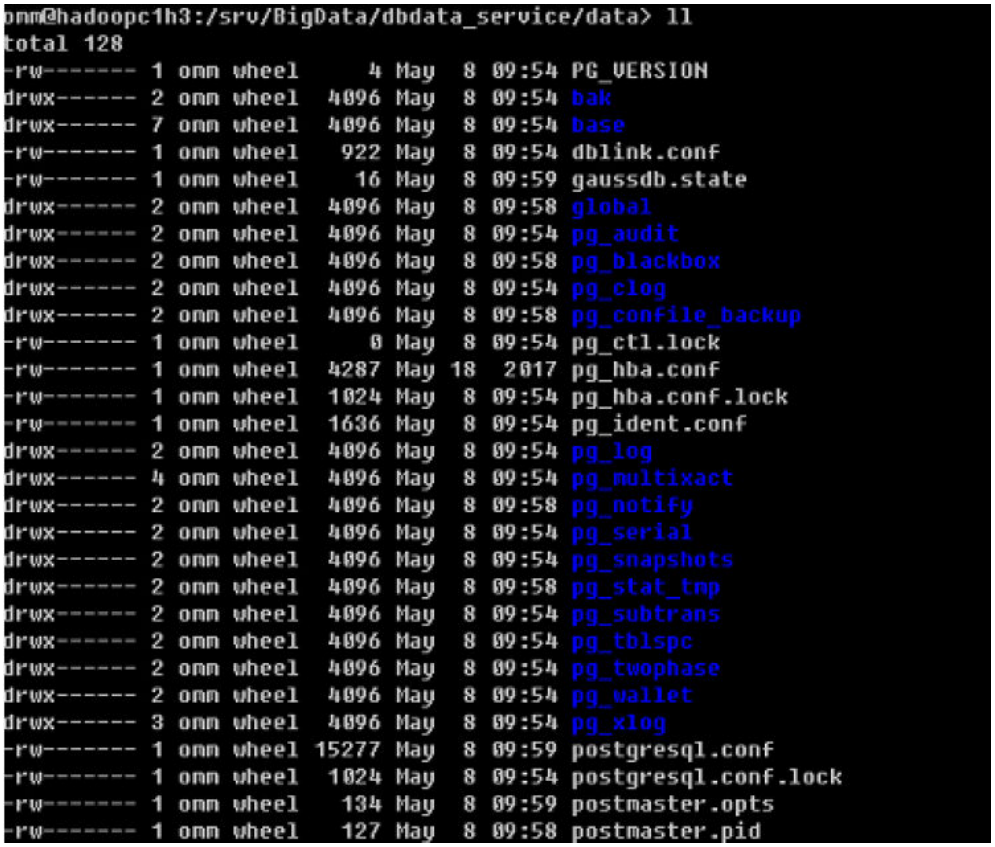
原因分析

1. 查看/var/log/Bigdata/dbservice/DB/gaussdb.log日志没有内容。
2. 查看/var/log/Bigdata/dbservice/scriptlog/preStartDBService.log日志，发现如下信息，判断为配置信息丢失。

```
The program "gaussdb" was found by "
/opt/Bigdata/MRS_xxx/install/dbservice/gaussdb/bin/gs_guc)
But not was not the same version as gs_guc.
Check your installation.
```



3. 比对主备DBServer节点/srv/BigData/dbdata_service/data目录下的配置文件发现差距比较大。



```
nm@hadoopc1h3:/srv/BigData/dbdata_service> cd data_bak/
nm@hadoopc1h3:/srv/BigData/dbdata_service/data_bak> ll
otal 64
rw------- 1 onm wheel   202 Feb 11 10:43 backup_label
rw------- 1 onm wheel     0 Feb 11 10:42 build_completed.start
rw------- 1 onm wheel    16 Apr 28 17:32 gaussdb.state
rw------- 1 onm wheel     7 Apr 28 17:32 gs_build.pid
rwx------ 2 onm wheel  4096 Feb 11 10:44 pg_audit
rwx------ 2 onm wheel  4096 Feb 11 10:41 pg_blackbox
rwx------ 2 onm wheel  4096 Feb 11 10:09 pg_confile_backup
rw------- 1 onm wheel     0 Apr 28 17:32 pg_ctl.lock
rw------- 1 onm wheel  4287 May 18  2017 pg_hba.conf
rwx------ 2 onm wheel  4096 Feb 11 10:43 pg_notify
rwx------ 2 onm wheel  4096 Feb 11 10:43 pg_xlog
rw------- 1 onm wheel 15155 May  7 15:33 postgresql.conf
rw------- 1 onm wheel  1024 May  7 15:33 postgresql.conf.lock
rw------- 1 onm wheel   134 Feb 11 10:42 postmaster.opts
```

解决办法

步骤1 把主节点**/srv/BigData/dbdata_service/data**的内容拷贝到备节点，保持文件权限和属组与主节点一样。

步骤2 修改postgresql.conf配置信息，localhost修改成本节点IP，remotehost修改成对端节点IP。

```
#------------------------------------------------------------------------------
# CUSTOMIZED OPTIONS
#------------------------------------------------------------------------------

# Add settings for extensions here
max_files_per_process = 300
unix_socket_directory = '/var/run/FusionInsight-DBService'
replconninfo1 = 'localhost=192.168.200.197 localport=20050 remotehost=192.168.200.196 remoteport=20050'
"postgresql.conf" 382L, 15277C
```

步骤3 登录Manager页面重启备DBServer节点。

----结束

16.6 使用 Flink

16.6.1 安装客户端执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file :"security.kerberos.login.keytab"

问题背景与现象

客户端安装成功，执行客户端命令例如yarn-session.sh命令报错，提示IllegalConfigurationException: Error while parsing YAML configuration file : "security.kerberos.login.keytab: "

```
[root@8-5-131-10 bin]# yarn-session.sh
2018-10-25 01:22:06,454 | ERROR | [main] | Error while trying to split key and value in configuration file /opt/flinkclient/Flink/flink/conf/flink-conf.yaml:80: "security.kerberos.login.keytab: " | org.apache.flink.configuration.GlobalConfiguration (GlobalConfiguration.java:160)
Exception in thread "main" org.apache.flink.configuration.IllegalConfigurationException: Error while parsing YAML configuration file :80: "security.kerberos.login.keytab: "
        at org.apache.flink.configuration.GlobalConfiguration.loadYAMLResource(GlobalConfiguration.java:161)
        at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:112)
        at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:79)
        at org.apache.flink.yarn.cli.FlinkYarnSessionCli.main(FlinkYarnSessionCli.java:482)
[root@8-5-131-10 bin]#
```

原因分析

在安全集群环境下，Flink需要进行安全认证。当前客户端未进行相关安全认证设置。

1. Flink整个系统有两种认证方式：
   - 使用kerberos认证：Flink yarn client、Yarn Resource Manager、JobManager、HDFS、TaskManager、Kafka和Zookeeper。
   - 使用YARN内部的认证机制：Yarn Resource Manager与Application Master（简称AM）。
2. 如果用户安装安全集群需要使用kerberos认证和security cookie认证。根据日志提示，发现配置文件中“security.kerberos.login.keytab :”配置项错误，未进行安全配置。

解决办法

步骤1 从MRS上下载用户keytab，并将keytab放到Flink客户端所在主机的某个文件夹下。

步骤2 在“flink-conf.yaml”上配置：

1. keytab路径。
   ```
   security.kerberos.login.keytab: /home/flinkuser/keytab/abc222.keytab
   ```
   说明
   - “/home/flinkuser/keytab/abc222.keytab”表示的是用户目录，为**步骤1**中放置目录。
   - 请确保客户端用户具备对应目录权限。
2. principal名。
   ```
   security.kerberos.login.principal: abc222
   ```
3. 对于HA模式，如果配置了ZooKeeper，还需要设置ZK kerberos认证相关的配置。配置如下：
   ```
   zookeeper.sasl.disable: false
   security.kerberos.login.contexts: Client
   ```
4. 如果用户对于Kafka client和Kafka broker之间也需要做kerberos认证，配置如下：
   ```
   security.kerberos.login.contexts: Client,KafkaClient
   ```

----结束

16.6.2 安装客户端修改配置后执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file

问题背景与现象

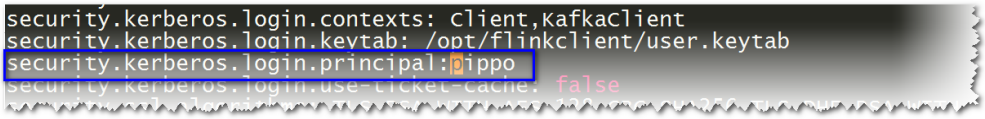
客户端安装成功，执行客户端命令例如yarn-session.sh命令报错，提示 IllegalConfigurationException: Error while parsing YAML configuration file :81: "security.kerberos.login.principal:pippo "

```
[root@8-5-131-10 bin]# yarn-session.sh
2018-10-25 19:27:01,397 | ERROR | [main] | Error while trying to split key and value in configuration file /opt/flinkclient/Flink/flink/conf/flink-conf.yaml:81: "security.kerberos.login.principal:pippo " | org.apache.flink.configuration.GlobalConfiguration (GlobalConfiguration.java:160)
Exception in thread "main" org.apache.flink.configuration.IllegalConfigurationException: Error while parsing YAML configuration file :81: "security.kerberos.login.principal:pippo "
```

```
    at org.apache.flink.configuration.GlobalConfiguration.loadYAMLResource(GlobalConfiguration.java:161)
    at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:112)
    at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:79)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli.main(FlinkYarnSessionCli.java:482)
```

原因分析

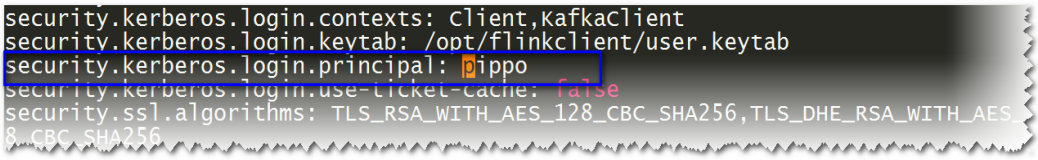
配置文件flink-conf.yaml中配置项"security.kerberos.login.principal:pippo”错误。



解决办法

修改flink-conf.yaml中配置。

注意：配置项名称和值之间存在空格。



16.6.3 创建 Flink 集群时执行 yarn-session.sh 命令失败

问题背景与现象

创建Fllink集群时，执行yarn-session.sh命令卡住一段时间后报错：

```
2018-09-20 22:51:16,842 | WARN  | [main] | Unable to get ClusterClient status from Application Client |
org.apache.flink.yarn.YarnClusterClient (YarnClusterClient.java:253)
org.apache.flink.util.FlinkException: Could not connect to the leading JobManager. Please check that the
JobManager is running.
    at org.apache.flink.client.program.ClusterClient.getJobManagerGateway(ClusterClient.java:861)
    at org.apache.flink.yarn.YarnClusterClient.getClusterStatus(YarnClusterClient.java:248)
    at org.apache.flink.yarn.YarnClusterClient.waitForClusterToBeReady(YarnClusterClient.java:516)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli.run(FlinkYarnSessionCli.java:717)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli$1.call(FlinkYarnSessionCli.java:514)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli$1.call(FlinkYarnSessionCli.java:511)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1729)
    at org.apache.flink.runtime.security.HadoopSecurityContext.runSecured(HadoopSecurityContext.java:41)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli.main(FlinkYarnSessionCli.java:511)
Caused by: org.apache.flink.runtime.leaderretrieval.LeaderRetrievalException: Could not retrieve the leader
gateway.
    at org.apache.flink.runtime.util.LeaderRetrievalUtils.retrieveLeaderGateway(LeaderRetrievalUtils.java:79)
    at org.apache.flink.client.program.ClusterClient.getJobManagerGateway(ClusterClient.java:856)
    ... 10 common frames omitted
Caused by: java.util.concurrent.TimeoutException: Futures timed out after [10000 milliseconds]
```

可能原因

Flink开启了SSL通信加密，却没有正确的配置SSL证书。

解决办法

针对MRS 2.x及之前版本，操作如下：

方法1：

关闭Flink SSL通信加密，修改客户端配置文件**conf/flink-conf.yaml**。

```
security.ssl.internal.enabled: false
```

方法2：

开启Flink SSL通信加密，security.ssl.internal.enabled 保持默认。正确配置SSL：

- 配置keystore或truststore文件路径为**相对路径**时，Flink Client执行命令的目录需要可以直接访问该相对路径

  ```
  security.ssl.internal.keystore: ssl/flink.keystore
  security.ssl.internal.truststore: ssl/flink.truststore
  ```

  在Flink的CLI yarn-session.sh命令中增加“-t”选项来传输keystore和truststore文件到各个执行节点。如：

  ```
  yarn-session.sh -t ssl/ 2
  ```

- 配置keystore或truststore文件路径为**绝对路径**时，需要在Flink Client以及各个节点的该绝对路径上放置keystore或truststore文件。

  ```
  security.ssl.internal.keystore: /opt/client/Flink/flink/conf/flink.keystore
  security.ssl.internal.truststore: /opt/client/Flink/flink/conf/flink.truststore
  ```

针对MRS3.x及之后版本，操作如下：

方法1：

关闭Flink SSL通信加密，修改客户端配置文件**conf/flink-conf.yaml**。

```
security.ssl.enabled: false
```

方法2：

开启Flink SSL通信加密，security.ssl.enabled 保持默认。正确配置SSL：

- 配置keystore或truststore文件路径为**相对路径**时，Flink Client执行命令的目录需要可以直接访问该相对路径

  ```
  security.ssl.keystore: ssl/flink.keystore
  security.ssl.truststore: ssl/flink.truststore
  ```

  在Flink的CLI yarn-session.sh命令中增加“-t”选项来传输keystore和truststore文件到各个执行节点。如：

  ```
  yarn-session.sh -t ssl/ 2
  ```

- 配置keystore或truststore文件路径为**绝对路径**时，需要在Flink Client以及各个节点的该绝对路径上放置keystore或truststore文件。

  ```
  security.ssl.keystore: /opt/Bigdata/client/Flink/flink/conf/flink.keystore
  security.ssl.truststore: /opt/Bigdata/client/Flink/flink/conf/flink.truststore
  ```

16.6.4 使用不同用户，执行 yarn-session 创建集群失败

问题背景与现象

使用Flink过程中，具有两个相同权限用户testuser和bdpuser。

使用用户testuser创建Flink集群正常，但是切换至bdpuser用户创建Fllink集群时，执行yarn-session.sh命令报错：

```
2019-01-02 14:28:09,098 | ERROR | [main] | Ensure path threw exception |
org.apache.flink.shaded.curator.org.apache.curator.framework.imps.CuratorFrameworkImpl
```

```
(CuratorFrameworkImpl.java:566)
org.apache.flink.shaded.zookeeper.org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /flink/application_1545397824912_0022
```

可能原因

高可用配置项未修改。由于在Flink的配置文件中，**high-availability.zookeeper.client.acl**默认为**creator**，仅创建者有权限访问，新用户无法访问ZooKeeper上的目录导致yarn-session.sh执行失败。

解决办法

步骤1 修改客户端配置文件**conf/flink-conf.yaml**中配置项**high-availability.zookeeper.path.root**，例如：

```
high-availability.zookeeper.path.root： flink2
```

步骤2 重新提交任务。

----结束

16.6.5 Flink 业务程序无法读取 NFS 盘上的文件

用户问题

Flink业务程序无法读取集群节点挂载的NFS盘上的文件。

问题现象

用户开发的Flink业务程序中需要读取用户定义的配置文件，该配置文件放在NFS盘上，NFS盘是挂载在集群节点上的，集群的所有节点均可以访问该盘。用户提交flink程序后，业务代码访问不到客户自定义的配置文件，导致业务程序启动失败。

原因分析

该问题的根因是NFS盘上的根目录权限不足，导致Flink程序启动后无法访问该目录。

MRS的Flink任务是在YARN运行，当集群为普通集群时，在YARN上运行任务的用户为yarn_user。用户自定义的配置文件如果在任务启动之后使用，则文件以及文件的父目录（NFS上的文件所在的父目录，非集群节点上的软连接），必须允许yarn_user可以访问，否则程序中无法获取文件内容。当集群为kerberos集群时，则文件的权限必须允许提交程序的用户访问。

处理步骤

步骤1 以root用户登录集群的Master节点。

步骤2 执行如下命令查看用户自定义配置文件所在父目录的权限。

ll <文件所在路径的父目录路径>

步骤3 进入NFS盘待访问文件所在目录，修改用户自定义配置文件所在父目录的权限为755。

chmod 755 -R /<文件所在路径的父目录路径>

步骤4 确认Core或者Task节点是否可以访问到该配置文件。

1. 以root用户登录Core/Task节点。
   如果当前集群已启用Kerberos认证，请以root用户登录Core节点。
2. 执行**su - yarn_user**命令切换到yarn_user用户。
   如果当前集群已启用Kerberos认证，请执行 **su - 提交作业的用户**命令切换用户。
3. 执行如下命令查看用户权限，文件所在路径请使用该文件的绝对路径。
   ll <文件所在路径>

----结束

建议与总结

当用户提交的任务中要访问自定义的配置文件时，特别是挂载NFS盘时，除了确认文件的权限之外，还要确认文件所在父目录的权限是否正确。NFS盘挂载到MRS集群节点上，一般会新建软连接到NFS目录，这个时候需要查看NFS上的目录权限是否正确。

16.6.6 自定义 Flink log4j 日志输出级别

用户问题

MRS 3.1.0集群自定义Flink log4j日志级别不生效。

问题现象

1. 在使用MRS 3.1.0集群Flink数据分析时，将$Flink_HOME/conf目录下的log4j.properties文件中日志级别修改为INFO级别日志。
2. 任务正常提交后，console未打印出INFO级别日志，输出的日志级别还是ERROR级别。

原因分析

修改$Flink_HOME/conf目录下的log4j.properties文件，控制的是JobManager和TaskManager的算子内的日志输出，输出的日志会打印到对应的yarn contain中，可以在yarn web ui查看对应日志。MRS 3.1.0及之后版本的Flink 1.12.0版本开始默认的日志框架是log4j2，配置的方式跟之前log4j的方式有区别，使用如log4j日志规则不会生效。

处理步骤

Log4j2详细日志规格配置参考开源官方文档：**http://logging.apache.org/log4j/2.x/manual/configuration.html#Properties**

16.7 使用 Flume

16.7.1 Flume 向 Spark Streaming 提交作业，提交到集群后报类找不到

用户问题

Flume向Spark Streaming提交作业，提交到集群后报类找不到。

问题现象

Spark Streaming代码打成jar包提交到集群后报类找不到错误，通过以下两种方式依然不生效。

1. 在提交Spark作业的时候使用**--jars** 命令引用类所在的jar包。
2. 将类所在的jar包引入Spark Streaming的jar包。

原因分析

执行Spark作业时无法加载部分jar，导致找不到class。

处理步骤

步骤1 使用 **--jars** 加载flume-ng-sdk-{version}.jar依赖包。

步骤2 同时修改**spark-default.conf**中两个配置项。

spark.driver.extraClassPath=$PWD/*:{加上原来配置的值}

spark.executor.extraClassPath =$PWD/*

步骤3 作业运行成功。如果还有报错，则需要排查还有哪个jar没有加载，再次执行步骤1和步骤2。

----结束

16.7.2 Flume 客户端安装失败

问题现象

安装Flume客户端失败，提示JAVA_HOME is null或flume has been installed。

```
CST 2016-08-31 17:02:51 [flume-client install]: JAVA_HOME is null in current user,please install the JDK and set the JAVA_HOME
CST 2016-08-31 17:02:51 [flume-client install]: check environment failed.
CST 2016-08-31 17:02:51 [flume-client install]: check param failed.
CST 2016-08-31 17:02:51 [flume-client install]: install flume client failed.
```

```
CST 2016-08-31 17:03:58 [flume-client install]: flume has been installed
CST 2016-08-31 17:03:58 [flume-client install]: check path failed.
CST 2016-08-31 17:03:58 [flume-client install]: check param failed.
CST 2016-08-31 17:03:58 [flume-client install]: install flume client failed.
```

原因分析

- Flume客户端安装时会检查环境变量，如果没有可用的JAVA，会报JAVA_HOME is null错误并且退出安装。
- 如果指定的目录下已经安装有flume，客户端安装时会报flume has been installed并退出安装。

解决办法

步骤1 如果报JAVA_HOME is null错误，需要使用命令：

export JAVA_HOME=*java路径*

设置JAVA_HOME，重新运行安装脚本。

步骤2 如果指定的目录下已经安装有Flume客户端，需要先卸载已经存在的Flume客户端，或指定其他目录安装。

----结束

16.7.3 Flume 客户端无法连接服务端

问题现象

安装Flume客户端并设置avro sink与服务端通信，发现无法连接Flume服务端。

原因分析

1. 服务端配置错误，监听端口启动失败，例如服务端avro source配置了错误的IP，或者已经被占用了的端口。查看Flume运行日志：
   ```
   2016-08-31 17:28:42,092 | ERROR | [lifecycleSupervisor-1-9] | Unable to start EventDrivenSourceRunner: { source:Avro source avro_source: { bindAddress: 10.120.205.7, port: 21154 } } - Exception follows. | org.apache.flume.lifecycle.LifecycleSupervisor $MonitorRunnable.run(LifecycleSupervisor.java:253)
   java.lang.RuntimeException: org.jboss.netty.channel.ChannelException: Failed to bind to: /192.168.205.7:21154
   ```
2. 若采用了加密传输，证书或密码错误。
   ```
   2016-08-31 17:15:59,593 | ERROR | [conf-file-poller-0] | Source avro_source has been removed due to an error during configuration | org.apache.flume.node.AbstractConfigurationProvider.loadSources(AbstractConfigurationProvider.java:388)
   org.apache.flume.FlumeException: Avro source configured with invalid keystore: /opt/Bigdata/MRS_XXX/install/FusionInsight-Flume-1.9.0/flume/conf/flume_sChat.jks
   ```
3. 客户端与服务端通信异常。
   ```
   PING 192.168.85.55 (10.120.85.55) 56(84) bytes of data.
   From 192.168.85.50 icmp_seq=1 Destination Host Unreachable
   From 192.168.85.50 icmp_seq=2 Destination Host Unreachable
   From 192.168.85.50 icmp_seq=3 Destination Host Unreachable
   From 192.168.85.50 icmp_seq=4 Destination Host Unreachable
   ```

解决办法

步骤1 设置为正确的IP，必须为本机的IP，如果端口被占用，重新配置一个空闲的端口。

步骤2 配置正确的证书路径。

步骤3 联系网络管理员，恢复网络。

----结束

16.7.4 Flume 数据写入组件失败

问题现象

Flume进程启动后，Flume数据无法写入到对应组件。（以下以服务端写入到HDFS为例）

原因分析

1. HDFS未启动或故障。查看Flume运行日志：
   ```
   2019-02-26 11:16:33,564 | ERROR | [SinkRunner-PollingRunner-DefaultSinkProcessor] | opreation the hdfs file errors. | org.apache.flume.sink.hdfs.HDFSEventSink.process(HDFSEventSink.java:414)
   2019-02-26 11:16:33,747 | WARN | [hdfs-CCCC-call-runner-4] | A failover has occurred since the start
   ```

```
of call #32795 ClientNamenodeProtocolTranslatorPB.getFileInfo over
192-168-13-88/192.168.13.88:25000 | org.apache.hadoop.io.retry.RetryInvocationHandler
$ProxyDescriptor.failover(RetryInvocationHandler.java:220)
2019-02-26 11:16:33,748 | ERROR | [hdfs-CCCC-call-runner-4] | execute hdfs error. {} |
org.apache.flume.sink.hdfs.HDFSEventSink$3.call(HDFSEventSink.java:744)
java.net.ConnectException: Call From 192-168-12-221/192.168.12.221 to 192-168-13-88:25000 failed
on connection exception: java.net.ConnectException: Connection refused; For more details see: http://
wiki.apache.org/hadoop/ConnectionRefused
```

2. hdfs sink未启动。查看Flume运行日志，发现“ flume current metrics”中并没有sink信息：

```
2019-02-26 11:46:05,501 | INFO | [pool-22-thread-1] | flume current metrics:{"CHANNEL.BBBB":
{"ChannelCapacity":"10000","ChannelFillPercentage":"0.0","Type":"CHANNEL","ChannelStoreSize":"0","
EventProcessTimedelta":"0","EventTakeSuccessCount":"0","ChannelSize":"0","EventTakeAttemptCount":
"0","StartTime":"1551152734999","EventPutAttemptCount":"0","EventPutSuccessCount":"0","StopTime
":"0"},"SOURCE.AAAA":
{"AppendBatchAcceptedCount":"0","EventAcceptedCount":"0","AppendReceivedCount":"0","MonTime":
"0","StartTime":"1551152735503","AppendBatchReceivedCount":"0","EventReceivedCount":"0","Type":"
SOURCE","TotalFilesCount":"1001","SizeAcceptedCount":"0","UpdateTime":"605410241202740","Appen
dAcceptedCount":"0","OpenConnectionCount":"0","MovedFilesCount":"1001","StopTime":"0"}} |
org.apache.flume.node.Application.getRestartComps(Application.java:467)
```

解决办法

步骤1 若Flume数据写入的组件未启动，启动对应组件；若组件异常，请联系服务技术支持。

步骤2 sink未启动，检查配置文件是否配置正确，若配置错误，则正确修改配置文件后重启Flume进程，如果配置正确，则查看日志错误信息，根据具体错误信息指定解决办法。

----结束

16.7.5 Flume 服务端进程故障

问题现象

Flume运行一段时间后，Manager界面Flume实例显示运行状态“故障”。

原因分析

Flume文件或文件夹权限异常，重启后Manager界面提示如下信息：

```
[2019-02-26 13:38:02]RoleInstance prepare to start failure [{ScriptExecutionResult=ScriptExecutionResult
[exitCode=126, output=, errMsg=sh: line 1: /opt/Bigdata/MRS_XXX/install/FusionInsight-Flume-1.9.0/
flume/bin/flume-manage.sh: Permission denied
```

解决办法

与运行正常的Flume节点进行文件和文件夹权限对比，更改错误文件或文件夹权限。

16.7.6 Flume 数据采集慢

问题现象

Flume启动后，Flume数据采集慢。

原因分析

1. Flume堆内存设置不合理，导致Flume进程一直处于频繁GC。查看Flume运行日志：
   2019-02-26T13:06:20.666+0800: 1085673.512: [Full GC:[CMS: 3849339k->3843458K(3853568K), 2.5817610 secs] 4153654K->3843458K(4160256K), [CMS Perm : 27335K->27335K(45592K),2.5820080 SECS] [Times: user=2.63, sys0.00, real=2.59 secs]
2. 用户业务配置的Spooldir source的deletePolicy策略是立即删除（immediate）。

解决办法

步骤1 适当调大堆内存（xmx）的值。

步骤2 将Spooldir source的deletePolicy策略更改为永不删除（never）。

----结束

16.7.7 Flume 启动失败

问题现象

安装Flume服务或重启Flume服务失败。

原因分析

1. Flume堆内存设置的值大于机器剩余内存，查看Flume启动日志：
   [CST 2019-02-26 13:31:43][INFO] [[checkMemoryValidity:124]] [GC_OPTS is invalid: Xmx(40960000MB) is bigger than the free memory(56118MB) in system.] [9928]
2. Flume文件或文件夹权限异常，界面或后台会提示如下信息：
   [2019-02-26 13:38:02]RoleInstance prepare to start failure [{ScriptExecutionResult=ScriptExecutionResult [exitCode=126, output=, errMsg=sh: line 1: /opt/Bigdata/MRS_XXX/install/FusionInsight-Flume-1.9.0/flume/bin/flume-manage.sh: Permission denied
3. JAVA_HOME配置错误，查看Flume agent启动日志：
   Info: Sourcing environment configuration script /opt/FlumeClient/fusioninsight-flume-1.9.0/conf/flume-env.sh
   + '[' -n '' ']'
   + exec /tmp/MRS-Client/MRS_Flume_ClientConfig/JDK/jdk-8u18/bin/java '-XX:OnOutOfMemoryError=bash /opt/FlumeClient/fusioninsight-flume-1.9.0/bin/out_memory_error.sh /opt/FlumeClient/fusioninsight-flume-1.9.0/conf %p' -Xms2G -Xmx4G -XX:CMSFullGCsBeforeCompaction=1 -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:+UseCMSCompactAtFullCollection -Dkerberos.domain.name=hadoop.hadoop.com -verbose:gc -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:/var/log/Bigdata//flume-client-1/flume/flume-root-20190226134231-%p-gc.log -Dproc_org.apache.flume.node.Application -Dproc_name=client -Dproc_conf_file=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf/properties.properties -Djava.security.krb5.conf=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf//krb5.conf -Djava.security.auth.login.config=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf//jaas.conf -Dzookeeper.server.principal=zookeeper/hadoop.hadoop.com -Dzookeeper.request.timeout=120000 -Dflume.instance.id=884174180 -Dflume.agent.name=clientName1 -Dflume.role=client -Dlog4j.configuration.watch=true -Dlog4j.configuration=log4j.properties -Dflume_log_dir=/var/log/Bigdata//flume-client-1/flume/ -Dflume.service.id=flume-client-1 -Dbeetle.application.home.path=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf/service -Dflume.called.from.service -Dflume.conf.dir=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf -Dflume.metric.conf.dir=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf -Dflume.script.home=/opt/FlumeClient/fusioninsight-flume-1.9.0/bin -cp '/opt/FlumeClient/fusioninsight-flume-1.9.0/conf:/opt/FlumeClient/fusioninsight-flume-1.9.0/lib/*:/opt/FlumeClient/fusioninsight-flume-1.9.0/conf/service/' -Djava.library.path=/opt/FlumeClient/fusioninsight-flume-1.9.0/plugins.d/native/native org.apache.flume.node.Application --conf-file /opt/FlumeClient/fusioninsight-flume-1.9.0/conf/properties.properties --name client
   /opt/FlumeClient/fusioninsight-flume-1.9.0/bin/flume-ng: line 233: /tmp/FusionInsight-Client/Flume/FusionInsight_Flume_ClientConfig/JDK/jdk-8u18/bin/java: No such file or directory

解决办法

步骤1 适当调大堆内存（xmx）的值。

步骤2 与正常启动Flume的节点进行文件和文件夹权限对比，更改错误文件或文件夹权限。

步骤3 重新配置JAVA_HOME。客户端替换${install_home}/fusioninsight-flume-*flume组件版本号*/conf/ENV_VARS文件中JAVA_HOME的值，服务端替换etc目录下ENV_VARS文件中JAVA_HOME的值。

其中JAVA_HOME的值可通过登录正常启动Flume的节点，执行**echo ${JAVA_HOME}**获取。

说明

${install_home}为Flume客户端的安装路径。

----结束

16.8 使用 HBase

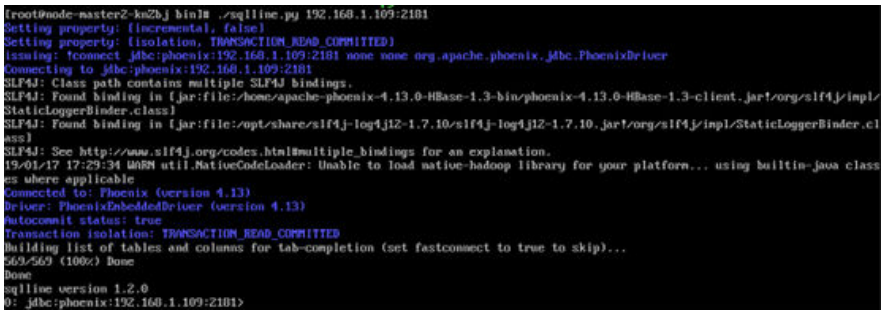
16.8.1 连接到 HBase 响应慢

用户问题

在相同的vpc网络下，外部集群通过Phoenix连接到HBase响应慢。

问题现象

在相同的vpc下，外部集群通过Phoenix连接到HBase时，响应太慢。



原因分析

客户配置了DNS服务，由于客户端连接到HBase先通过DNS来解析服务器端，所以导致响应慢。

处理步骤

步骤1 以root用户登录Master节点。

步骤2 执行**vi /etc/resolv.conf**，打开resolv.conf文件，注释掉DNS服务器地址。例如，#1.1.1.1

----结束

16.8.2 HBase 用户认证失败

用户问题

HBase用户认证失败。

问题现象

客户侧HBase用户认证失败，报错信息如下：

```
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: login failed with hbaseuser and /usr/local/linoseyc/hbase-tomcat/webapps/bigdata_hbase/WEB-INF/classes/user.keytab.
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: perhaps cause 1 is (wrong password) keytab file and user not match, you can kinit -k -t keytab user in client server to check.
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: perhaps cause 2 is (clock skew) time of local server and remote server not match, please check ntp to remote server.
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: perhaps cause 3 is (aes256 not support) aes256 not support by default jdk/jre, need copy local_policy.jar and US_export_policy.jar from remote server in path ${BIGDATA_HOME}/jdk/jre/lib/security.
```

原因分析

客户使用的JDK中的jar包与MRS服务认证的jar包版本不一致。

处理步骤

步骤1 以root登录集群Master1节点。

步骤2 执行如下命令，查看MRS服务认证的jar包。

ll /opt/share/local_policy/local_policy.jar

ll /opt/Bigdata/jdk{version}/jre/lib/security/local_policy.jar

步骤3 将步骤2中的jar包下载到本地。

步骤4 将下载的jar包替换到本地JDK目录/opt/Bigdata/jdk/jre/lib/security。

步骤5 执行**cd /opt/client/HBase/hbase/bin**命令，进入到HBase的bin目录。

步骤6 执行**sh start-hbase.sh**命令，重启HBase组件。

----结束

16.8.3 端口被占用导致 RegionServer 启动失败

问题现象

Manager页面监控发现RegionServer状态为Restoring。

原因分析

1. 通过查看RegionServer日志（/var/log/Bigdata/hbase/rs/hbase-omm-xxx.log）。
2. 使用**lsof -i:21302**（MRS1.7.X及以后端口号是16020）查看到pid，然后根据pid查看到相应的进程，发现RegionServer的端口被DFSZkFailoverController占用。

3. 查看“/proc/sys/net/ipv4/ip_local_port_range”显示为“9000 65500”，临时端口范围与MRS产品端口范围重叠，因为安装时未进行preinstall操作。

解决办法

步骤1 执行**kill -9** *DFSZkFailoverController的pid*， 使得其重启后绑定其它端口，然后重启Restoring的RegionServer。

----结束

16.8.4 节点剩余内存不足导致 HBase 启动失败

问题现象

HBase的RegionServer服务一直是Restoring状态。

原因分析

1. 查看RegionServer的日志（“/var/log/Bigdata/hbase/rs/hbase-omm-XXX.out”），发现显示以下打印信息：
   There is insufficient memory for the Java Runtime Environment to continue.
2. 使用**free**指令查看，该节点确实没有足够内存。

解决办法

步骤1 现场进行排查内存不足原因，确认是否有某些进程占用过多内存，或者由于服务器自身内存不足。

----结束

16.8.5 HDFS 性能差导致 HBase 服务不可用告警

问题现象

HBase组件断断续续上报服务不可用告警。

原因分析

该问题多半为HDFS性能较慢，导致健康检查超时，从而导致监控告警。可通过以下方式判断：

1. 首先查看HMaster日志（“/var/log/Bigdata/hbase/hm/hbase-omm-xxx.log”），确认HMaster日志中没有频繁打印“system pause”或“jvm”等GC相关信息。
2. 然后可以通过下列三种方式确认原因为HDFS性能慢造成告警产生。
   a. 使用客户端验证，通过**hbase shell**进入hbase命令行后，执行**list**验证需要运行多久。
   b. 开启HDFS的debug日志，然后查看下层目录很多的路径（**hadoop fs -ls /XXX/XXX**），验证需要运行多久。
   c. 打印HMaster进程jstack：

      su - omm

jps

jstack pid

3. 如下图所示，Jstack显示一直卡在DFSClient.listPaths。

图 16-20 异常

```
java.lang.Thread.State: WAITING (on object monitor)
    at java.lang.Object.wait(Native Method)
    at java.lang.Object.wait(Object.java:503)
    at org.apache.hadoop.ipc.Client.call(Client.java:1396)
    - locked <0x00000000b9268a38> (a org.apache.hadoop.ipc.Client$Call)
    at org.apache.hadoop.ipc.Client.call(Client.java:1363)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:206)
    at com.sun.proxy.$Proxy13.getListing(Unknown Source)
    at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.getListing(ClientNamenodeProtocolTra
    at sun.reflect.GeneratedMethodAccessor24.invoke(Unknown Source)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:187)
    at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:102)
    at com.sun.proxy.$Proxy14.getListing(Unknown Source)
    at sun.reflect.GeneratedMethodAccessor24.invoke(Unknown Source)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.hbase.fs.HFileSystem$1.invoke(HFileSystem.java:294)
    at com.sun.proxy.$Proxy17.getListing(Unknown Source)
    at org.apache.hadoop.hdfs.DFSClient.listPaths(DFSClient.java:1767)
    at org.apache.hadoop.hdfs.DFSClient.listPaths(DFSClient.java:1750)
    at org.apache.hadoop.hdfs.DistributedFileSystem.listStatusInternal(DistributedFileSystem.java:691)
    at org.apache.hadoop.hdfs.DistributedFileSystem.access$600(DistributedFileSystem.java:102)
    at org.apache.hadoop.hdfs.DistributedFileSystem$15.doCall(DistributedFileSystem.java:753)
    at org.apache.hadoop.hdfs.DistributedFileSystem$15.doCall(DistributedFileSystem.java:749)
    at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
    at org.apache.hadoop.hdfs.DistributedFileSystem.listStatus(DistributedFileSystem.java:749)
    at org.apache.hadoop.fs.FileSystem.listStatus(FileSystem.java:1483)
```

解决办法

步骤1 如果确认是HDFS性能慢导致告警，需要排除是否为旧版本中Impala运行导致HDFS性能慢或者是否为集群最初部署时JournalNode部署不正确（部署过多，大于3个）。

----结束

16.8.6 参数不合理导致 HBase 启动失败

问题现象

修改部分参数后，无法正常启动HBase。

原因分析

1. 查看HMaster日志（/var/log/Bigdata/hbase/hm/hbase-omm-xxx.log）显示，hbase.regionserver.global.memstore.size + hfile.block.cache.size总和大于0.8导致启动不成功，因此需要调整参数配置值总和低于0.8。

```
ava HotSpot(TM) 64-Bit Server VM warning: ignoring option PermSize=128M; support was removed in 8.0
ava HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=128M; support was removed in 8.0
ava HotSpot(TM) 64-Bit Server VM warning: UseCMSCompactAtFullCollection is deprecated and will likely be removed in a future release.
ava HotSpot(TM) 64-Bit Server VM warning: CMSFullGCsBeforeCompaction is deprecated and will likely be removed in a future release.
NFO: Watching file:/opt/huawei/Bigdata/etc/3_14_RegionServer/log4j.properties for changes with interval : 60000
xception in thread "main" java.lang.RuntimeException: Current heap configuration for MemStore and BlockCache exceeds the threshold required for successful cluster operation. The combined value cannot exceed
0.8. Please check the settings for hbase.regionserver.global.memstore.size and hfile.block.cache.size in your configuration. hbase.regionserver.global.memstore.size is 0.6 hfile.block.cache.size is 0.25
        at org.apache.hadoop.hbase.io.util.HeapMemorySizeUtil.checkForClusterFreeMemoryLimit(HeapMemorySizeUtil.java:64)
        at org.apache.hadoop.hbase.HBaseConfiguration.addHbaseResources(HBaseConfiguration.java:82)
        at org.apache.hadoop.hbase.HBaseConfiguration.create(HBaseConfiguration.java:96)
        at org.apache.hadoop.hbase.regionserver.HRegionServer.main(HRegionServer.java:2663)
```

2. 查看HMaster和RegionServer的out日志（/var/log/Bigdata/hbase/hm/hbase-omm-xxx.out/var/log/Bigdata/hbase/rs/hbase-omm-xxx.out），提示Unrecognized VM option。

```
Unrecognized VM option
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. Program will exit.
```

检查GC_OPTS相关参数存在多余空格，如-D sun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFE。

解决办法

步骤1 针对memstore、cache修改配置参数后，重启HBase服务成功。

步骤2 针对GC_OPTS配置错误，修改参数后重启HBase服务成功。

----结束

16.8.7 残留进程导致 Regionsever 启动失败

问题现象

HBase服务启动失败，健康检查报错。

原因分析

查看启动HBase服务时manager页面的详细打印信息，提示the previous process is not quit。

解决办法

步骤1 登录节点，后台通过执行**ps -ef | grep HRegionServer**发现确实存在一个残留的进程。

步骤2 确认进程可以终止后，使用kill命令终止该进程（如果kill无法终止该进程，需要使用kill -9来强制终止该进程）。

步骤3 重新启动HBase服务成功。

----结束

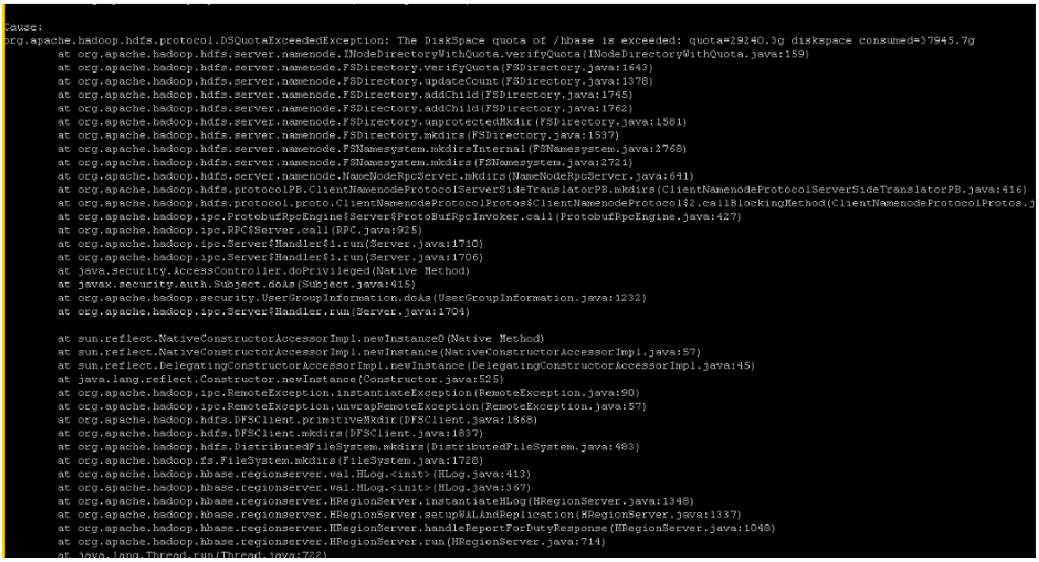
16.8.8 HDFS 上设置配额导致 HBase 启动失败

问题现象

HBase启动失败。

原因分析

查看HMaster日志信息（“/var/log/Bigdata/hbase/hm/hbase-omm-xxx.log”），出现如下异常，The DiskSpace quota of /hbase is exceeded。



解决办法

步骤1 通过后台使用**df -h**命令查看数据盘目录空间已满，因此需要删除无用的数据来进行应急恢复。

步骤2 后续需要扩容节点来解决数据目录空间不足问题。

----结束

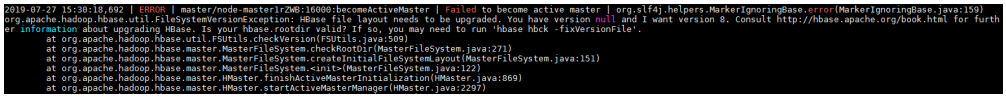
16.8.9 HBase version 文件损坏导致启动失败

问题背景

HBase启动失败。

原因分析

1. HBase启动时会读取hbase.version文件，但是日志显示读取存在异常。



2. 通过**hadoop fs -cat /hbase/hbase.version**命令发现文件不能正常查看，该文件损坏。

解决办法

步骤1 执行**hbase hbck -fixVersionFile**命令修复文件。

步骤2 如**步骤1**不能解决，从同版本的其他集群中获取hbase.version文件上传进行替换。

步骤3 重新启动HBase服务。

----结束

16.8.10 无业务情况下，RegionServer 占用 CPU 高

问题背景

无业务情况下，RegionServer占用CPU较高。

原因分析

1. 通过**top**命令获取RegionServer的进程使用CPU情况信息，查看CPU使用率高的进程号。
2. 根据RegionServer的进程编号，获取该进程下线程使用CPU情况。

   top -H -p <PID>（根据实际RegionServer的进程ID进行替换），具体如下图所示，发现部分线程CPU使用率均达到80%。

```
PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+ COMMAND
75706 omm     20   0 6879444   1.0g  25612 S  90.4  1.6   0:00.00 java
75716 omm     20   0 6879444   1.0g  25612 S  90.4  1.6   0:04.74 java
75720 omm     20   0 6879444   1.0g  25612 S  88.6  1.6   0:01.93 java
75721 omm     20   0 6879444   1.0g  25612 S  86.8  1.6   0:01.99 java
75722 omm     20   0 6879444   1.0g  25612 S  86.8  1.6   0:01.94 java
75723 omm     20   0 6879444   1.0g  25612 S  86.8  1.6   0:01.96 java
75724 omm     20   0 6879444   1.0g  25612 S  86.8  1.6   0:01.97 java
75725 omm     20   0 6879444   1.0g  25612 S  81.5  1.6   0:02.06 java
75726 omm     20   0 6879444   1.0g  25612 S  79.7  1.6   0:02.01 java
75727 omm     20   0 6879444   1.0g  25612 S  79.7  1.6   0:01.95 java
75728 omm     20   0 6879444   1.0g  25612 S  78.0  1.6   0:01.99 java
```

3. 根据RegionServer的进程编号，获取线程堆栈信息。

   jstack 12345 >allstack.txt（根据实际RegionServer的进程ID进行替换）
4. 将需要的线程ID转换为16进制格式：

   printf "%x\n" 30648

   输出结果TID为77b8。
5. 根据输出16进制TID，在线程堆栈中进行查找，发现在执行compaction操作。

```
"regionserver/ahbd-hbase-dat1/12.2.168.1:21302-longCompactions-1482676601478" #1641 prio=5 os_prio=0 tid=0x00007fa614563000 nid=0x77b8 runnable [0x0
   java.lang.Thread.State: RUNNABLE
    at org.apache.hadoop.io.compress.snappy.SnappyCompressor.compressBytesDirect(Native Method)
    at org.apache.hadoop.io.compress.snappy.SnappyCompressor.compress(SnappyCompressor.java:228)
    at org.apache.hadoop.io.compress.BlockCompressorStream.compress(BlockCompressorStream.java:149)
    at org.apache.hadoop.io.compress.BlockCompressorStream.finish(BlockCompressorStream.java:142)
    at org.apache.hadoop.hbase.io.encoding.HFileBlockDefaultEncodingContext.compressAfterEncoding(HFileBlockDefaultEncodingContext.java:219)
    at org.apache.hadoop.hbase.io.encoding.HFileBlockDefaultEncodingContext.compressAndEncrypt(HFileBlockDefaultEncodingContext.java:132)
    at org.apache.hadoop.hbase.io.hfile.HFileBlock$Writer.finishBlock(HFileBlock.java:989)
    at org.apache.hadoop.hbase.io.hfile.HFileBlock$Writer.ensureBlockReady(HFileBlock.java:961)
    at org.apache.hadoop.hbase.io.hfile.HFileBlock$Writer.finishBlockAndWriteHeaderAndData(HFileBlock.java:1077)
```

6. 对其它线程执行相同操作，发现均为compactions线程。

```
"regionserver/ahbd-hbase-dat1/12.2.168.1:21302-longCompactions-1482676601473" #1629 prio=5 os_prio=0 tid=0x00007fa61454d800 nid=0x77a0 runnable
   java.lang.Thread.State: RUNNABLE
    at org.apache.hadoop.hdfs.DFSOutputStream.writeChunk(DFSOutputStream.java:425)
    - locked <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSOutputSummer.writeChecksumChunks(FSOutputSummer.java:214)
    at org.apache.hadoop.fs.FSOutputSummer.flushBuffer(FSOutputSummer.java:165)
    - locked <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSOutputSummer.flushBuffer(FSOutputSummer.java:146)
    - eliminated <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSOutputSummer.write1(FSOutputSummer.java:137)
    at org.apache.hadoop.fs.FSOutputSummer.write(FSOutputSummer.java:112)
    - locked <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSDataOutputStream$PositionCache.write(FSDataOutputStream.java:58)
    at java.io.DataOutputStream.write(DataOutputStream.java:107)
    - locked <0x00000004de9535c8> (a org.apache.hadoop.hdfs.client.HdfsDataOutputStream)
    at java.io.FilterOutputStream.write(FilterOutputStream.java:97)
```

解决办法

属于正常现象。

发现消耗CPU较高线程均为HBase的compaction，其中部分线程调用Snappy压缩处理，部分线程调用HDFS读写数据。当前每个Region数据量和数据文件多，且采用Snappy压缩算法，因此执行compaction时会使用大量CPU导致CPU较高。

定位办法

步骤1 使用**top**命令查看 CPU使用率高的进程号。

步骤2 查看此进程中占用CPU高的线程。

使用命令**top -H -p <PID>**即可打印出某进程<PID>下的线程的CPU耗时信息。

一般某个进程如果出现问题，是因为某个线程出现问题了，获取查询到的占用CPU最高的线程号。

或者使用命令**ps -mp <PID> -o THREAD,tid,time | sort -rn**。

观察回显可以得到CPU最高的线程号。

步骤3 获取出现问题的线程的堆栈。

java问题使用jstack工具是最有效，最可靠的。

到java/bin目录下有 jstack工具，获取进程堆栈，并输出到本地文件。

jstack <PID> > allstack.txt

获取线程堆栈，并输出到本地文件。

步骤4 将需要的线程ID转换为16进制格式。

printf "%x\n" <PID>

回显结果为线程ID，即 TID。

步骤5 使用命令获得TID,并输出到本地文件。

jstack <PID> | grep <TID> > Onestack.txt

如果只是在命令行窗口查看，可以使用命名：

jstack <PID> | grep <TID> -A 30

-A 30意思是显示30行。

----结束

16.8.11 HBase 启动失败，RegionServer 日志中提示 FileNotFoundException 异常

问题背景

HBase启动失败，RegionServer一直处于Restoring状态。

原因分析

1. 查看RegionServer的日志（/var/log/Bigdata/hbase/rs/hbase-omm-XXX.log），发现显示以下打印信息：

```
| ERROR | RS_OPEN_REGION-ab-dn01:21302-2 | ABORTING region server ab-dn01,21302,1487663269375: The coprocessor org.apache.kylin.storage.hbase.cube.v2.coprocessor.endpoint.CubeVisitService threw java.io.FileNotFoundException: File does not exist: hdfs://hacluster/kylin/kylin_metadata/coprocessor/kylin-coprocessor-1.6.0-SNAPSHOT-0.jar |
org.apache.hadoop.hbase.regionserver.HRegionServer.abort(HRegionServer.java:2123)
java.io.FileNotFoundException: File does not exist: hdfs://hacluster/kylin/kylin_metadata/coprocessor/
```

```
kylin-coprocessor-1.6.0-SNAPSHOT-0.jar
at org.apache.hadoop.hdfs.DistributedFileSystem$25.doCall(DistributedFileSystem.java:1399)
at org.apache.hadoop.hdfs.DistributedFileSystem$25.doCall(DistributedFileSystem.java:1391)
at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
at org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(DistributedFileSystem.java:1391)
at org.apache.hadoop.fs.FileUtil.copy(FileUtil.java:340)
at org.apache.hadoop.fs.FileUtil.copy(FileUtil.java:292)
at org.apache.hadoop.fs.FileSystem.copyToLocalFile(FileSystem.java:2038)
at org.apache.hadoop.fs.FileSystem.copyToLocalFile(FileSystem.java:2007)
at org.apache.hadoop.fs.FileSystem.copyToLocalFile(FileSystem.java:1983)
at org.apache.hadoop.hbase.util.CoprocessorClassLoader.init(CoprocessorClassLoader.java:168)
at
org.apache.hadoop.hbase.util.CoprocessorClassLoader.getClassLoader(CoprocessorClassLoader.java:250)
at org.apache.hadoop.hbase.coprocessor.CoprocessorHost.load(CoprocessorHost.java:224)
at
org.apache.hadoop.hbase.regionserver.RegionCoprocessorHost.loadTableCoprocessors(RegionCoprocess
orHost.java:365)
at
org.apache.hadoop.hbase.regionserver.RegionCoprocessorHost.<init>(RegionCoprocessorHost.java:227)
at org.apache.hadoop.hbase.regionserver.HRegion.<init>(HRegion.java:783)
at org.apache.hadoop.hbase.regionserver.HRegion.<init>(HRegion.java:689)
at sun.reflect.GeneratedConstructorAccessor22.newInstance(Unknown Source)
at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:4
5)
at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
at org.apache.hadoop.hbase.regionserver.HRegion.newHRegion(HRegion.java:6312)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6622)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6594)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6550)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6501)
at
org.apache.hadoop.hbase.regionserver.handler.OpenRegionHandler.openRegion(OpenRegionHandler.ja
va:363)
at
org.apache.hadoop.hbase.regionserver.handler.OpenRegionHandler.process(OpenRegionHandler.java:12
9)
at org.apache.hadoop.hbase.executor.EventHandler.run(EventHandler.java:129)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
```

2. 使用客户端hdfs命令查看，如下文件不存在。

   hdfs://hacluster/kylin/kylin_metadata/coprocessor/kylin-coprocessor-1.6.0-SNAPSHOT-0.jar

3. HBase在配置协处理器时，一定要保证对应的jar包路径没有问题，否则HBase会无法启动。

解决办法

使用Kylin对接MRS，确保Kylin相关jar包存在。

16.8.12 HBase 启动后原生页面显示 RegionServer 个数多于实际个数

问题背景

HBase启动后，HMaster原生页面显示RegionServer个数多于实际RegionServer个数。

查看HMaster原生页面，显示有4个RegionServer在线，如下图示：

ServerName	Start time	Requests Per Second	Num. Regions
controller-192-168-1-1,21302,1494933959261	Tue May 16 19:25:59 CST 2017	0	19
controller-192-168-1-2,21302,1494933957536	Tue May 16 19:25:57 CST 2017	0	24
controller-192-168-1-3,21302,1494933958592	Tue May 16 19:25:58 CST 2017	0	16
eth0,21302,1494933958592	Tue May 16 19:25:58 CST 2017	0	0
Total:4		0	59

原因分析

如下图可以看出，第三行hostname为controller-192-168-1-3节点和第四行hostname为eth0节点为同一RegionServer上报的信息，登录相应节点，查看/etc/hosts文件，发现，对应同一ip，配置两个hostname。如下：

```
# special IPv6 addresses
::1             localhost ipv6-localhost ipv6-loopback

fe00::0         ipv6-localnet

ff00::0         ipv6-mcastprefix
ff02::1         ipv6-allnodes
ff02::2         ipv6-allrouters
ff02::3         ipv6-allhosts
11.1.1.3        eth2 eth2
#192.168.1.3     eth0 eth0
192.168.2.3     eth1 eth1
10.130.87.37    eth3 eth3
192.168.1.102 controller
1.1.1.1 hadoop.hadoop.com
192.168.1.2 controller-192-168-1-2
192.168.1.1 controller-192-168-1-1
192.168.1.3 controller-192-168-1-3
```

解决办法

登录RegionServer所在节点，修改/etc/hosts文件，同一ip只能对应同一hostname。

16.8.13 RegionServer 实例异常，处于 Restoring 状态

问题背景

HBase启动失败，RegionServer一直处于Restoring状态。

原因分析

查看异常的RegionServer实例的运行日志（/var/log/Bigdata/hbase/rs/hbase-omm-XXX.log），发现显示以下打印信息ClockOutOfSyncException...，Reported time is too far out of sync with master

```
2017-09-18 11:16:23,636 | FATAL | regionserver21302 | Master rejected startup because clock is out of sync |
org.apache.hadoop.hbase.regionserver.HRegionServer.reportForDuty(HRegionServer.java:2059)
org.apache.hadoop.hbase.ClockOutOfSyncException: org.apache.hadoop.hbase.ClockOutOfSyncException:
Server nl-bi-fi-datanode-24-65,21302,1505726180086 has been rejected; Reported time is too far out of
sync with master.  Time difference of 152109ms > max allowed of 30000ms
at org.apache.hadoop.hbase.master.ServerManager.checkClockSkew(ServerManager.java:354)
...

...
2017-09-18 11:16:23,858 | ERROR | main | Region server exiting |
org.apache.hadoop.hbase.regionserver.HRegionServerCommandLine.start(HRegionServerCommandLine.java:
70)
java.lang.RuntimeException: HRegionServer Aborted
```

该日志说明异常的RegionServer实例和HMaster实例的时差大于允许的时差值30s（由参数hbase.regionserver.maxclockskew控制，默认30000ms），导致RegionServer实例异常。

解决办法

调整异常节点时间，确保节点间时差小于30s。

16.8.14 新安装的集群 HBase 启动失败

问题背景

新安装的集群HBase启动失败，查看RegionServer日志报如下错误：

```
2018-02-24 16:53:03,863 | ERROR | regionserver/host3/187.6.71.69:21302 | Master passed us a different
hostname to use; was=host3, but now=187-6-71-69 |
org.apache.hadoop.hbase.regionserver.HRegionServer.handleReportForDutyResponse(HRegionServer.java:138
6)
```

原因分析

/etc/hosts中同一个IP地址配置了多个主机名映射关系。

解决办法

步骤1 修改/etc/host中IP与主机名的映射关系，配置正确。

步骤2 重新启动HBase组件。

----结束

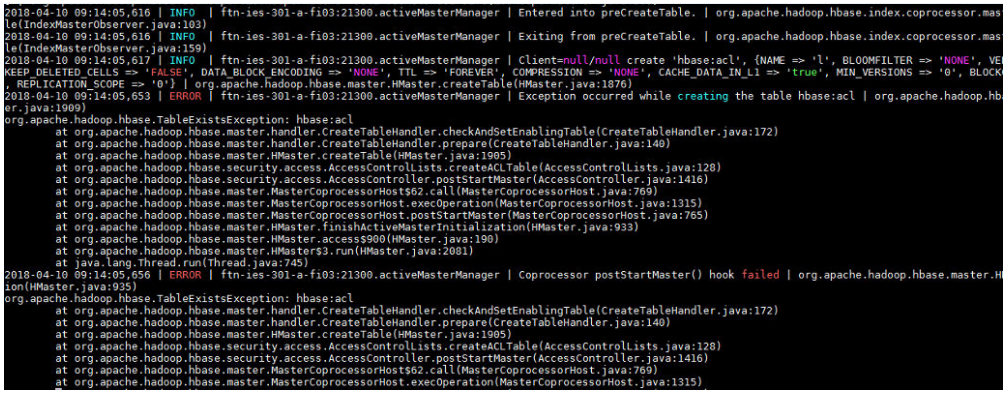
16.8.15 acl 表目录丢失导致 HBase 启动失败

问题背景与现象

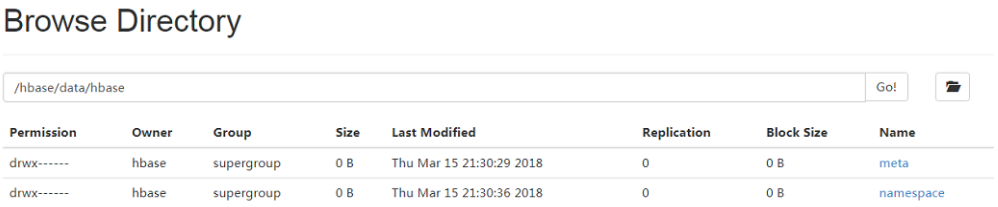
集群HBase启动失败

原因分析

1. 查看HBase的HMaster日志，报如下错误：



2. 检查HDFS上HBase的路径发现acl表路径丢失。



解决办法

步骤1 停止HBase组件。

步骤2 在HBase客户端使用hbase用户登录认证，执行如下命令。

例如：

```
hadoop03:~ # source /opt/client/bigdata_env
hadoop03:~ # kinit hbase
Password for hbase@HADOOP.COM:
hadoop03:~ # hbase zkcli
```

步骤3 删除zk中acl表信息。

例如：

```
[zk: hadoop01:24002,hadoop02:24002,hadoop03:24002(CONNECTED) 0] deleteall /hbase/table/hbase:acl
[zk: hadoop01:24002,hadoop02:24002,hadoop03:24002(CONNECTED) 0] deleteall /hbase/table-lock/
hbase:acl
```

步骤4 启动HBase组件。

----结束

16.8.16 集群上下电之后 HBase 启动失败

问题背景与现象

集群的ECS关机重启后，HBase启动失败。

原因分析

查看HMaster的运行日志，发现有报大量的如下错误：

```
2018-03-26 11:10:54,185 | INFO | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
```

```
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:00,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:06,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:10,787 | INFO  | RpcServer.reader=9,bindAddress=hadoopc1h3,port=21300 | Kerberos
principal name is hbase/hadoop.hadoop.com@HADOOP.COM | org.apache.hadoop.hbase
.ipc.RpcServer$Connection.readPreamble(RpcServer.java:1532)
2018-03-26 11:11:12,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:18,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
```

节点上下电，RegionServer的wal分裂失败导致。

解决办法

步骤1 停止HBase组件。

步骤2 通过hdfs fsck命令检查/hbase/WALs文件的健康状态。

hdfs fsck /hbase/WALs

输出如下表示文件都正常，如果有异常则需要先处理异常的文件，再执行后面的操作。

```
The filesystem under path '/hbase/WALs' is HEALTHY
```

步骤3 备份/hbase/WALs文件。

hdfs dfs -mv /hbase/WALs /hbase/WALs_old

步骤4 新建/hbase/WALs目录。

hdfs dfs -mkdir /hbase/WALs

必须保证路径权限是hbase:hadoop。

步骤5 启动HBase组件。

----结束

16.8.17 文件块过大导致 HBase 数据导入失败

问题现象

导入数据到hbase报错：NotServingRegionException。

原因分析

当一个block size大于2G时，hdfs在seek的时候会出现读取异常，持续频繁写入regionserver时出现了full gc，且时间比较久，导致hmaster与regionserver之间的心跳异常，然后hmaster把regionserver标记为dead状态，强制重启了Regionserver，重启后触发servercrash机制开始回滚wal日志。现在这个splitwal的文件已经达到将近2.1G，且其仅有一个block块，导致hdfs seek异常，引起splitwal失败，regionserver检测到当前这个wal日志还需要split，又会触发splitwal日志的机制进行回滚，就这样在split与split失败之间不停循环，导致无法上线该regionserver节点上的region，最后出现查询该RS上某一个region时会报region not online的异常。

处理步骤

步骤1 在“HMaster Web UI”右侧，单击“HMaster (主)”进入HBase Web UI界面。

步骤2 在“Procedures”页签查看问题节点。

步骤3 以root用户登录问题节并执行**hdfs dfs -ls**命令查看所有块信息。

步骤4 执行**hdfs dfs -mkdir**命令新建目录用于存放问题块。

步骤5 执行**hdfs dfs -mv**将问题块转移至新建目录位置。

----结束

建议与总结

以下两点可供参考：

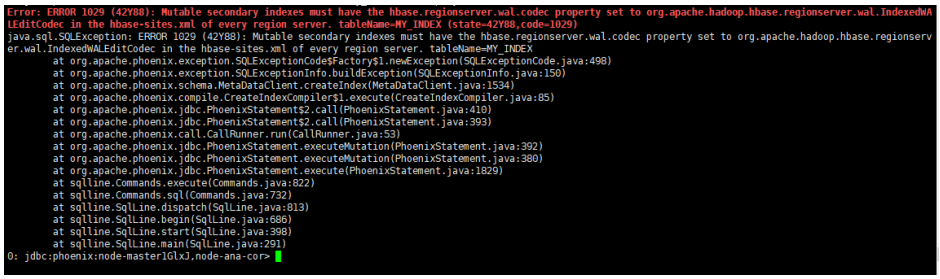
- 数据块损坏，通过**hdfs fsck /tmp -files -blocks -racks**命令检查block数据块健康信息。
- region正在分裂时对数据的操作会抛NotServingRegionException异常。

16.8.18 使用 Phoenix 创建 HBase 表后，向索引表中加载数据报错

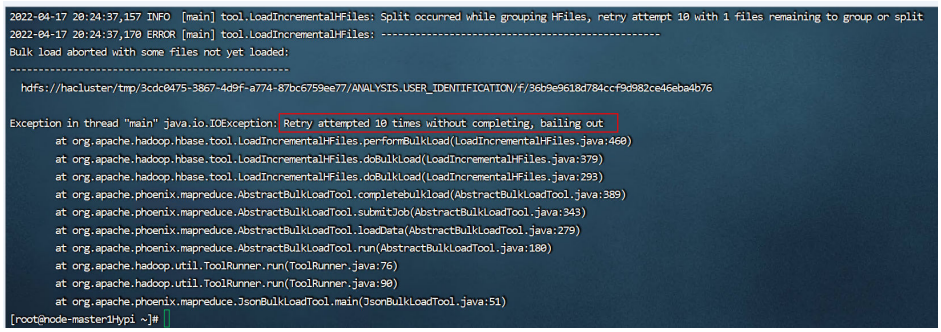
问题背景与现象

使用Phoenix创建HBase表后，使用命令向索引表中加载数据报错：

- MRS 2.x及之前版本：Mutable secondary indexes must have the hbase.regionserver.wal.codec property set to org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec in the hbase-sites.xml of every region server. tableName=MY_INDEX (state=42Y88,code=1029)



- MRS 3.x及之后版本：Exception in thread "main" java.io.IOException: Retry attempted 10 times without completing, bailing out



处理步骤

步骤1 MRS 2.x及之前版本，操作步骤如下：

1. 使用**admin**用户登录MRS Manager界面，选择“服务管理 > HBase > 服务配置”，将“参数类别”的“基础配置”切换为“全部配置”，选择“HMaster > 自定义”，给参数“hbase.hmaster.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项。
2. 选择“RegionServer > 自定义”，给参数“hbase.regionserver.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项，单击“保存配置”，输入当前用户密码，单击“确定”，保存配置。
3. 单击“服务状态”，选择“更多 > 重启服务”，输入当前用户密码，单击“确定”，重启HBase服务。

步骤2 MRS 3.x及之后版本，操作步骤如下：

1. 使用**admin**用户登录FusionInsight Manager，选择“集群 > 服务 > HBase > 配置 > 全部配置 > RegionServer > 自定义”，给参数“hbase.regionserver.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项。
2. 选择“HMaster > 自定义”，给参数“hbase.hmaster.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项。
3. 单击“保存”，在弹出的对话框中单击“确定”，保存配置。

4. 单击“概览”，选择“更多 > 重启服务”，输入当前用户密码，单击“确定”，重启HBase服务。

----结束

16.8.19 在 MRS 集群客户端无法执行 hbase shell 命令

用户问题

在MRS集群客户端无法执行**hbase shell**命令。

原因分析

- 执行**hbase shell**命令前未配置环境变量。
- 当前MRS集群未安装HBase客户端。

处理步骤

步骤1 使用**root**用户登录安装客户端的节点，切换到客户端安装目录，查看是否安装了HBase客户端。

- 是，执行**步骤2**。
- 否，下载并安装HBase客户端。

步骤2 执行以下命令配置环境变量。

source bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

步骤4 执行HBase组件的客户端命令。

hbase shell

----结束

16.8.20 HBase shell 客户端在使用中有 INFO 信息打印在控制台导致显示混乱

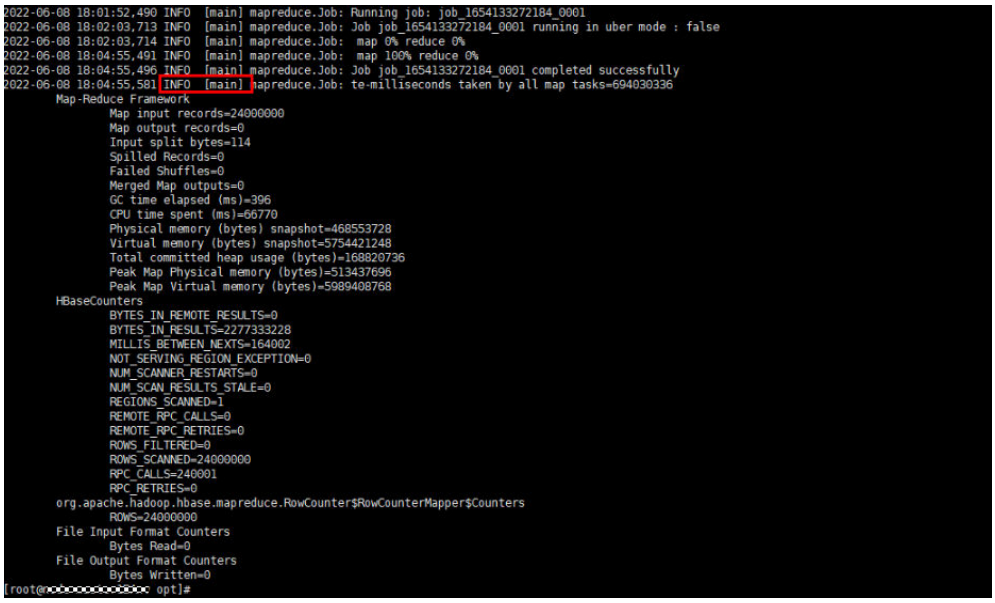
用户问题

在使用HBase shell中会有INFO信息打印在控制台，导致显示混乱，用户在输入命令前需要按回车符清理控制台。如下图所示：

原因分析

- HBase客户端中默认日志打印设置为“INFO,console”，所以在使用期间会有INFO日志输出到控制台，影响HBase shell 窗口的显示。

- HBase客户端命令繁多，例如：**hbase shell**、**hbase hbck**、**hbase org.apache.hadoop.hbase.mapreduce.RowCounter**等，且后续还会增加。部分命令的输出为INFO打印，如果直接把INFO关闭会导致部分命令输出结果丢失。例如：RowCounter输出结果为INFO类型：



处理步骤

步骤1 使用**root**用户登录安装HBase客户端的节点。

步骤2 在“*HBase客户端安装目录*/HBase/component_env”文件中添加如下信息：

export HBASE_ROOT_LOGGER=INFO,RFA

把日志输出到日志文件中，后期如果使用**hbase org.apache.hadoop.hbase.mapreduce.RowCounter**等命令，执行结果请在日志文件“*HBase客户端安装目录*/HBase/hbase/logs/hbase.log”中查看。

步骤3 切换到HBase客户端安装目录，执行以下命令使配置生效。

cd *HBase客户端安装目录*

source HBase/component_env

----结束

16.8.21 RegionServer 剩余内存不足导致 HBase 服务启动失败

用户问题

RegionServer剩余内存不足导致HBase服务启动失败。

原因分析

RegionServer启动时节点剩余内存不足，导致无法启动实例。排查步骤如下：

1. 登录Master节点，到“/var/log/Bigdata”查找HBase相关日志，HMaster的日志中报错“connect regionserver timeout ”。

2. 登录到1中HMaster连接不上的RegionServer节点，到“/var/log/Bigdata”查找HBase相关日志，RegionServer报错“error=’Cannot allocate memory’(errno=12)”。
3. 根据2报错判断由于RegionServer内存不足导致RegionServer启动失败。

处理步骤

步骤1 登录报错的RegionServer节点，执行以下命令查看节点剩余内存：

free -g

步骤2 执行**top**命令查看节点内存使用情况。

步骤3 根据**top**提示结束内存占用多的进程（内存占用多并且非MRS自身组件的进程），并重新启动HBase服务。

说明

集群的Core节点除了MRS组件运行占用外，Yarn上的作业还会被分配到节点运行，占用节点内存。若是由于Yarn作业占用内存多导致组件无法正常启动时，建议扩容Core节点。

----结束

16.9 使用 HDFS

16.9.1 修改集群 HDFS 服务的 NameNode RPC 端口后，NameNode 都变为备状态

用户问题

通过页面更改NameNode的RPC端口，随后重启HDFS服务，出现所有NameNode一直是备状态，导致集群异常。

问题现象

所有NameNode都是备状态，导致集群异常。

原因分析

集群安装启动后，如果修改NameNode的RPC端口，则需要重新格式化Zkfc服务来更新zookeeper上的节点信息。

处理步骤

步骤1 登录Manager，停止HDFS服务。

说明

在停止HDFS时，建议不要停止相关服务。

步骤2 停止成功后，登录到被修改了RPC端口的Master节点。

说明

如果两个Master节点都被修改了RPC端口，则只需登录其中一个修改即可。

步骤3 执行**su - omm**命令切换到omm用户。

说明

如果是安全集群，需要执行**kinit hdfs**命令进行认证。

步骤4 执行如下命令，将环境变量脚本加载到环境中。

cd ${BIGDATA_HOME}/*MRS_X.X.X*/*1_8*_**Zkfc/etc**

source ${BIGDATA_HOME}/*MRS_X.X.X***/install/FusionInsight-Hadoop-3.1.1/hadoop/sbin/exportENV_VARS.sh**

说明

命令中的“*MRS_X.X.X*”和“1_8”根据实际版本而定。

步骤5 加载完成后，执行如下命令，格式化Zkfc。

cd ${HADOOP_HOME}/bin

./hdfs zkfc -formatZK

步骤6 格式化成功后，在Manager页面“重启”HDFS服务。

说明

如果更改了NameNode的RPC端口，则之前安装的所有客户端都需要刷新配置文件。

----结束

16.9.2 通过公网 IP 连接主机，使用 HDFS 客户端报错

用户问题

通过公网IP连接主机，不能使用HDFS客户端，运行HDFS提示**-bash: hdfs: command not found**。

问题现象

通过公网IP连接主机，不能使用HDFS客户端，运行HDFS提示**-bash: hdfs: command not found**。

原因分析

用户登录Master节点执行命令之前，未设置环境变量。

处理步骤

步骤1 以**root**用户登录任意一个Master节点。

步骤2 执行**source /opt/client/bigdata_env**命令，设置环境变量。

步骤3 执行**hdfs**命令即可成功使用HDFS客户端。

----结束

16.9.3 使用 Python 远程连接 HDFS 的端口失败

用户问题

使用Python远程连接HDFS的端口失败，如何解决?

问题现象

用户使用Python远程连接HDFS的50070端口失败。

原因分析

HDFS开源3.0.0以下版本的默认端口为50070，3.0.0及以上的默认端口为9870。用户使用的端口和HDFS版本不匹配导致连接端口失败。

步骤1 登录集群的主Master节点。

步骤2 执行**su - omm**命令，切换到omm用户。

步骤3 执行**/opt/Bigdata/om-0.0.1/sbin/queryVersion.sh**命令，查看集群中的HDFS版本号。

根据版本号确认开源组件的端口号。

步骤4 执行**netstat -anp|grep ${port}**命令，查看组件的默认端口号是否存在。

如果不存在，说明用户修改了默认的端口号。请修改为默认端口，再重新连接HDFS。

如果存在，请联系技术服务。

说明

- ${ port }：表示与组件版本相对应的组件默认端口号。
- 如果用户修改了默认端口号，请使用修改后的端口号连接HDFS。不建议修改默认端口号。

----结束

16.9.4 HDFS 容量使用达到 100%，导致上层服务 HBase、Spark 等上报服务不可用

用户问题

集群的HDFS容量使用达到100%，HDFS服务状态为只读，导致上层服务HBase、Spark等上报服务不可用告警。

问题现象

HDFS使用容量100%，磁盘容量只使用85%左右，HDFS服务状态为只读，导致上层服务HBase、Spark等上报服务不可用。

原因分析

当前NodeManager和DataNode共数据盘使用，MRS默认预留15%的数据磁盘空间给非HDFS使用，可通过HDFS参数**dfs.datanode.du.reserved.percentage**修改百分比来控制具体的磁盘占比。

当HDFS磁盘使用100%之后，可通过降低**dfs.datanode.du.reserved.percentage**百分比来恢复业务，再进行磁盘扩容。

处理步骤

步骤1 登录集群任意Master节点。

步骤2 执行**source /opt/client/bigdata_env**命令，初始化环境变量。

说明

如果是安全集群，则需要执行**kinit -kt <keytab file> <pricipal name>** 进行认证。

步骤3 执行**hdfs dfs –put ./startDetail.log /tmp**命令，测试HDFS写文件失败。

```
19/05/12 10:07:32 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /tmp/startDetail.log._COPYING_ could only be replicated to 0 nodes instead of minReplication (=1).  There are 3 datanode(s) running and no node(s) are excluded in this operation.
```

步骤4 执行**hdfs dfsadmin -report**命令，检查HDFS使用容量，发现已经达到100%。

```
Configured Capacity: 5389790579100 (4.90 TB)
Present Capacity: 5067618628404 (4.61 TB)
DFS Remaining: 133350196 (127.17 MB)
DFS Used: 5067485278208 (4.61 TB)
DFS Used%: 100.00%
Under replicated blocks: 10
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0
Pending deletion blocks: 0
```

步骤5 当HDFS使用容量达到100%时，通过HDFS参数**dfs.datanode.du.reserved.percentage**修改百分比来控制具体的磁盘占比。

1. 登录Manager进入服务配置页面。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > HDFS> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 >HDFS > 配置”。
2. 选择“全部配置”，在搜索框中搜索**dfs.datanode.du.reserved.percentage**。
3. 修改此参数的取值为“10”。

步骤6 修改完成后，扩容Core节点的磁盘个数。。

----结束

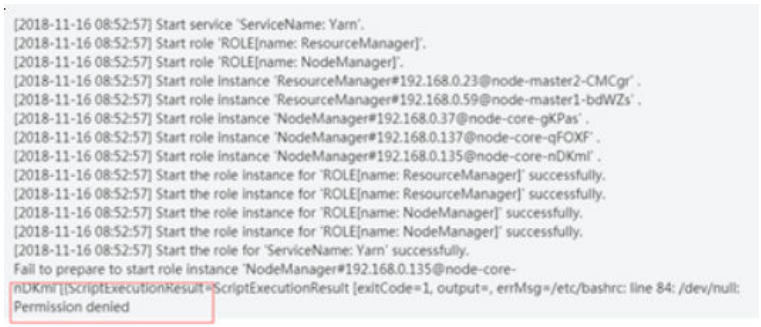
16.9.5 启动 HDFS 和 Yarn 报错

用户问题

启动HDFS和Yarn时报错。

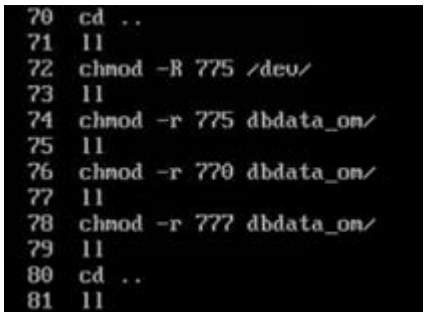
问题现象

无法启动HDFS、Yarn服务组件，报错内容：/dev/null Permission denied。



原因分析

客户修改了虚机系统的/dev/null的权限值为775。



处理步骤

步骤1 以root用户登录集群的任意一个Master节点。

步骤2 登录成功后，执行**chmod 666 /dev/null**命令，修改/dev/null的权限值为666。

步骤3 执行**ls -al /dev/null**命令，查看修改的/dev/null权限值是否为666，如果不是，需要修改为666。

步骤4 修改成功后，重新启动HDFS和Yarn组件。

----结束

16.9.6 HDFS 权限设置问题

用户问题

在使用MRS服务的时候，某个用户可以在其他用户的HDFS目录下面删除或者创建文件。

问题现象

在使用MRS服务时，某个用户可以在其他用户的HDFS目录下面删除或者创建文件。

原因分析

客户配置的用户具有ficommon组的权限，所以可以对HDFS任意操作。需要移除用户的ficommon组权限。

处理步骤

步骤1 以root用户登录集群的Master节点。

步骤2 执行**id ${用户名}**命令，显示用户组信息，确认是否有ficommon组权限。

如果存在ficommon组权限，请执行**步骤3**。如果不存在，请联系技术服务。

说明

${用户名}：出现HDFS权限设置问题的用户名。

步骤3 执行**gpasswd -d ${用户名} ficommon**命令，删除该用户的ficommon组权限。

说明

${用户名}：出现HDFS权限设置问题的用户名。

步骤4 执行成功后，登录Manager修改参数。

MRS Manager界面操作（适用MRS 3.x之前版本）：

1. 登录MRS Manager，在MRS Manager页面，选择“服务管理 > HDFS > 服务配置”。
2. “参数类别”选择“全部配置”，在搜索框中输入“dfs.permissions.enabled”，修改为“true”。
3. 修改完成后，单击“保存配置”，并重启HDFS服务。

FusionInsight Manager界面操作（适用MRS 3.x及之后版本）：

1. 登录FusionInsight Manager。选择“集群 > 服务 > HDFS > 配置 > 全部配置”。
2. 在搜索框中输入“dfs.permissions.enabled”，修改为“true”。
3. 修改完成后，单击“保存”，并重启HDFS服务。

MRS集群详情页操作：

1. 登录MRS控制台，选择“组件管理 > HDFS > 服务配置”。
2. “参数类别”选择“全部配置”，在搜索框中输入“dfs.permissions.enabled”，修改为“true”。
3. 修改完成后，单击“保存配置”，并重启HDFS服务。

----结束

16.9.7 HDFS 的 DataNode 一直显示退服中

用户问题

HDFS的DataNode一直显示退服中。

问题现象

HDFS的某个DataNode退服（或者对Core节点进行缩容）任务失败，但是DataNode在任务失败后一直处于退服中的状态。

原因分析

在对HDFS的某个DataNode进行退服（或者对core节点进行缩容）过程中，因为Master节点重启或者nodeagent进程意外退出等情况出现，使得退服（或缩容）任务失败，并且没有进行黑名单清理。此时DataNode节点会一直处于退服中的状态，需要人工介入进行黑名单清理。

处理步骤

步骤1 进入服务实例界面。

MRS Manager界面操作：

登录MRS Manager，在MRS Manager页面，选择“服务管理 > HDFS > 实例”。

FusionInsight Manager界面操作：

对于MRS 3.x及后续版本集群：也可登录FusionInsight Manager。选择“集群 > 服务 > HDFS > 实例”。

也可登录MRS控制台，选择“组件管理 > HDFS > 实例”。

步骤2 查看HDFS服务实例状态，找到一直处于退服中的DataNode，复制这个DataNode的IP地址。

步骤3 登录Master1节点的后台，执行**cd ${BIGDATA_HOME}/MRS_*/1_*_NameNode/etc/**命令进入黑名单目录。

步骤4 执行**sed -i "/^*IP*$/d" excludeHosts**命令清理黑名单中的故障DataNode信息，该命令中IP替换为**步骤2**中查询到的故障DataNode的IP地址，其中不能有空格。

步骤5 如果有两个Master节点，请在Master2节点上同样执行**步骤3**和**步骤4**。

步骤6 在Master1节点执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤7 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

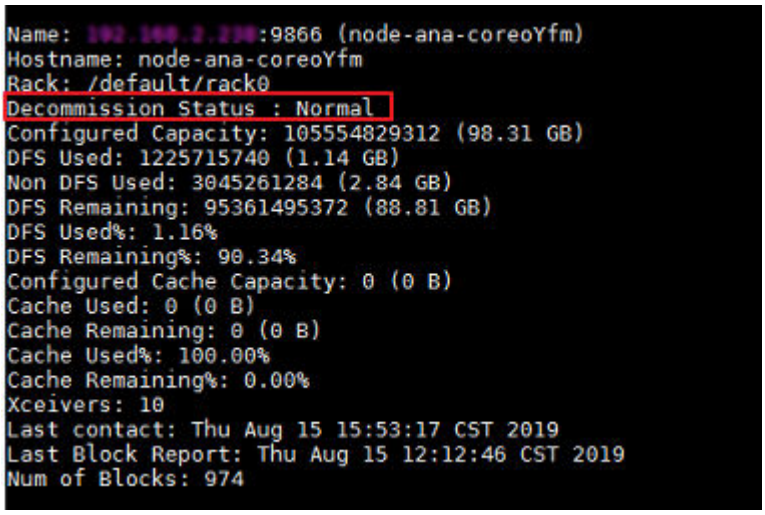
例如, **kinit admin**

步骤8 在Master1节点执行如下命令刷新HDFS的黑名单。

hdfs dfsadmin -refreshNodes

步骤9 使用命令**hdfs dfsadmin -report**来查看各个DataNode的状态，确认中查到的IP对应的DataNode已经恢复为**Normal**状态。

图 16-21 DataNode 的状态



步骤10 进入服务实例界面。

MRS Manager界面操作：

登录MRS Manager，在MRS Manager页面，选择“服务管理 > HDFS > 实例”。

FusionInsight Manager界面操作：

对于MRS 3.x及后续版本集群：可登录FusionInsight Manager。选择“集群 > 服务 > HDFS > 实例”。

登录MRS控制台，选择“组件管理 > HDFS > 实例”。

步骤11 勾选一直处于退服中的DataNode实例，单击“更多 > 重启实例”。

步骤12 等待重启完成，确认DataNode是否恢复正常。

----结束

建议与总结

尽量不要在退服（或缩容）过程中重启节点等高危操作。

参考信息

无

16.9.8 内存不足导致 HDFS 启动失败

问题背景与现象

重启HDFS后，HDFS的状态是Bad，且NameNode实例状态常常异常，并且花很久没有退出安全模式。

原因分析

1. 在NameNode运行日志（/var/log/Bigdata/hdfs/nn/hadoop-omm-namendoe-XXX.log）中搜索“WARN”，可以看到有大量时间在垃圾回收，如下例中耗时较长63s。

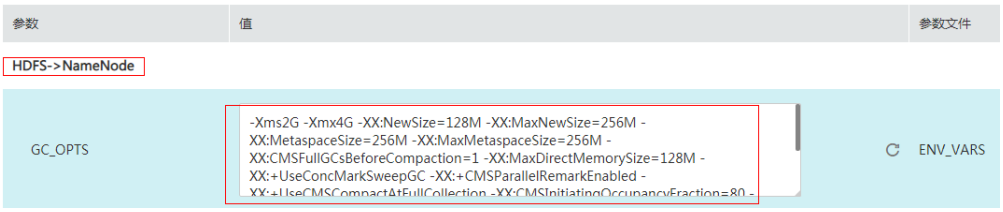
   ```
   2017-01-22 14:52:32,641 | WARN  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@1b39fd82 |
   Detected pause in JVM or host machine (eg GC): pause of approximately 63750ms
   GC pool 'ParNew' had collection(s): count=1 time=0ms
   GC pool 'ConcurrentMarkSweep' had collection(s): count=1 time=63924ms | JvmPauseMonitor.java:189
   ```

2. 分析NameNode日志“/var/log/Bigdata/hdfs/nn/hadoop-omm-namendoe-XXX.log”，可以看到NameNode在等待块上报，且总的Block个数过多，如下例中是3629万。

   ```
   2017-01-22 14:52:32,641 | INFO  | IPC Server handler 8 on 25000 | STATE* Safe mode ON.
   The reported blocks 29715437 needs additional 6542184 blocks to reach the threshold 0.9990 of total
   blocks 36293915.
   ```

3. 打开Manager页面，查看NameNode的GC_OPTS参数配置如下：

   图 16-22 查看 NameNode 的 GC_OPTS 参数配置

4. NameNode内存配置和数据量对应关系参考**表16-2**。

   表 16-2 NameNode 内存配置和数据量对应关系

文件对象数量	参考值
10,000,000	“-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M”
20,000,000	“-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G”
50,000,000	“-Xms32G -Xmx32G -XX:NewSize=2G -XX:MaxNewSize=3G”
100,000,000	“-Xms64G -Xmx64G -XX:NewSize=4G -XX:MaxNewSize=6G”
200,000,000	“-Xms96G -Xmx96G -XX:NewSize=8G -XX:MaxNewSize=9G”
300,000,000	“-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G”

解决办法

步骤1 按照规格修改NameNode的内存参数，如这里3600万block，将内存参数调整为“-Xms32G -Xmx32G -XX:NewSize=2G -XX:MaxNewSize=3G”。

步骤2 重启一个NameNode，确认该NameNode可以正常启动。

步骤3 重启另一个NameNode，确认页面状态恢复。

----结束

16.9.9 ntpdate 修改时间导致 HDFS 出现大量丢块

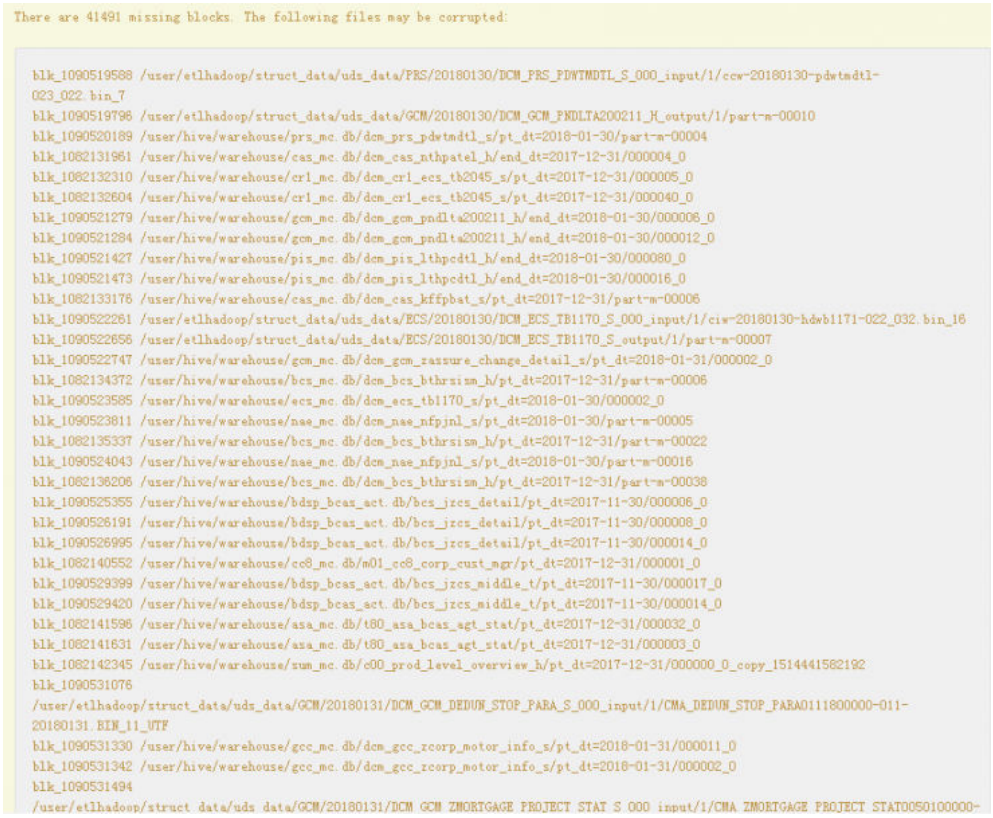
问题背景与现象

1. 用ntpdate修改了集群时间，修改时未停止集群，修改后HDFS进入安全模式，无法启动。
2. 退出安全模式后启动，hfck检查丢了大概1T数据。

原因分析

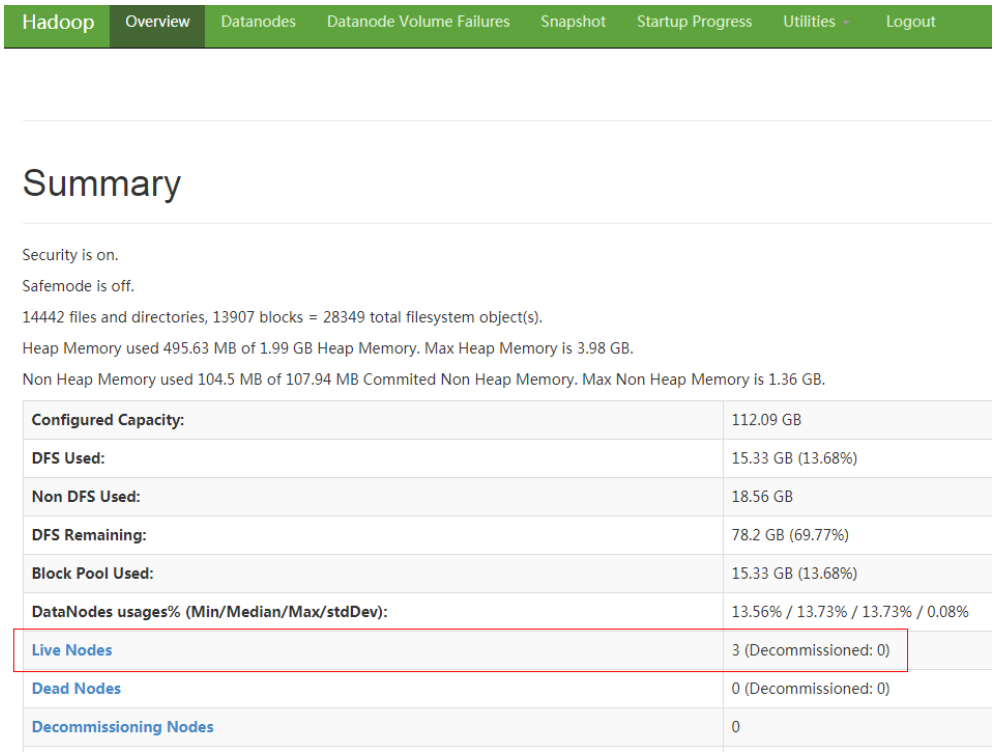
1. 查看NameNode原生页面发现有大量的块丢失。

图 16-23 块丢失



2. 查看原生页面 Datanode Information 发现显示的DataNode节点数和实际的相差10个节点。

图 16-24 查看 DataNode 节点数



3. 查看DateNode运行日志“/var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-主机名.log”，发现如下错误信息。

   重要错误信息 Clock skew too great

图 16-25 DateNode 运行日志错误

```
    at org.apache.hadoop.ipc.Client.call(Client.java:1486)
    at org.apache.hadoop.ipc.Client.call(Client.java:1447)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:229)
    at com.sun.proxy.$Proxy14.versionRequest(Unknown Source)
    at org.apache.hadoop.hdfs.protocolPB.DatanodeProtocolClientSideTranslatorPB.versionRequest(DatanodeProtocolClientSideTranslatorPB.java:273)
    at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.retrieveNamespaceInfo(BPServiceActor.java:187)
    at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.connectToNNAndHandshake(BPServiceActor.java:237)
    at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.run(BPServiceActor.java:689)
    at java.lang.Thread.run(Thread.java:745)
Caused by: GSSException: No valid credentials provided (Mechanism level: Clock skew too great (37))
    at sun.security.jgss.krb5.Krb5Context.initSecContext(Krb5Context.java:770)
    at sun.security.jgss.GSSContextImpl.initSecContext(GSSContextImpl.java:248)
    at sun.security.jgss.GSSContextImpl.initSecContext(GSSContextImpl.java:179)
    at com.sun.security.sasl.gsskerb.GssKrb5Client.evaluateChallenge(GssKrb5Client.java:192)
    ... 20 more
Caused by: KrbException: Clock skew too great (37)
    at sun.security.krb5.KrbKdcRep.check(KrbKdcRep.java:88)
    at sun.security.krb5.KrbTgsRep.<init>(KrbTgsRep.java:87)
    at sun.security.krb5.KrbTgsReq.getReply(KrbTgsReq.java:259)
    at sun.security.krb5.KrbTgsReq.sendAndGetCreds(KrbTgsReq.java:270)
    at sun.security.krb5.internal.CredentialsUtil.serviceCreds(CredentialsUtil.java:302)
    at sun.security.krb5.internal.CredentialsUtil.acquireServiceCreds(CredentialsUtil.java:120)
    at sun.security.krb5.Credentials.acquireServiceCreds(Credentials.java:458)
    at sun.security.jgss.krb5.Krb5Context.initSecContext(Krb5Context.java:693)
```

解决办法

步骤1 修改在原生页面查看不到的10个数据节点的时间。

步骤2 在Manager页面重启对应的DataNode实例。

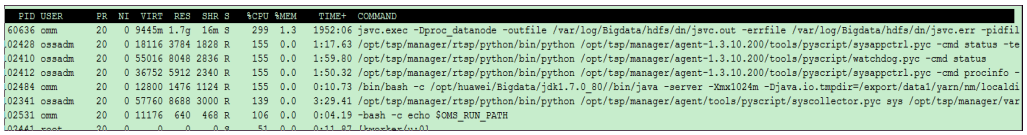
----结束

16.9.10 DataNode 概率性出现 CPU 占用接近 100%，导致节点丢失（ssh 连得很慢或者连不上）

问题背景与现象

DataNode概率性出现CPU占用接近100%，导致节点丢失。

图 16-26 DataNode 出现 CPU 占用接近 100%



原因分析

1. DataNode有许多写失败的日志。

   图 16-27 DataNode 写失败的日志

   ```
   2015-08-31 11:29:34,184 | ERROR | DataXceiver for client DFSClient_NONMAPREDUCE_1675952887_23 at /192.168.8.40:44514 [Receiving block
   BP-125271511-192.168.8.29-1440656260530:blk_1074766997_1034914] | TSP21:25009:DataXceiver error processing WRITE_BLOCK operation  src:
   /192.168.8.40:44514 dst: /192.168.8.64:25009 | DataXceiver.java:258
   java.io.IOException: Premature EOF from inputStream
           at org.apache.hadoop.io.IOUtils.readFully(IOUtils.java:194)
           at org.apache.hadoop.hdfs.protocol.datatransfer.PacketReceiver.doReadFully(PacketReceiver.java:213)
           at org.apache.hadoop.hdfs.protocol.datatransfer.PacketReceiver.doRead(PacketReceiver.java:134)
           at org.apache.hadoop.hdfs.protocol.datatransfer.PacketReceiver.receiveNextPacket(PacketReceiver.java:109)
           at org.apache.hadoop.hdfs.server.datanode.BlockReceiver.receivePacket(BlockReceiver.java:446)
           at org.apache.hadoop.hdfs.server.datanode.BlockReceiver.receiveBlock(BlockReceiver.java:707)
           at org.apache.hadoop.hdfs.server.datanode.DataXceiver.writeBlock(DataXceiver.java:748)
           at org.apache.hadoop.hdfs.protocol.datatransfer.Receiver.opWriteBlock(Receiver.java:124)
           at org.apache.hadoop.hdfs.protocol.datatransfer.Receiver.processOp(Receiver.java:71)
           at org.apache.hadoop.hdfs.server.datanode.DataXceiver.run(DataXceiver.java:240)
           at java.lang.Thread.run(Thread.java:745)
   2015-08-31 11:29:35,147 | INFO  | DataXceiver for client DFSClient_NONMAPREDUCE_-402997805_1 at /192.168.8.30:59449 [Sending block BP-
   125271511-192.168.8.29-1440656260530:blk_1074181856_446655] | src: /192.168.8.64:25009, dest: /192.168.8.30:59449, bytes: 16826, op:
   HDFS_READ, cliID: DFSClient_NONMAPREDUCE_-402997805_1, offset: 0, srvID: 9d1d30a5-046d-438b-83c9-2c6c54c6bd12, blockid: BP-125271511-
   192.168.8.29-1440656260530:blk_1074181856_446655, duration: 78832 | BlockSender.java:738
   2015-08-31 11:29:35,269 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg
   GC): pause of approximately 7480ms
   No GCs detected | JvmPauseMonitor.java:172
   2015-08-31 11:29:36,985 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg
   GC): pause of approximately 1215ms
   No GCs detected | JvmPauseMonitor.java:172
   2015-08-31 11:29:43,067 | INFO  | DataXceiver for client DFSClient_NONMAPREDUCE_1675952887_23 at /192.168.8.33:35530 [Receiving block
   BP-125271511-192.168.8.29-1440656260530:blk_1074767006_1034923] | Exception for BP-125271511-192.168.8.29-
   1440656260530:blk_1074767006_1034923 | BlockReceiver.java:742
   java.io.IOException: Premature EOF from inputStream
   ```

2. 短时间内写入大量文件导致这种情况，因此DataNode内存不足。

   图 16-28 写入大量文件导致 DataNode 内存不足

   ```
   Line 153101: 2015-08-31 11:24:29,313 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 1199ms
   Line 153132: 2015-08-31 11:24:42,689 | WARN  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 11273ms
   Line 153135: 2015-08-31 11:24:45,810 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 1005ms
   Line 153138: 2015-08-31 11:24:49,801 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 1067ms
   Line 153155: 2015-08-31 11:25:10,167 | WARN  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 12323ms
   ```

解决办法

步骤1 检查DataNode内存配置，以及机器剩余内存是否充足。

步骤2 增加DataNode内存，并重启DataNode。

----结束

16.9.11 单 NameNode 长期故障，如何使用客户端手动 checkpoint

问题背景与现象

在备NameNode长期异常的情况下，会积攒大量的editlog，此时如果重启HDFS或者主NameNode，主NameNode会读取大量的未合并的editlog，导致耗时启动较长，甚至启动失败。

原因分析

备NameNode会周期性做合并editlog，生成fsimage文件的过程叫做checkpoint。备NameNode在新生成fsimage后，会将fsimage传递到主NameNode。

说明

由于“备NameNode会周期性做合并editlog”，因此当备NameNode异常时，无法合并editlog，因此主NameNode在下次启动的时候，需要加载较多editlog，需要大量内存，并且耗时较长。

合并元数据的周期由以下参数确定，即如果NameNode运行30分钟或者HDFS操作100万次，均会执行checkpoint。

- dfs.namenode.checkpoint.period：checkpoint周期，默认1800s。
- dfs.namenode.checkpoint.txns：执行指定操作次数后执行checkpoint，默认1000000。

解决办法

在重启前，主动执行异常checkpoint合并主NameNode的元数据。

步骤1 停止业务。

步骤2 获取主NameNode的主机名。

步骤3 在客户端执行如下命令：

source /opt/client/bigdata_env

kinit *组件用户*

说明：/opt/client 需要换为实际客户端的安装路径。

步骤4 执行如下命令，让主NameNode进入安全模式，其中linux22换为主NameNode的主机名。

hdfs dfsadmin -fs linux22:25000 -safemode enter

```
linux16:/opt/fi_client #  hdfs dfsadmin -fs linux22:25000   -safemode enter
17/04/26 18:38:30 WARN fs.FileSystem: "linux22:25000" is a deprecated filesystem name. Use "hdfs://linux22:25000/" instead.
17/04/26 18:38:32 INFO hdfs.PeerCache: SocketCache disabled.
Safe mode is ON
```

步骤5 执行如下命令，在主NameNode，合并editlog。

hdfs dfsadmin -fs linux22:25000 -saveNamespace

```
linux16:/opt/fi_client #  hdfs dfsadmin -fs linux22:25000   -saveNamespace
17/04/26 18:38:54 WARN fs.FileSystem: "linux22:25000" is a deprecated filesystem name. Use "hdfs://linux22:25000/" instead.
17/04/26 18:38:56 INFO hdfs.PeerCache: SocketCache disabled.
Save namespace successful
```

步骤6 执行如下命令，让主NameNode离开安全模式。

hdfs dfsadmin -fs linux22:25000 -safemode leave

```
linux16:/opt/fi_client #  hdfs dfsadmin -fs linux22:25000  -safemode leave
17/04/26 18:39:07 WARN fs.FileSystem: "linux22:25000" is a deprecated filesystem name. Use "hdfs://linux22:25000/" instead.
17/04/26 18:39:09 INFO hdfs.PeerCache: SocketCache disabled.
Safe mode is OFF
```

步骤7 检查是否真的合并完成。

cd /srv/BigData/namenode/current

检查先产生的fsimage是否是当前时间的，若是则表示已经合并完成

```
-rw-------  1 omm wheel 1048576 Apr 26 18:43 edits_inprogress_0000000000002083018
-rw-------  1 omm wheel  736657 Apr 26 15:46 fsimage_0000000000002071390
-rw-------  1 omm wheel      62 Apr 26 15:46 fsimage_0000000000002071390.md5
-rw-------  1 omm wheel  736657 Apr 26 16:46 fsimage_0000000000002075405
-rw-------  1 omm wheel      62 Apr 26 16:46 fsimage_0000000000002075405.md5
-rw-------  1 omm wheel  736410 Apr 26 17:46 fsimage_0000000000002079398
-rw-------  1 omm wheel      62 Apr 26 17:46 fsimage_0000000000002079398.md5
-rw-------  1 omm wheel       8 Apr 26 18:42 seen_txid
linux-20:/srv/BigData/namenode/current #
linux-20:/srv/BigData/namenode/current #
```

----结束

16.9.12 文件读写常见故障

问题背景与现象

当用户在HDFS上执行写操作时，出现“Failed to place enough replicas:expected…”信息。

原因分析

- DataNode的数据接受器不可用。

  此时DataNode会有如下日志：

  ```
  2016-03-17 18:51:44,721 | WARN |
  org.apache.hadoop.hdfs.server.datanode.DataXceiverServer@5386659f |
  hadoopc1h2:25009:DataXceiverServer: | DataXceiverServer.java:158
  java.io.IOException: Xceiver count 4097 exceeds the limit of concurrent xcievers: 4096
  at org.apache.hadoop.hdfs.server.datanode.DataXceiverServer.run(DataXceiverServer.java:140)
  at java.lang.Thread.run(Thread.java:745)
  ```

- DataNode的磁盘空间不足。
- DataNode的心跳有延迟。

解决办法

- 如果DataNode的数据接收器不可用，通过在Manager页面，增加HDFS参数“dfs.datanode.max.transfer.threads”的值解决。
- 如果没有足够的硬盘空间或者CPU，试着增加新的数据节点或确保资源是可用的（磁盘空间或CPU）。
- 如果网络问题，确保网络是可用的。

16.9.13 文件最大打开句柄数设置太小导致读写文件异常

问题背景与现象

文件最大打开句柄数设置太小，导致文件句柄不足。写文件到HDFS很慢，或者写文件失败。

原因分析

1. DataNode日志“/var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-XXX.log”，存在异常提示java.io.IOException: Too many open files。

```
2016-05-19 17:18:59,126 | WARN  |
org.apache.hadoop.hdfs.server.datanode.DataXceiverServer@142ff9fa |
YSDN12:25009:DataXceiverServer:  |
org.apache.hadoop.hdfs.server.datanode.DataXceiverServer.run(DataXceiverServer.java:160)
java.io.IOException: Too many open files
at sun.nio.ch.ServerSocketChannelImpl.accept0(Native Method)
at sun.nio.ch.ServerSocketChannelImpl.accept(ServerSocketChannelImpl.java:241)
at sun.nio.ch.ServerSocketAdaptor.accept(ServerSocketAdaptor.java:100)
at org.apache.hadoop.hdfs.net.TcpPeerServer.accept(TcpPeerServer.java:134)
at org.apache.hadoop.hdfs.server.datanode.DataXceiverServer.run(DataXceiverServer.java:137)
at java.lang.Thread.run(Thread.java:745)
```

2. 如果某个DataNode日志中打印“Too many open files”，说明该节点文件句柄不足，导致打开文件句柄失败，然后就会重试往其他DataNode节点写数据，最终表现为写文件很慢或者写文件失败。

解决办法

步骤1 执行**ulimit -a**命令查看有问题节点文件句柄数最多设置是多少，如果很小，建议修改成640000。

图 16-29 查看文件句柄数

```
[omm@189-39-150-167 ~]$ ulimit -a
core file size          (blocks, -c) 0
data seg size           (kbytes, -d) unlimited
scheduling priority             (-e) 0
file size               (blocks, -f) unlimited
pending signals                 (-i) 256551
max locked memory       (kbytes, -l) 64
max memory size         (kbytes, -m) unlimited
open files                      (-n) 640000
pipe size            (512 bytes, -p) 8
POSIX message queues     (bytes, -q) 819200
real-time priority              (-r) 0
stack size              (kbytes, -s) 10240
cpu time               (seconds, -t) unlimited
max user processes              (-u) 60000
virtual memory          (kbytes, -v) unlimited
file locks                      (-x) unlimited
```

步骤2 执行**vi /etc/security/limits.d/90-nofile.conf**命令编辑这个文件，修改文件句柄数设置。如果没有这个文件，可以新建一个文件，并按照下图内容修改。

图 16-30 修改文件句柄数

```
*          hard    nofile    640000
*          soft    nofile    640000
~
```

步骤3 重新打开一个终端窗口，用**ulimit -a**命令查看是否修改成功，如果没有，请重新按照上述步骤重新修改。

步骤4 从Manager页面重启DataNode实例。

----结束

16.9.14 客户端写文件 close 失败

问题背景与现象

客户端写文件close失败，客户端提示数据块没有足够副本数。

客户端日志：

```
2015-05-27 19:00:52.811 [pool-2-thread-3] ERROR: /tsp/nedata/collect/UGW/ugwufdr/
20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp143272380633
8 close hdfs sequence file fail (SequenceFileInfoChannel.java:444)
java.io.IOException: Unable to close file because the last block does not have enough number of replicas.
at org.apache.hadoop.hdfs.DFSOutputStream.completeFile(DFSOutputStream.java:2160)
at org.apache.hadoop.hdfs.DFSOutputStream.close(DFSOutputStream.java:2128)
at org.apache.hadoop.fs.FSDataOutputStream$PositionCache.close(FSDataOutputStream.java:70)
at org.apache.hadoop.fs.FSDataOutputStream.close(FSDataOutputStream.java:103)
at com.xxx.pai.collect2.stream.SequenceFileInfoChannel.close(SequenceFileInfoChannel.java:433)
at com.xxx.pai.collect2.stream.SequenceFileWriterToolChannel
$FileCloseTask.call(SequenceFileWriterToolChannel.java:804)
at com.xxx.pai.collect2.stream.SequenceFileWriterToolChannel
$FileCloseTask.call(SequenceFileWriterToolChannel.java:792)
at java.util.concurrent.FutureTask.run(FutureTask.java:262)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
at java.lang.Thread.run(Thread.java:745)
```

原因分析

1. HDFS客户端开始写Block。

   例如：HDFS客户端是在2015-05-27 18:50:24,232开始写/20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp1432723806338的。其中分配的块是blk_1099105501_25370893。

   ```
   2015-05-27 18:50:24,232 | INFO  | IPC Server handler 30 on 25000 | BLOCK* allocateBlock: /
   20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp1432723
   806338. BP-1803470917-192.168.57.33-1428597734132
   blk_1099105501_25370893{blockUCState=UNDER_CONSTRUCTION, primaryNodeIndex=-1,
   replicas=[ReplicaUnderConstruction[[DISK]DS-b2d7b7d0-f410-4958-8eba-6deecbca2f87:NORMAL|
   RBW], ReplicaUnderConstruction[[DISK]DS-76bd80e7-ad58-49c6-bf2c-03f91caf750f:NORMAL|RBW]]}
   |
   org.apache.hadoop.hdfs.server.namenode.FSNamesystem.saveAllocatedBlock(FSNamesystem.java:3166
   )
   ```

2. 写完之后HDFS客户端调用了fsync。

   ```
   2015-05-27 19:00:22,717 | INFO  | IPC Server handler 22 on 25000 | BLOCK* fsync:
   20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp1432723
   806338 for DFSClient_NONMAPREDUCE_-120525246_15 |
   org.apache.hadoop.hdfs.server.namenode.FSNamesystem.fsync(FSNamesystem.java:3805)
   ```

3. HDFS客户端调用close关闭文件，NameNode收到客户端的close请求之后就会检查最后一个块的完成状态，只有当有足够的DataNode上报了块完成才可用关闭文件，检查块完成的状态是通过checkFileProgress函数检查的，打印如下：

   ```
   2015-05-27 19:00:27,603 | INFO  | IPC Server handler 44 on 25000 | BLOCK* checkFileProgress:
   blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
   replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
   RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
   ```

```
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:28,005 | INFO  | IPC Server handler 45 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:28,806 | INFO  | IPC Server handler 63 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:30,408 | INFO  | IPC Server handler 43 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:33,610 | INFO  | IPC Server handler 37 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:40,011 | INFO  | IPC Server handler 37 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
```

4. NameNode打印了多次checkFileProgress是由于HDFS客户端多次尝试close文件，但是由于当前状态不满足要求，导致close失败， HDFS客户端retry的次数是由参数dfs.client.block.write.locateFollowingBlock.retries决定的，该参数默认是5，所以在NameNode的日志中看到了6次checkFileProgress打印。
5. 但是再过0.5s之后，DataNode就上报块已经成功写入。

```
2015-05-27 19:00:40,608 | INFO  | IPC Server handler 60 on 25000 | BLOCK* addStoredBlock:
blockMap updated: 192.168.10.21:25009 is added to
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
size 11837530 |
org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.logAddStoredBlock(BlockManager.java
:2393)
2015-05-27 19:00:48,297 | INFO  | IPC Server handler 37 on 25000 | BLOCK* addStoredBlock:
blockMap updated: 192.168.10.10:25009 is added to blk_1099105501_25370893 size 11837530 |
org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.logAddStoredBlock(BlockManager.java
:2393)
```

6. DataNode上报块写成功通知延迟的原因可能有：网络瓶颈导致、CPU瓶颈导致。
7. 如果此时再次调用close或者close的retry的次数增多，那么close都将返回成功。建议适当增大参数dfs.client.block.write.locateFollowingBlock.retries的值，默认值为5次，尝试的时间间隔为400ms、800ms、1600ms、3200ms、6400ms，12800ms，那么close函数最多需要25.2秒才能返回。

解决办法

步骤1 规避办法：

可以通过调整客户端参数dfs.client.block.write.locateFollowingBlock.retries的值来增加retry的次数，可以将值设置为6，那么中间睡眠等待的时间为400ms、800ms、

1600ms、3200ms、6400ms、12800ms，也就是说close函数最多要50.8秒才能返回。

----结束

备注说明

一般出现上述现象，说明集群负载很大，通过调整参数只是临时规避这个问题，建议还是降低集群负载。例如：避免把所有CPU都分配MR跑任务。

16.9.15 文件错误导致上传文件到 HDFS 失败

问题背景与现象

用**hadoop dfs -put**把本地文件拷贝到HDFS上，有报错。

上传部分文件后，报错失败，从NameNode原生页面看，临时文件大小不再变化。

原因分析

1. 查看NameNode日志“/var/log/Bigdata/hdfs/nn/hadoop-omm-namenode-主机名.log”，发现该文件一直在被尝试写，直到最终失败。

```
2015-07-13 10:05:07,847 | WARN  | org.apache.hadoop.hdfs.server.namenode.LeaseManager
$Monitor@36fea922 | DIR* NameSystem.internalReleaseLease: Failed to release lease for file /hive/
order/OS_ORDER._8.txt._COPYING_. Committed blocks are waiting to be minimally replicated. Try
again later. | FSNamesystem.java:3936
2015-07-13 10:05:07,847 | ERROR | org.apache.hadoop.hdfs.server.namenode.LeaseManager
$Monitor@36fea922 | Cannot release the path /hive/order/OS_ORDER._8.txt._COPYING_ in the lease
[Lease.  Holder: DFSClient_NONMAPREDUCE_-1872896146_1, pendingcreates: 1] |
LeaseManager.java:459
org.apache.hadoop.hdfs.protocol.AlreadyBeingCreatedException: DIR*
NameSystem.internalReleaseLease: Failed to release lease for file /hive/order/
OS_ORDER._8.txt._COPYING_. Committed blocks are waiting to be minimally replicated. Try again
later.
at FSNamesystem.internalReleaseLease(FSNamesystem.java:3937)
```

2. 根因分析：被上传的文件损坏，因此会上传失败。
3. 验证办法：cp或者scp被拷贝的文件，也会失败，确认文件本身已损坏。

解决办法

步骤1 文件本身损坏造成的此问题，采用正常文件进行上传。

----结束

16.9.16 界面配置 dfs.blocksize 后 put 数据，block 大小还是原来的大小

问题背景与现象

界面配置“dfs.blocksize”，将其设置为268435456，put数据，block大小还是原来的大小。

原因分析

客户端的“hdfs-site.xml”文件中的dfs.blocksize大小没有更改，以客户端配置为准。

解决办法

步骤1 确保“dfs.blocksize”为512的倍数。

步骤2 重新下载安装客户端或者更改客户端配置。

步骤3 dfs.blocksize是客户端配置，以客户端为准。若客户端不配置，以服务端为准。

----结束

16.9.17 读取文件失败，FileNotFoundException

问题背景与现象

有MapReduce任务所有map任务均成功，但reduce任务失败，查看日志发现报异常“FileNotFoundException...No lease on...File does not exist”。

```
Error: org.apache.hadoop.ipc.RemoteException(java.io.FileNotFoundException): No lease on /user/sparkhive/warehouse/daas/dsp/output/_temporary/1/_temporary/attempt_1479799053892_17075_r_000007_0/part-r-00007 (inode 6501287): File does not exist. Holder DFSClient_attempt_1479799053892_17075_r_000007_0_-1463597952_1 does not have any open files.
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkLease(FSNamesystem.java:3350)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.completeFileInternal(FSNamesystem.java:3442)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.completeFile(FSNamesystem.java:3409)
at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.complete(NameNodeRpcServer.java:789)
```

原因分析

FileNotFoundException...No lease on...File does not exist，该日志说明文件在操作的过程中被删除了。

1. 搜索HDFS的NameNode的审计日志（Active NameNode的/var/log/Bigdata/audit/hdfs/nn/hdfs-audit-namenode.log）搜索文件名，确认文件的创建时间。
2. 搜索文件创建到出现异常时间范围的NameNode的审计日志，搜索该文件是否被删除或者移动到其他目录。
3. 如果该文件没有被删除或者移动，可能是该文件的父目录，或者更上层目录被删除或者移动，需要继续搜索上层目录。如本样例中，是文件的父目录的父目录被删除。

   ```
   2017-05-31 02:04:08,286 | INFO  | IPC Server handler 30 on 25000 | allowed=true ugi=appUser@HADOOP.COM (auth:TOKEN) ip=/192.168.1.22  cmd=delete  src=/user/sparkhive/warehouse/daas/dsp/output/_temporary  dst=null  perm=null proto=rpc | FSNamesystem.java:8189
   ```

   说明

   - 如上日志说明：192.168.1.22 节点的appUser用户删除了/user/sparkhive/warehouse/daas/dsp/output/_temporary。
   - 可以使用**zgrep "文件名" *.zip**命令搜索zip包的内容。

解决办法

步骤1 需要排查业务，确认为何该文件或者文件的父目录被删除。

----结束

16.9.18 HDFS 写文件失败，item limit of / is exceeded

问题背景与现象

客户端或者上层组件日志报往HDFS的某目录写文件失败，报错为

The directory item limit of /tmp is exceeded: limit=5 items=5。

原因分析

1. 查看客户端或者NameNode运行日志“/var/log/Bigdata/hdfs/nn/hadoop-omm-namenode-XXX.log”存在异常提示The directory item limit of /tmp is exceeded:。该错误的含义为/tmp目录的文件数超过1048576的限制。
   ```
   2018-03-14 11:18:21,625 | WARN  | IPC Server handler 62 on 25000 | DIR* NameSystem.startFile: /tmp/test.txt The directory item limit of /tmp is exceeded: limit=1048576 items=1048577 | FSNamesystem.java:2334
   ```
2. 该限制是dfs.namenode.fs-limits.max-directory-items参数，定义单个目录下不含递归的最大目录数或者文件数，默认值1048576，取值范围1～6400000。

解决办法

步骤1 确认该目录不含递归拥有100万以上文件目录是否正常，如果正常，可以将HDFS的参数dfs.namenode.fs-limits.max-directory-items调大并且重启HDFS NameNode生效。

步骤2 如果该目录下拥有100万文件不正常，需要清理不需要的文件。

----结束

16.9.19 调整 shell 客户端日志级别

- **临时调整**，关闭该shell客户端窗口后，日志会还原为默认值。
  a. 执行**export HADOOP_ROOT_LOGGER**命令可以调整客户端日志级别。
  b. 执行**export HADOOP_ROOT_LOGGER=*日志级别*,console**，可以调整shell客户端的日志级别。
     export HADOOP_ROOT_LOGGER=DEBUG,console，调整为DEBUG。
     export HADOOP_ROOT_LOGGER=ERROR,console，调整为ERROR。
- **永久调整**
  a. 在HDFS客户端环境变量配置文件“/opt/client/HDFS/component_env”（其中“/opt/client”需要改为实际客户端路径）增加“export HADOOP_ROOT_LOGGER=*日志级别*,console”。
  b. 执行**source /opt/client/bigdata_env**。
  c. 重新执行客户端命令。

16.9.20 读文件失败 No common protection layer

问题背景与现象

shell客户端或者其他客户端操作HDFS失败，报“**No common protection layer between client and server**”。

在集群外的机器，执行任意hadoop命令，如**hadoop fs -ls /**均失败，最底层的报错为"**No common protection layer between client and server**"。

```
2017-05-13 19:14:19,060 | ERROR | [pool-1-thread-1] |  Server startup failure  |
org.apache.sqoop.core.SqoopServer.initializeServer(SqoopServer.java:69)
org.apache.sqoop.common.SqoopException: MAPRED_EXEC_0028:Failed to operate HDFS - Failed to get the
file /user/loader/etl_dirty_data_dir status
        at org.apache.sqoop.job.mr.HDFSClient.fileExist(HDFSClient.java:85)
...
        at java.lang.Thread.run(Thread.java:745)
Caused by: java.io.IOException: Failed on local exception: java.io.IOException: Couldn't setup connection for
loader/hadoop@HADOOP.COM to loader37/10.162.0.37:25000; Host Details : local host is:
"loader37/10.162.0.37"; destination host is: "loader37":25000;
        at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:776)
...
        ... 10 more
Caused by: java.io.IOException: Couldn't setup connection for loader/hadoop@HADOOP.COM to
loader37/10.162.0.37:25000
        at org.apache.hadoop.ipc.Client$Connection$1.run(Client.java:674
        ... 28 more
Caused by: javax.security.sasl.SaslException: No common protection layer between client and server
        at com.sun.security.sasl.gsskerb.GssKrb5Client.doFinalHandshake(GssKrb5Client.java:251)
...
        at org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:720)
```

原因分析

1. HDFS的客户端和服务端数据传输走的rpc协议，该协议有多种加密方式，由hadoop.rpc.protection参数控制。
2. 如果客户端和服务端的hadoop.rpc.protection参数的配置值不一样，即会报**No common protection layer between client and server**错误。

   说明

   hadoop.rpc.protection参数表示数据可通过以下任一方式在节点间进行传输。

   - privacy：指数据在鉴权及加密后再传输。这种方式会降低性能。
   - authentication：指数据在鉴权后直接传输，不加密。这种方式能保证性能但存在安全风险。
   - integrity：指数据直接传输，即不加密也不鉴权。 为保证数据安全，请谨慎使用这种方式。

解决办法

步骤1 重新下载客户端，如果是应用程序，更新应用程序中的配置文件。

----结束

16.9.21 HDFS 目录配额（quota）不足导致写文件失败

问题背景与现象

给某目录设置quota后，往目录中写文件失败，出现如下问题“**The DiskSpace quota of** /tmp/tquota2 is **exceeded**”

```
[omm@189-39-150-115 client]$ hdfs dfs -put switchuser.py  /tmp/tquota2
put: The DiskSpace quota of /tmp/tquota2 is exceeded: quota = 157286400 B = 150 MB but diskspace
consumed = 402653184 B = 384 MB
```

可能原因

目录配置的剩余的空间小于写文件实际需要的空间。

原因分析

1. HDFS支持设置某目录的配额，即限制某目录下的文件最多占用空间大小，例如如下命令是设置/tmp/tquota 目录最多写入150MB的文件（文件大小*副本数）。

   hadoop dfsadmin -setSpaceQuota 150M /tmp/tquota2

2. 使用如下命令可以查看目录设置的配额情况，SPACE_QUOTA是设置的空间配额，REM_SPACE_QUOTA是当前剩余的空间配额。

   hdfs dfs -count -q -h -v /tmp/tquota2

   图 16-31 查看目录设置的配额

   ```
   hdfs dfs -count -q -h -v /tmp/tquota2
   QUOTA REM_QUOTA SPACE_QUOTA REM_SPACE_QUOTA  DIR_COUNT   FILE_COUNT CONTENT_SIZE PATHNAME
   none  inf       150M        150M                1           0            0  /tmp/tquota2
   ```

3. 日志分析，如下日志说明写入文件需要消耗384M，但是当前的空间配额是150M，因此空间不足。写文件前，需要的剩余空间是：块大小*副本数，128M*3副本=384M。

   ```
   [omm@189-39-150-115 client]$
   [omm@189-39-150-115 client]$ hdfs dfs -put switchuser.py  /tmp/tquota2
   put: The DiskSpace quota of /tmp/tquota2 is exceeded: quota = 157286400 B = 150 MB but diskspace consumed = 402653184 B = 384 MB
   ```

解决办法

步骤1 增加配额大小，即重新设置目录的配额大小。

hadoop dfsadmin -setSpaceQuota 150G /*目录名*

步骤2 清空配额。

hdfs dfsadmin -clrSpaceQuota /*目录名*

----结束

16.9.22 执行 balance 失败，Source and target differ in block-size

问题背景与现象

执行distcp跨集群拷贝文件时，出现部分文件拷贝失败“ Source and target differ in block-size. Use -pb to preserve block-sizes during copy. ”

```
Caused by: java.io.IOException: Check-sum mismatch between hdfs://10.180.144.7:25000/kylin/kylin_default_instance_prod/parquet/f2e72874-f01c-45ff-b219-207f3a5b3fcb/c769cd2d-575a-4459-837b-a19dd7b20c27/339114721280/0.parquettar and hdfs://10.180.180.194:25000/kylin/kylin_default_instance_prod/parquet/f2e72874-f01c-45ff-b219-207f3a5b3fcb/.distcp.tmp.attempt_1523424430246_0004_m_000019_2. Source and target differ in block-size. Use -pb to preserve block-sizes during copy. Alternatively, skip checksum-checks altogether, using -skipCrc. (NOTE: By skipping checksums, one runs the risk of masking data-corruption during file-transfer.)        at org.apache.hadoop.tools.mapred.RetriableFileCopyCommand.compareCheckSums(RetriableFileCopyCommand.java:214)
```

可能原因

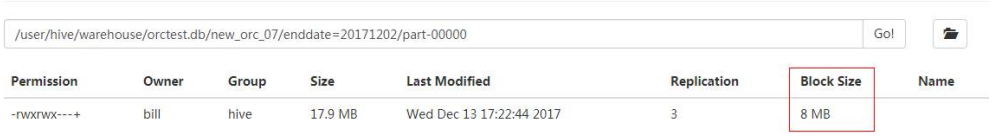
distcp默认拷贝文件时不记录原block大小导致在原文件block.size不是128M时校验失败，需要在distcp命令增加-pb参数。

原因分析

1. HDFS在写的时候有设置块大小，默认128M，某些组件或者业务程序写入的文件可能不是128M，如8M。

```
<name>dfs.blocksize</name>
<value>134217728</value>
```

图 16-32 某些组件或者业务程序写入的文件大小



2. distcp 从源集群读文件后写入新集群，默认是使用的MapReduce任务中的dfs.blocksize，默认128M。
3. 在distcp写完文件后，会基于块的物理大小做校验，因为该文件在新旧集群中block.size不一致，因此拆分大小不一致，导致校验失败。

   如以上文件，在旧集群是17.9/8MB = 3个block， 在新集群 17.9/128M = 1个block. 因此实际在磁盘的物理大小因分割而导致校验失败。

解决办法

distcp时，增加**-pb**参数。该参数作用为distcp时候保留block大小，确保新集群写入文件blocksize和老集群一致。

图 16-33 distcp 时保留 block 大小

```
[root@189-39-235-118 clientu10]#hadoop distcp -pb  hdfs://haclusterX/user hdfs://hacluster/tmp/test
```

16.9.23 查询或者删除文件失败，父目录可以看见此文件（不可见字符）

问题背景与现象

使用HDFS的shell客户端查询或者删除文件失败，父目录可以看见此文件。

图 16-34 父目录文件列表

```
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 01:44 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 16:45 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp2
[root@dggtsp355-or FusionInsight_client]# hadoop fs -ls /user/hive/warehouse/datalake_dwi_barpsit.db
Found 4 items
drwxrwxr-x   - datalab90020_639_w hive          0 2018-04-11 12:05 /user/hive/warehouse/datalake_dwi_barpsit.db/bak_v_tp_mp_aut_input
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-11 11:16 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 01:44 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 16:45 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp2
[root@dggtsp355-or FusionInsight_client]# hadoop fs -rm -r /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
rm: `/user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input': No such file or directory
[root@dggtsp355-or FusionInsight_client]# hdfs dfs -rm -r /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
rm: `/user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input': No such file or directory
[root@dggtsp355-or FusionInsight_client]#
[root@dggtsp355-or FusionInsight_client]#
[root@dggtsp355-or FusionInsight_client]# hdfs dfs -ls /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
ls: `/user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input': No such file or directory
[root@dggtsp355-or FusionInsight_client]#
[root@dggtsp355-or FusionInsight_client]#
```

原因分析

可能是该文件写入时有异常，写入了不可见字符。可以将该文件名重定向写入本地文本中，使用**vi**命令打开。

hdfs dfs -ls *父目录* **> /tmp/t.txt**

vi /tmp/t.txt

然后输入命令“**:set list**”将文件名的不可见字符显示出来。如这里显示出文件名中包含“**^M**”不可见字符。

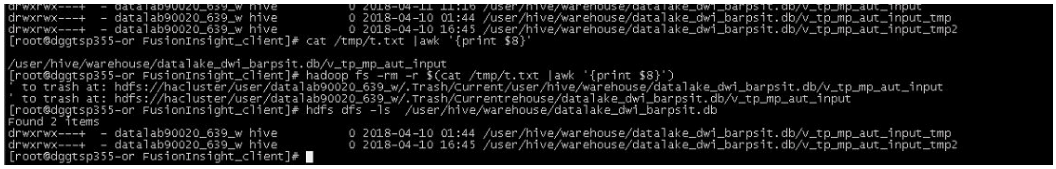
图 16-35 显示不可见字符

解决办法

步骤1 使用shell命令读到文本中记录的该文件名，确认如下命令输出的是该文件在HDFS中的全路径。

cat /tmp/t.txt |awk '{print $8}'

图 16-36 文件路径



步骤2 使用如下命令删除该文件。

hdfs dfs -rm $(cat /tmp/t.txt |awk '{print $8}')

步骤3 查看确认该文件已被删除。

hdfs dfs -ls *父目录*

----结束

16.9.24 非 HDFS 数据残留导致数据分布不均衡

问题背景与现象

数据出现不均衡，某磁盘过满而其他磁盘未写满。

HDFS DataNode数据存储目录配置为“/export/data1/dfs--/export/data12/dfs”，看到的现象是大量数据都是存储到了“/export/data1/dfs”，其他盘的数据比较均衡。

原因分析

磁盘为卸载重装，有一个目录在上次卸载时未卸载干净，即添加的磁盘，未格式化，残留历史垃圾数据。

解决办法

手动清理未卸载干净的数据。

16.9.25 客户端安装在数据节点导致数据分布不均衡

问题背景与现象

HDFS的DataNode数据分布不均匀，在某节点上磁盘使用率很高，甚至达到100%，其他节点空闲很多。

原因分析

客户端安装在该节点，根据HDFS数据副本机制，第一个副本会存放在本地机器，最终导致节点磁盘被占满，而其他节点空闲很多。

解决办法

步骤1 针对已有不平衡的数据，执行balance脚本均衡数据。

/opt/client/HDFS/hadoop/sbin/start-balancer.sh -threshold 10

其中 /opt/client是实际的客户端安装目录。

步骤2 针对新写入数据，将客户端安装在没有安装DataNode的节点。

----结束

16.9.26 节点内 DataNode 磁盘使用率不均衡处理指导

问题背景与现象

单个节点内DataNode的各磁盘使用率不均匀。

例如：

```
189-39-235-71:~ # df -h
Filesystem  Size  Used Avail Use% Mounted on
/dev/xvda  360G  92G  250G  28% /
/dev/xvdb  700G  900G  200G  78% /srv/BigData/hadoop/data1
/dev/xvdc  700G  900G  200G  78% /srv/BigData/hadoop/data2
/dev/xvdd  700G  900G  200G  78% /srv/BigData/hadoop/data3
/dev/xvde  700G  900G  200G  78% /srv/BigData/hadoop/data4
/dev/xvdf  10G  900G  890G  2% /srv/BigData/hadoop/data5
189-39-235-71:~ #
```

可能原因

部分磁盘故障，更换为新盘，因此新盘使用率低。

增加了磁盘个数，如原先4个数据盘，现扩容为5个数据盘。

原因分析

DataNode节点内写block磁盘时，有两种策略“轮询”和“优先写剩余磁盘空间多的磁盘”，默认是“轮询”。

参数说明：dfs.datanode.fsdataset.volume.choosing.policy

可选值：

- 轮询：
  org.apache.hadoop.hdfs.server.datanode.fsdataset.RoundRobinVolumeChoosingPolicy
- 优先写剩余空间多的磁盘：
  org.apache.hadoop.hdfs.server.datanode.fsdataset.AvailableSpaceVolumeChoosingPolicy

解决办法

将DataNode选择磁盘策略的参数 dfs.datanode.fsdataset.volume.choosing.policy 的值改为：
org.apache.hadoop.hdfs.server.datanode.fsdataset.AvailableSpaceVolumeChoosingPolicy，保存并重启受影响的服务或实例。

让DataNode根据磁盘剩余空间大小，优先选择磁盘剩余空间多的节点存储数据副本。

说明

- 针对新写入到本DataNode的数据会优先写磁盘剩余空间多的磁盘。
- 部分磁盘使用率较高，依赖业务逐渐删除在HDFS中的数据（老化数据）来逐渐降低。

16.9.27 执行 balance 常见问题定位方法

问题 1：报没权限（Access denied）执行 balance

问题详细：执行start-balancer.sh，“hadoop-root-balancer-主机名.out”日志显示“Access denied for user test1. Superuser privilege is required”

```
cat /opt/client/HDFS/hadoop/logs/hadoop-root-balancer-host2.out
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  Bytes Being Moved
INFO: Watching file:/opt/client/HDFS/hadoop/etc/hadoop/log4j.properties for changes with interval : 60000
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.AccessControlException): Access denied for user test1.
Superuser privilege is required
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkSuperuserPrivilege(FSPermissionChecker.java:122)
at
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkSuperuserPrivilege(FSNamesystem.java:5916)
```

问题根因：

执行balance需要使用管理员账户

解决方法

- 安全版本
  使用hdfs或者其他属于supergroup组的用户认证后，执行balance
- 普通版本
  执行HDFS的balance命令前，需要在客户端执行su - hdfs命令。

问题 2：执行 balance 失败，/system/balancer.id 文件异常

问题详细：

在HDFS客户端启动一个Balance进程，该进程被异常停止后，再次执行Balance操作，操作会失败。

```
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.protocol.RecoveryInProgressException): Failed to APPEND_FILE /system/balancer.id for DFSClient because lease recovery is in progress. Try again later.
```

问题根因：

通常，HDFS执行Balance操作结束后，会自动释放“/system/balancer.id”文件，可再次正常执行Balance。

但在上述场景中，由于第一次的Balance操作是被异常停止的，所以第二次进行Balance操作时，“/system/balancer.id”文件仍然存在，则会触发append /system/balancer.id操作，进而导致Balance操作失败。

解决方法

方法1：等待硬租期超过1小时后，原有客户端释放租约，再执行第二次Balance操作。

方法2：删除HDFS中的“/system/balancer.id”文件，再执行下次Balance操作。

16.9.28 HDFS 显示磁盘空间不足，其实还有 10%磁盘空间

问题背景与现象

1. 出现“HDFS磁盘空间使用率超过阈值”告警。
2. 查看HDFS页面，查看磁盘空间使用率非常高。

原因分析

HDFS中配置了dfs.datanode.du.reserved.percentage参数：每个磁盘的保留空间所占磁盘百分比。DataNode会保留这么多可用空间，以备其他组件如Yarn的NodeManager运行计算时，或者预留升级时使用。

因为预留了10%的磁盘，当磁盘使用率达到90%的时候，HDFS的DataNode即会认为没有可用磁盘空间。

解决办法

步骤1 扩容，在HDFS DataNode磁盘到80%，即需要及时扩容。

步骤2 如不能及时扩容，需要删除HDFS中的不需要数据，释放磁盘空间。

----结束

16.9.29 普通集群在 Core 节点安装 hdfs 客户端，使用时报错

用户问题

普通集群在Core节点新建用户安装使用客户端报错。

问题现象

普通集群在Core节点新建用户安装使用客户端报错如下：

```
2020-03-14 19:16:17,166 WARN shortcircuit.DomainSocketFactory: error creating DomainSocket
java.net.ConnectException: connect(2) error: Permission denied when trying to connect to '/var/run/MRS-
```

```
HDFS/dn_socket'
at org.apache.hadoop.net.unix.DomainSocket.connect0(Native Method)
at org.apache.hadoop.net.unix.DomainSocket.connect(DomainSocket.java:256)
at org.apache.hadoop.hdfs.shortcircuit.DomainSocketFactory.createSocket(DomainSocketFactory.java:168)
at org.apache.hadoop.hdfs.client.impl.BlockReaderFactory.nextDomainPeer(BlockReaderFactory.java:799)
...
```

原因分析

用户使用 **useradd** 命令来创建用户，此用户默认用户组不包含“ficommmon”用户组，导致在使用hdfs的get命令的时候出现上述报错。

处理步骤

使用命令**usermod -a -G ficommon username** 为用户添加用户组“ficommon”。

16.9.30 集群外节点安装客户端使用 hdfs 上传文件失败

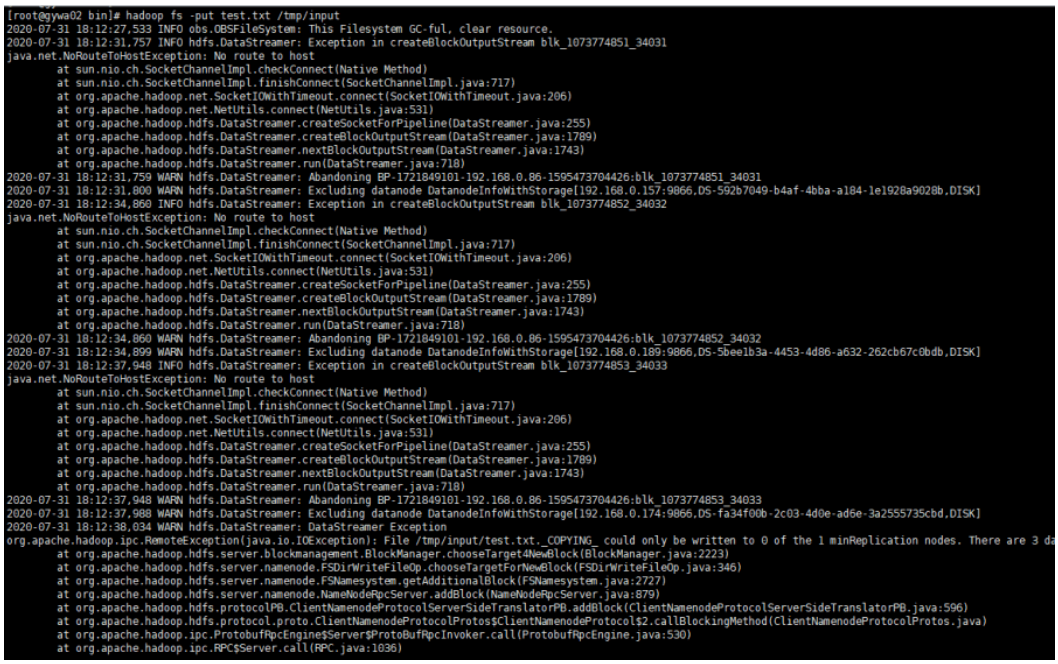
用户问题

集群外节点安装客户端使用hdfs上传文件失败

问题现象

在集群节点上安装客户端，在该客户端使用hdfs命令上传一个文件，报如下错误：

图 16-37 上传文件报错



原因分析

从错误截图可以看到报错是no route to host，且报错信息里面有192.168的ip，也即客户端节点到集群的DN节点的内网路由不通，导致上传文件失败。

处理步骤

在客户端节点的客户端目录下，找到HDFS的客户端配置目录hdfs-site.xml文件，在配置文件中增加配置项dfs.client.use.datanode.hostname，并将该配置设置为true。

16.9.31 HDFS 写并发较大时，报副本不足的问题

问题背景与现象

用户运行作业时写文件到HDFS，偶现写文件失败的情况。

操作日志如下：

```
105 | INFO  | IPC Server handler 23 on 25000 | IPC Server handler 23 on 25000, call
org.apache.hadoop.hdfs.protocol.ClientProtocol.addBlock from 192.168.1.96:47728 Call#1461167 Retry#0 |
Server.java:2278
java.io.IOException: File /hive/warehouse/000000_0.835bf64f-4103 could only be replicated to 0 nodes
instead of minReplication (=1).  There are 3 datanode(s) running and 3 node(s) are excluded in this
operation.
```

原因分析

- HDFS写文件的预约机制：无论文件是10M还是1G，开始写的每个块都会被预约128M。如果需要写入一个10M的文件，HDFS会预约一个块来写，当文件写完后，这个块只占实际大小10M，释放多余预约的118M空间。如果需要写入一个1G的文件，HDFS还是会预约一个块来写，这个块写完后再开启下一个块，文件写完后，实际占用1G磁盘，释放多余预约的空间。
- 该异常通常是因为业务写文件的并发量太高，预约写Block的磁盘空间不足，导致写文件失败。

解决办法

步骤1 登录HDFS的WebUI页面，进入DataNode的JMX页面。

1. 在HDFS原生界面，选择Datanodes页面。
2. 找到对应的DataNode节点，单击Http Address地址进入DataNode详情。
3. 将url的“datanode.html”改为“jmx”就能获取到DataNode的JMX信息。

步骤2 搜索“XceiverCount”指标，当该指标的值*Block块的大小超过DataNode磁盘的容量，就说明预约写Block的磁盘空间不足。

步骤3 发生该问题，通常有以下两种方法来解决：

方法一：降低业务的并发度。

方法二：减少业务写文件的数目，将多个文件合并成一个文件来写。

----结束

16.9.32 HDFS 客户端无法删除超长目录

问题背景与现象

执行**hadoop fs -rm -r -f obs://***<obs_path>*命令，删除OBS超长目录出现如下报错：

```
2022-02-28 17:12:45,605 INFO internal.RestStorageService: OkHttp cost 19 ms to apply http request
2022-02-28 17:12:45,606 WARN internal.RestStorageService: Request failed, Response code: 400; Request
```

```
ID: 0000017F3F9A8545401491602FC8CAD9; Request path: http://wordcount01-fcq.obs.xxx.****.example.com/
user%2Froot%2F.Trash%2FCurrent
%2Ftest1%2F12345678901234567890123456789012345678901234567890123456789012345678901234567
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456
2022-02-28 17:12:45,606 WARN services.AbstractClient: Storage|1|HTTP+XML|getObjectMetadata||||
2022-02-28 17:12:45|2022-02-28 17:12:45|||400|
2022-02-28 17:12:45,607 INFO log.AccessLogger: 2022-02-28 17:12:45 605|
com.obs.services.internal.RestStorageService|executeRequest|560|OkHttp cost 19 ms to apply http request
2022-02-28 17:12:45 606|com.obs.services.internal.RestStorageService|handleThrowable|221|Request failed,
Response code: 400; Request ID: 0000017F3F9A8545401491602FC8CAD9; Request path: http://wordcount01-
fcq.obs.xxx.****.example.com/user%2Froot%2F.Trash%2FCurrent
%2Ftest1%2F12345678901234567890123456789012345678901234567890123456789012345678901234567
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
890123456789012345678901234567890123456789012345678901234567890123456
2022-02-28 17:12:45 606|com.obs.services.AbstractClient|doActionWithResult|404|Storage|1|HTTP+XML|
getObjectMetadata||||2022-02-28 17:12:45|2022-02-28 17:12:45|||400|
```

原因分析

使用rm命令从HDFS中删除某些内容时，文件或目录不会立即被清除，它们将被移动到回收站Current目录（ /user/*${username}*/.Trash/current ）中。

解决办法

使用skipTrash命令可以跳过HDFS回收站，直接删除。使用前先设置HDFS客户端配置项“dfs.client.skipTrash.enabled=true”。

步骤1 以**root**用户登录集群任一Master节点。

步骤2 执行如下命令编辑HDFS用到的hdfs-site.xml文件。

vim *客户单安装目录*/**HDFS/hadoop/etc/hadoop/hdfs-site.xml**

步骤3 在hdfs-site.xml文件中增加如下内容。

```
<property>
<name>dfs.client.skipTrash.enabled</name>
<value>true</value>
</property>
```

步骤4 执行以下命令直接删除OBS超长目录。

hadoop fs -rm -r -f –skipTrash obs://*<obs_path>*

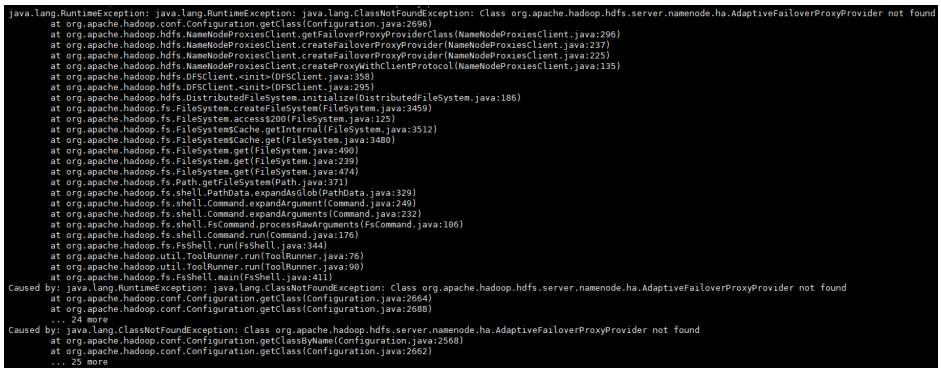
步骤5 登录集群其他Master节点，执行**步骤2**~**步骤4**，直到集群所有Master节点操作完成。

----结束

16.9.33 集群外节点访问 MRS HDFS 报错

问题背景与现象

集群外节点访问MRS HDFS的时候报错：Class org.apache.hadoop.hdfs.server.namenode.ha.AdaptiveFailoverProxyProvider not found。



原因分析

出现这个报错可能的场景有：

- 开源HDFS客户端访问MRS集群的HDFS时报错。
- 使用jar包连接MRS集群的HDFS（包括提交任务时连接HDFS）时报错。

解决办法

方法一：

步骤1 找到命令或者jar包使用的HDFS配置文件hdfs-site.xml。

步骤2 修改“dfs.client.failover.proxy.provider.hacluster”参数配置项如下。

```
<property>
<name>dfs.client.failover.proxy.provider.hacluster</name>
<value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider</value>
</property>
```

说明

也可以将上述配置项删除。

步骤3 保存文件后，重新访问MRS HDFS。

----结束

方法二：

步骤1 从Maven库下载MRS集群版本对应的hadoop-plugins版本。

步骤2 将下载的jar包添加到命令或者jar包的依赖中。

----结束

16.10 使用 Hive

16.10.1 Hive 各个日志里都存放了什么信息?

审计日志

首先，对于审计日志来说，记录了某个时间点某个用户从哪个IP发起对HiveServer或者MetaStore的请求以及记录执行的语句是什么。

如下的HiveServer审计日志，表示在2016-02-01 14:51:22 用户user_chen向HiveServer发起了show tables请求，客户端IP为192.168.1.18。

```
2016-02-01 14:51:22,335 | INFO  | HiveServer2-Handler-Pool: Thread-37815 | UserN
ame=user_chen  UserIP=192.168.1.18  Time=2016/02/01 14:51:22  Opertai
on=ExecuteStatement  stmt={show tables}  Resource=  Result= Detail=
| org.apache.hive.service.cli.thrift.ThriftCLIService.logAuditEvent(ThriftCLISer
vice.java:350)
```

如下面MetaStore审计日志，表示在2016-01-29 11:31:15 用户hive向MetaStore发起shutdown请求，客户端ip为192.168.1.18。

```
2016-01-29 11:31:15,451 | INFO  | pool-6-thread-70648 | ugi=hive/hadoop.hadoop.c
om@HADOOP.COM  ip=/192.168.1.18  cmd=Shutting down the object store...
 | org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.logAuditEvent(HiveM
etaStore.java:375)
```

通常情况下，审计日志对定位实际错误信息并无太大帮助。但在遇到诸如下述类问题时，需要查看审计日志：

1. 如客户端发起请求，但是迟迟未得到响应。由于不确定到底是任务是卡在客户端还是服务端，可以通过审计日志查看。如果审计日志根本没有相关信息，那么说明卡死在客户端；如审计日志有相关打印，那么就需要去运行日志里看到底程序卡在哪一步了。
2. 查看指定时间段的任务请求个数。可通过审计日志查看在指定时间段到底有多少个请求。

HiveServer 运行日志

简言之，HiveServer负责接收客户端请求（SQL语句），然后编译、执行（提交到YARN或运行local MR）、与MetaStore交互获取元数据信息等。HiveServer运行日志记录了一个SQL完整的执行过程。

通常情况下，当遇到SQL语句运行失败，首先要看的就是HiveServer运行日志。

MetaStore 运行日志

通常情况下，当遇到查看HiveServer运行日志时，如遇到MetaException或者连接MetaStore失败，需要去看MetaStore运行日志。

GC 日志查看

HiveServer和MetaStore均有GC日志，当遇到GC问题可以查看GC日志以快速定位是否是GC导致。如，当遇到HiveServer或MetaStore频繁重启就需要去看下对应的GC日志了。

16.10.2 Hive 启动失败问题的原因有哪些?

Hive启动失败最常见的原因是metastore实例无法连接上DBservice。可以查看metastore日志中具体的错误信息。目前总结连不上DBservice原因主要有：

可能原因 1

DBservice没有初始化好Hive的元数据库hivemeta。

处理步骤 1

步骤1 执行以下命令：

source /opt/Bigdata/MRS_*XXX*/install/dbservice/.dbservice_profile

gsql -h *DBservice浮动IP* **-p 20051 -d hivemeta -U hive -W HiveUser@**

步骤2 如果不能正确进入交互界面，说明数据库初始化失败。如果报如下错误说明在DBservice所在的节点的配置文件可能丢失了hivemeta的配置。

```
org.postgresql.util.PSQLException: FATAL: no pg_hba.conf entry for host "192.168.0.146", database "HIVEMETA"。
```

步骤3 编辑“/srv/BigData/dbdata_service/data/pg_hba.conf”，在文件最后面追加**host hivemeta hive 0.0.0.0/0 sha256**配置。

步骤4 执行**source /opt/Bigdata/MRS_*XXX*/install/dbservice/.dbservice_profile**命令配置环境变量。

步骤5 执行**gs_ctl -D $GAUSSDATA reload #**命令使修改后的配置生效。

----结束

可能原因 2

DBservice的浮动IP配置有误，导致metastore节点IP无法正确连接浮动IP，或者是在与该ip建立互信的时候失败导致metastore启动失败。

处理步骤 2

DBservice的浮动IP配置需要同网段内没有被使用过的ip，也就是在配置前ping不通的ip，请修改DBService浮动IP配置。

16.10.3 安全集群执行 set 命令的时候报 Cannot modify xxx at runtime.

问题现象

执行set命令时报以下错误：

```
0: jdbc:hive2://192.168.1.18:21066/> set mapred.job.queue.name=QueueA;
 Error: Error while processing statement: Cannot modify mapred.job.queue.name at list of params that are allowed to be modified at runtime (state=42000,code=1)
```

处理步骤

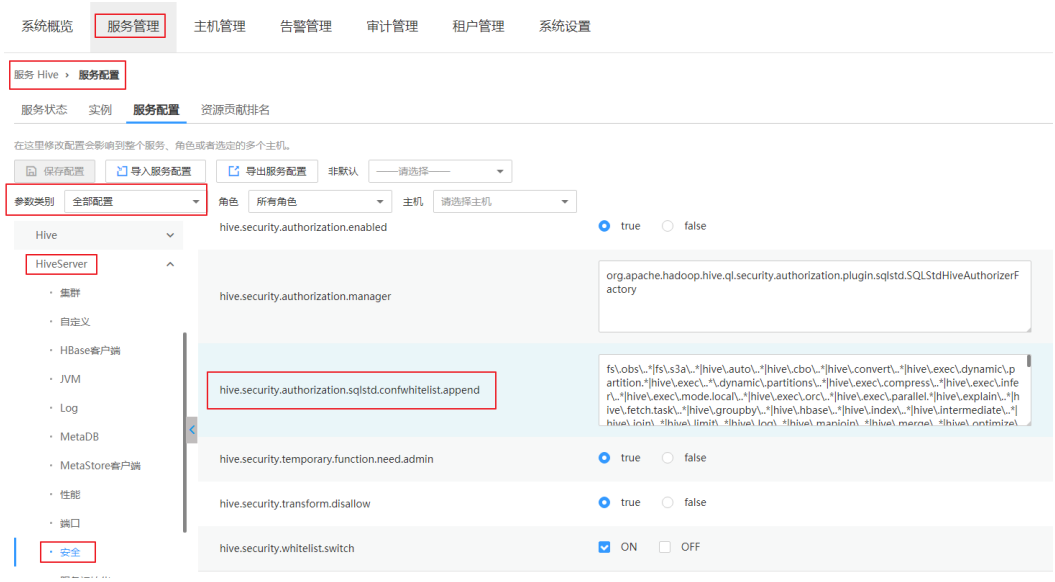
方案1：

步骤1 登录Manager界面，修改Hive参数。

- MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Hive > 服务配置 > 全部配置 > HiveServer > 安全”。
- FusionInsight Manager界面操作：登录FusionInsight Manager页面，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置 > 全部配置 > HiveServer > 安全”。

步骤2 将需要执行的命令参数添加到配置项hive.security.authorization.sqlstd.confwhitelist.append中。

步骤3 单击保存并重启HiveServer后即可。如下图所示：



----结束

方案2：

步骤1 登录Manager界面，修改Hive参数。

- MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Hive > 服务配置 > 全部配置 > HiveServer > 安全”。
- FusionInsight Manager界面操作：登录FusionInsight Manager页面，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置 > 全部配置 > HiveServer > 安全”。

步骤2 找到选项hive.security.whitelist.switch，选择OFF，单击保存并重启HiveServer即可。

----结束

16.10.4 怎样在 Hive 提交任务的时候指定队列?

问题现象

怎样在Hive提交任务的时候指定队列?

处理步骤

步骤1 在执行语句前通过如下参数设置任务队列，例如，提交任务至队列QueueA。

```
set mapred.job.queue.name=QueueA;
select count(*) from rc;
```

说明

队列的名称区分大小写，如写成queueA，Queuea均无效；且该队列为叶子队列，不支持提交任务到非叶子队列。

步骤2 提交任务后，可在Yarn页面看到，如下任务已经提交到队列QueueA。

User:	admin
Name:	select count(*) from rc(Stage-1)
Application Type:	MAPREDUCE
Application Tags:	
YarnApplicationState:	FINISHED
Queue:	QueueA
FinalStatus Reported by AM:	SUCCEEDED
Started:	Thu Mar 03 09:01:58 +0800 2016
Elapsed:	1mins, 0sec
Tracking URL:	History
Log Aggregation Status	Status
Diagnostics:	

----结束

16.10.5 客户端怎么设置 Map/Reduce 内存?

问题现象

客户端怎么设置Map/Reduce内存?

处理步骤

Hive在执行SQL语句前，可以通过set命令来设置Map/Reduce相关客户端参数。

以下为与Map/Reduce内存相关的参数：

```
set mapreduce.map.memory.mb=4096;// 每个Map Task需要的内存量
set mapreduce.map.java.opts=-Xmx3276M; // 每个Map Task 的JVM最大使用内存
set mapreduce.reduce.memory.mb=4096; // 每个Reduce Task需要的内存量
set mapreduce.reduce.java.opts=-Xmx3276M; // 每个Reduce Task 的JVM最大使用内存
set mapred.child.java.opts=-Xms1024M -Xmx3584M;//此参数为全局参数，既对Map和Reduce统一设置
```

说明

参数设置只对当前session有效。

16.10.6 如何在导入表时指定输出的文件压缩格式

问题现象

如何在导入表时指定输出的文件压缩格式?

处理步骤

当前Hive支持以下几种压缩格式：

```
org.apache.hadoop.io.compress.BZip2Codec
org.apache.hadoop.io.compress.Lz4Codec
org.apache.hadoop.io.compress.DeflateCodec
org.apache.hadoop.io.compress.SnappyCodec
org.apache.hadoop.io.compress.GzipCodec
```

- 如需要全局设置，即对所有表都进行压缩，可以在Manager页面对Hive的服务配置参数进行如下全局配置：
  - hive.exec.compress.output设置为true
  - mapreduce.output.fileoutputformat.compress.codec设置为org.apache.hadoop.io.compress.BZip2Codec

    说明

    hive.exec.compress.output参数必须设置为true，才能使下边的参数选项生效。
- 如需在session级设置，只需要在执行命令前增加如下设置即可：

```
set hive.exec.compress.output=true;
set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.SnappyCodec;
```

16.10.7 desc 描述表过长时，无法显示完整

问题现象

desc描述表过长时，如何让描述显示完整?

处理步骤

步骤1 启动Hive的beeline时，设置参数maxWidth=20000即可，例如：

```
[root@192-168-1-18 logs]# beeline --maxWidth=20000
scan complete in 3ms
Connecting to
......
Beeline version 1.1.0 by Apache Hive
```

步骤2 （可选）通过**beeline -help**命令查看关于客户端显示的设置。如下：

```
 -u <database url>              the JDBC URL to connect to
 -n <username>                  the username to connect as
 -p <password>                  the password to connect as
 -d <driver class>              the driver class to use
 -i <init file>                 script file for initialization
 -e <query>                     query that should be executed
 -f <exec file>                 script file that should be executed
 --hiveconf property=value      Use value for given property
 --color=[true/false]           control whether color is used for display
 --showHeader=[true/false]      show column names in query results
 --headerInterval=ROWS;         the interval between which heades are displayed
 --fastConnect=[true/false]     skip building table/column list for tab-completion
 --autoCommit=[true/false]      enable/disable automatic transaction commit
 --verbose=[true/false]         show verbose error messages and debug info
 --showWarnings=[true/false]    display connection warnings
 --showNestedErrs=[true/false]  display nested errors
 --numberFormat=[pattern]       format numbers using DecimalFormat pattern
 --force=[true/false]           continue running script even after errors
 --maxWidth=MAXWIDTH            the maximum width of the terminal
 --maxColumnWidth=MAXCOLWIDTH   the maximum width to use when displaying columns
 --silent=[true/false]          be more silent
 --autosave=[true/false]        automatically save preferences
 --outputformat=[table/vertical/csv2/tsv2/dsv/csv/tsv]  format mode for result display
```

```
                                Note that csv, and tsv are deprecated - use csv2, tsv2 instead
 --truncateTable=[true/false]     truncate table column when it exceeds length
 --delimiterForDSV=DELIMITER      specify the delimiter for delimiter-separated values output format
(default: |)
 --isolation=LEVEL                set the transaction isolation level
 --nullemptystring=[true/false]   set to true to get historic behavior of printing null as empty string
 --socketTimeOut=n                socket connection timeout interval, in second. The default value is 300.
```

----结束

16.10.8 增加分区列后再 insert 数据显示为 NULL

问题现象

1. 执行如下命令创建表
   ```
   create table test_table(
   col1 string,
   col2 string
   )
   PARTITIONED BY(p1 string)
   STORED AS orc tblproperties('orc.compress'='SNAPPY');
   ```
2. 修改表结构，添加分区并插入数据
   ```
   alter table test_table add partition(p1='a');
   insert into test_table partition(p1='a') select col1,col2 from temp_table;
   ```
3. 修改表结构，添加列并插入数据
   ```
   alter table test_table add columns(col3 string);
   insert into test_table partition(p1='a') select col1,col2,col3 from temp_table;
   ```
4. 查询test_table表数据，返回结果中列col3的值全为NULL
   ```
   select * from test_table where p1='a'
   ```
5. 新添加表分区，并插入数据
   ```
   alter table test_table add partition(p1='b');
   insert into test_table partition(p1='b') select col1,col2,col3 from temp_table;
   ```
6. 查询test_table表数据，返回结果中列col3有不为NULL的值
   ```
   select * from test_table where p1='b'
   ```

原因分析

在alter table时默认选项为RESTRICT，RESTRICT只会更改元数据，不会修改在此操作之前创建的partition的表结构，而只会修改在此之后创建的新的partition，所以查询时旧的partition中的值全为NULL。

处理步骤

add column时加入cascade关键字即可，例如：

```
alter table test_table add columns(col3 string) cascade;
```

16.10.9 创建新用户，执行查询时报无权限

问题现象

创建了新用户，但是执行查询的时候报无权限的错。

```
Error: Error while compiling statement: FAILED: HiveAccessControlException Permission denied: Principal
[name=hive, type=USER] does not have following privileges for operation QUERY [[SELECT] on Object
[type=TABLE_OR_VIEW, name=default.t1]] (state=42000,code=40000)
```

原因分析

创建的新用户没有Hive组件的操作权限。

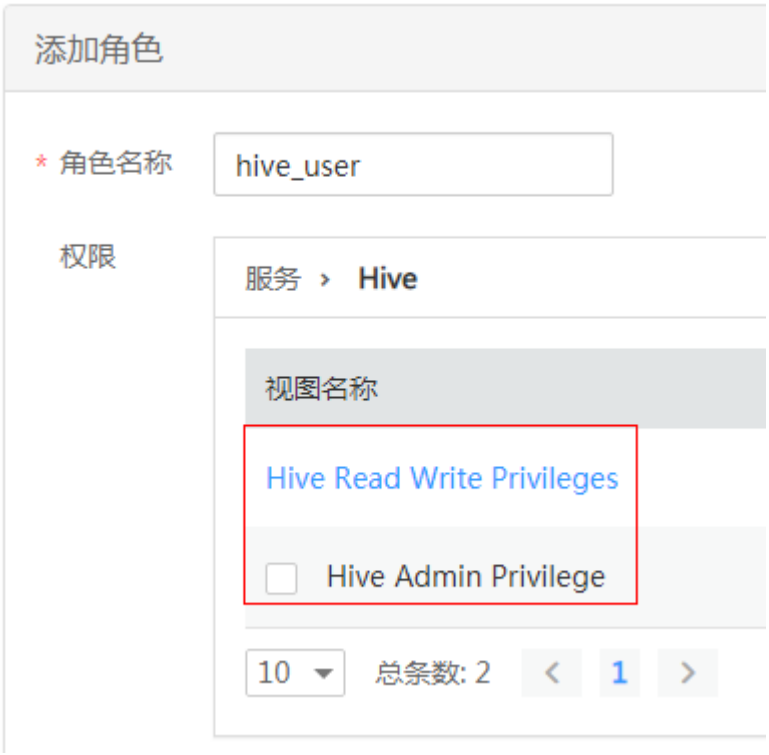
解决方案

MRS Manager界面操作：

步骤1 登录MRS Manager页面，选择“系统配置 > 角色管理 > 添加角色”。

步骤2 输入角色名称。

步骤3 在“权限”区域选择Hive，出现Hive管理员权限和Hive表的读写权限。



步骤4 选择“Hive Read Write Privileges”Hive表的读写权限，此时显示列Hive中的所有数据库。

步骤5 勾选角色需要的权限并单击“确定”完成角色创建。

步骤6 在MRS Manager页面，选择“系统配置 > 用户管理”。

步骤7 在已创建的新用户对应的“操作”列单击“修改”。

步骤8 单击“选择添加的用户组”，如需使用Hive服务，必须添加Hive组。

步骤9 单击“选择并绑定角色”，勾选**步骤5**中已创建的角色。

步骤10 单击“确定”完成用户权限的配置。

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，输入“角色名称”和“描述”。

步骤3 设置角色“配置资源权限”，选择“Hive读写权限”Hive表的读写权限，此时显示列Hive中的所有数据库。

步骤4 勾选角色需要的权限并单击“确定”完成角色创建。

步骤5 在FusionInsight Manager页面，选择“系统 > 权限 > 用户”。

步骤6 在已创建的新用户对应的“操作”列单击“修改”。

步骤7 单击“用户组”右侧的“添加”，如需使用Hive服务，必须添加Hive组。

步骤8 单击“角色”右侧的“添加”，勾选4中已创建的角色。

步骤9 单击“确定”完成用户权限的配置。

----结束

16.10.10 执行 SQL 提交任务到指定队列报错

问题现象

执行SQL提交任务到Yarn报如下错误：

```
Failed to submit application_1475400939788_0033 to YARN :
org.apache.hadoop.security.AccessControlException: User newtest cannot submit applications to queue
root.QueueA
```

原因分析

当前登录的用户无YARN队列提交权限。

解决方案

用户无YARN队列提交权限，需要赋予YARN相应队列的提交权限。在Manager页面的“系统 >权限 > 用户”中给用户绑定队列提交权限的角色。

16.10.11 执行 load data inpath 命令报错

问题现象

执行load data inpath报如下错误：

- 错误1：
  ```
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have
  following privileges on Object [type=DFS_URI, name=hdfs://hacluster/tmp/input/mapdata] for
  operation LOAD : [OBJECT OWNERSHIP]
  ```
- 错误2：
  ```
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have
  following privileges on Object [type=DFS_URI, name=hdfs://hacluster/tmp/input/mapdata] for
  operation LOAD : [INSERT, DELETE]
  ```
- 错误3：
  ```
  SemanticException [Error 10028]: Line 1:17 Path is not legal ''file:///tmp/input/mapdata'': Move from:
  file:/tmp/input/mapdata to: hdfs://hacluster/user/hive/warehouse/tmp1 is not valid. Please check that
  values for params "default.fs.name" and "hive.metastore.warehouse.dir" do not conflict.
  ```

原因分析

当前登录的用户不具备操作此目录的权限或者文件目录格式不正确。

解决方案

Hive对load data inpath命令有如下权限要求，请对照下述要求是否满足：

- 文件的owner需要为执行命令的用户。
- 当前用户需要对该文件有读、写权限。
- 当前用户需要对该文件的目录有执行权限。
- 由于load操作会将该文件移动到表对应的目录中，所以要求当前用户需要对表的对应目录有写权限。
- 要求文件的格式与表指定的存储格式相同。如创建表时指定stored as rcfile，但是文件格式为txt，则不符合要求。
- 文件必须是HDFS上的文件，不可以用file://的形式指定本地文件系统上的文件。
- 文件名不能以下横线（ _ ）或点（ . ）开头，以这些开头的文件会被忽略。

  如下所示，如果用户test_hive load数据，正确的权限如下：

  ```
  [root@192-168-1-18 duan]# hdfs dfs -ls /tmp/input2
  16/03/21 14:45:07 INFO hdfs.PeerCache: SocketCache disabled.
  Found 1 items
  -rw-r--r--   3 test_hive hive          6 2016-03-21 14:44 /tmp/input2/input.txt
  ```

16.10.12 执行 load data local inpath 命令报错

问题现象

执行load data local inpath报如下错误：

- 错误1：

  ```
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have following privileges on Object [type=LOCAL_URI, name=file:/tmp/input/mapdata] for operation LOAD : [SELECT, INSERT, DELETE]
  ```

- 错误2：

  ```
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have following privileges on Object [type=LOCAL_URI, name=file:/tmp/input/mapdata] for operation LOAD : [OBJECT OWNERSHIP]
  ```

- 错误3：

  ```
  SemanticException Line 1:23 Invalid path ''/tmp/input/mapdata'': No files matching path file:/tmp/input/mapdata
  ```

原因分析

当前登录的用户不具备操作此目录的权限或者在HiveServer所在节点上没有此目录。

解决方案

说明

通常不建议使用本地文件加载数据到hive表。 建议先将本地文件放入HDFS，然后从集群中加载数据。

Hive对load data local inpath命令有如下权限要求，请对照下述要求是否满足：

- 由于所有的命令都是发送到主HiveServer上去执行的，所以要求此文件在HiveServer节点上。
- HiveServer进程是以操作系统上的omm用户启动的，所以要求omm用户对此文件有读权限，对此文件的目录有读、执行权限。
- 文件的owner需要为执行命令的用户。
- 当前用户需要对该文件有读、写权限。
- 要求文件的格式与表指定的存储格式相同。如创建表时指定stored as rcfile，但是文件格式为txt，则不符合要求。
- 文件名不能以下横线（_）或点（.）开头，以这些开头的文件会被忽略。

16.10.13 执行 create external table 报错

问题现象

执行命令：create external table xx(xx int) stored as textfile location '/tmp/aaa/aaa'，报以下错误：

```
Permission denied. Principal [name=fantasy, type=USER] does not have following privileges on Object [type=DFS_URI, name=/tmp/aaa/aaa] for operation CREATETABLE : [SELECT, INSERT, DELETE, OBJECT OWNERSHIP] (state=42000,code=40000)
```

原因分析

当前登录的用户不具备该目录或者其父目录的读写权限。创建外部表时，会判断当前用户对指定的目录以及该目录下其它目录和文件是否有读写权限，如果该目录不存在，会去判断其父目录，依次类推。如果一直不满足就会报权限不足。而不是报指定的目录不存在。

解决方案

请确认当前用户为路径“/tmp/aaa/aaa”的owner有读写权限，如果该路径不存在，确认对其父路径有读写权限。

16.10.14 在 beeline 客户端执行 dfs -put 命令报错

问题现象

执行命令：

dfs -put /opt/kv1.txt /tmp/kv1.txt

报以下错误：

```
Permission denied. Principal [name=admin, type=USER] does not have following privileges onObject[type=COMMAND_PARAMS,name=[-put, /opt/kv1.txt, /tmp/kv1.txt]] for operation DFS : [ADMIN PRIVILEGE] (state=,code=1)
```

原因分析

当前登录的用户不具备操作此命令的权限。

解决方案

如果登录的当前用户具有admin角色，请用set role admin来切换成admin角色操作。
如果不具备admin角色，在Manager页面中给用户绑定对应角色的权限。

16.10.15 执行 set role admin 报无权限

问题现象

执行命令：

set role admin

报下述错误：

```
0: jdbc:hive2://192.168.42.26:21066/> set role admin;
Error: Error while processing statement: FAILED: Execution Error, return code 1 from
org.apache.hadoop.hive.ql.exec.DDLTask. dmp_B doesn't belong to role admin (state=08S01,code=1)
```

原因分析

当前登录的用户不具有Hive的admin角色的权限。

解决方案

步骤1 登录Manager。

- MRS 3.x之前版本，执行**步骤7**。
- MRS 3.x及之后版本，选择“集群 > 服务 > Hive”，在服务“概览”页面右上角单击“更多”，查看“启用Ranger鉴权”是否置灰。
  - 是，执行**步骤2**。
  - 否，执行**步骤7**。

步骤2 选择“集群 > 服务 > Ranger”，单击“基本信息”区域中的“RangerAdmin”，进入Ranger WebUI界面。

步骤3 单击右上角用户名后，选择“Log Out”，退出当前用户后使用**rangeradmin**用户登录。

步骤4 在首页中单击“Settings”，选择“Roles”。

步骤5 单击“Role Name”为“admin”的角色，在“Users”区域，单击“Select User”，选择指定用户名。

步骤6 单击Add Users按钮，在对应用户名所在行勾选“Is Role Admin”，单击“Save”保存配置，操作结束。

步骤7 选择“系统 > 权限 > 角色”，添加一个拥有Hive管理员权限的角色。

步骤8 在FusionInsight Manager页面，选择“系统 > 权限 > 用户 ”。

步骤9 在指定用户对应的“操作”列单击“修改”。

步骤10 为用户绑定拥有Hive管理员权限的角色，并单击“确定”完成权限添加。

----结束

16.10.16 通过 beeline 创建 UDF 时候报错

问题现象

执行命令：

create function fn_test3 as 'test.MyUDF' using jar 'hdfs:///tmp/udf2/MyUDF.jar'

报以下错误：

```
Error: Error while compiling statement: FAILED: HiveAccessControlException Permission denied: Principal [name=admin, type=USER] does not have following privileges for operation CREATEFUNCTION [[ADMIN PRIVILEGE] on Object [type=DATABASE, name=default], [ADMIN PRIVILEGE] on Object [type=FUNCTION, name=default.fn_test3]] (state=42000,code=40000)
```

原因分析

Hive中创建永久函数需要特殊的role admin。

解决方案

在执行语句前执行**set role admin命令**即可解决。

16.10.17 Hive 服务健康状态和 Hive 实例健康状态的区别

问题现象

Hive服务健康状态和Hive实例健康状态的区别是什么?

解决方案

Hive服务的健康状态（也就是在services界面看到的健康状态）有Good，Bad，Partially Healthy，Unknown四种状态，四种状态除了取决于Hive本身服务的可用性（会用简单的sql来检测Hive服务的可用性），还取决于Hive服务所依赖的其他组件的服务状态。

Hive实例分为Hiveserver和Metastore两种，健康状态有Good，Concerning，Unknown三种状态，这三种状态是通过jmx通信来判定，与实例通信正常时为Good，通信异常时为Concerning，无法通信时为Unknown。

16.10.18 Hive 中的告警有哪些以及触发的场景

Hive 中的告警

告警ID	告警级别	可自动清除	告警名称	告警类型
16000	Minor	TRUE	Percentage of Sessions Connected to the HiveServer to Maximum Number Allowed Exceeds the Threshold	故障告警
16001	Minor	TRUE	Hive Warehouse Space Usage Exceeds the Threshold	故障告警
16002	Minor	TRUE	The Successful Hive SQL Operations Lower than The Threshold	故障告警
16004	Critical	TRUE	Hive Service Unavailable	故障告警

告警触发场景

- 16000：当连接HiveServer的session数占允许连接总数的比率超过设定的阈值的时候触发告警。如连接的session数为9，总连接数为12，设定的阈值为70%，9/12 > 70%便触发告警。
- 16001：当Hive使用的HDFS容量占分配给Hive的HDFS总容量的比率超过设定的阈值时触发告警。如分配给Hive的是500G，Hive已经使用400G，设定的阈值时75%，400/500 > 75%便触发告警。
- 16002：当执行SQL的成功率低于设定的阈值时变触发告警。如你执行了4条失败了2条，设定的阈值为60%，成功率2/4 < 60%便触发告警。
- 16004：Hive服务的健康状态变为Bad时触发告警。

说明

- MRS Manager界面操作：告警的阈值和告警的级别以及触发告警的时间段可以在MRS Manager界面的“系统设置 > 阈值配置”中设定。FusionInsight Manager界面操作：告警的阈值和告警的级别以及触发告警的时间段可以在FusionInsight Manager界面的“运维 > 告警 > 阈值设置”中设定。
- Hive运行相关的指标可以在Hive监控界面查看。

16.10.19 Shell 客户端连接提示"authentication failed"

问题现象

安全集群中，HiveServer服务正常的情况下，Shell客户端中执行beeline命令失败，界面提示“authentication failed”，如下：

```
Debug is true storeKey false useTicketCache true useKeyTab false doNotPrompt false ticketCache is null isInitiator true KeyTab is null refreshKrb5Config is false principal is null tryFirstPass is false useFirstPass is false storePass is false clearPass is false
Acquire TGT from Cache
Credentials are no longer valid
Principal is null
null credentials from Ticket Cache
[Krb5LoginModule] authentication failed
No password provided
```

可能原因

- 客户端用户没有进行安全认证
- kerberos认证超期

解决方案

步骤1 登录Hive客户端所在节点。

步骤2 执行**source** *集群客户端安装目录*/**bigdata_env**命令。

可通过**klist**命令查看本地是否有有效票据，如下信息表明票据在16年12月24日14:11:42生效，将在16年12月25日14:11:40失效。在此期间可以使用该票据，其他时间则该票据无效。

```
klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: xxx@HADOOP.COM
Valid starting     Expires            Service principal
12/24/16 14:11:42  12/25/16 14:11:40  krbtgt/HADOOP.COM@HADOOP.COM
```

步骤3 执行**kinit** *username*进行认证，然后再使用客户端。

----结束

16.10.20 客户端提示访问 ZooKeeper 失败

问题现象

安全集群中，HiveServer服务正常的情况下，通过jdbc接口连接HiveServer执行sql时报出ZooKeeper认证异常"The ZooKeeper client is AuthFailed"，如下：

```
14/05/19 10:52:00 WARN utils.HAClientUtilDummyWatcher: The ZooKeeper client is AuthFailed
 14/05/19 10:52:00 INFO utils.HiveHAClientUtil: Exception thrown while reading data from znode.The
```

```
possible reason may be connectionless. This is recoverable. Retrying..
 14/05/19 10:52:16 WARN utils.HAClientUtilDummyWatcher: The ZooKeeper client is AuthFailed
 14/05/19 10:52:32 WARN utils.HAClientUtilDummyWatcher: The ZooKeeper client is AuthFailed
 14/05/19 10:52:32 ERROR st.BasicTestCase: Exception: Could not establish connection to active hiveserver
 java.sql.SQLException: Could not establish connection to active hiveserver
```

或者报出无法读取"Hiveserver2 configs from ZooKeeper"，如下：

```
Exception in thread "main" java.sql.SQLException: org.apache.hive.jdbc.ZooKeeperHiveClientException:
Unable to read HiveServer2 configs from ZooKeeper
at org.apache.hive.jdbc.HiveConnection.<init>(HiveConnection.java:144)
at org.apache.hive.jdbc.HiveDriver.connect(HiveDriver.java:105)
at java.sql.DriverManager.getConnection(DriverManager.java:664)
at java.sql.DriverManager.getConnection(DriverManager.java:247)
at JDBCExample.main(JDBCExample.java:82)
Caused by: org.apache.hive.jdbc.ZooKeeperHiveClientException: Unable to read HiveServer2 configs from
ZooKeeper
at
org.apache.hive.jdbc.ZooKeeperHiveClientHelper.configureConnParams(ZooKeeperHiveClientHelper.java:100)
at org.apache.hive.jdbc.Utils.configureConnParams(Utils.java:509)
at org.apache.hive.jdbc.Utils.parseURL(Utils.java:429)
at org.apache.hive.jdbc.HiveConnection.<init>(HiveConnection.java:142)
... 4 more
Caused by: org.apache.zookeeper.KeeperException$ConnectionLossException: KeeperErrorCode =
ConnectionLoss for /hiveserver2
at org.apache.zookeeper.KeeperException.create(KeeperException.java:99)
at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
at org.apache.zookeeper.ZooKeeper.getChildren(ZooKeeper.java:2374)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl$3.call(GetChildrenBuilderImpl.java:214)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl$3.call(GetChildrenBuilderImpl.java:203)
at org.apache.curator.RetryLo，op.callWithRetry(RetryLoop.java:107)
at
org.apache.curator.framework.imps.GetChildrenBuilderImpl.pathInForeground(GetChildrenBuilderImpl.java:2
00)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl.forPath(GetChildrenBuilderImpl.java:191)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl.forPath(GetChildrenBuilderImpl.java:38)
```

可能原因

- 客户端连接HiveServer时，HiveServer的地址是从ZooKeeper中自动获取，当ZooKeeper连接认证异常时，无法从ZooKeeper中获取正确的HiveServer地址。
- 在连接zookeeper认证时，需要客户端传入krb5.conf，principal，keytab等相关信息。认证失败有如下几种：
  - user.keytab路径写错。
  - user.principal写错。
  - 集群做过切换域名操作但客户端拼接url时使用旧的principal。
  - 有防火墙相关设置，导致客户端本身无法通过kerberos认证，Kerberos需要开放的端口有21730(TCP)、21731(TCP/UDP)、21732(TCP/UDP)。

解决方案

步骤1 确保用户可以正常读取客户端节点相关路径下的user.keytab文件。

步骤2 确保用户的user.principal与指定的keytab文件对应。

可通过**klist -kt keytabpath/user.keytab**查看。

步骤3 如果集群有做过切换域名操作，需要保证url中使用的principal字段是新域名。

如默认为hive/hadoop.hadoop.com@HADOOP.COM，当集群有切换域名的操作时，该字段需要进行相关修改。如域名为abc.com时，则此处应填写hive/hadoop.abc.com@ABC.COM。

步骤4 确保可以正常的认证连接HiveServer。

在客户端执行以下命令

source *客户端安装目录*/**bigdata_env**

kinit *username*

然后再使用客户端执行**beeline**，确保可以正常运行。

----结束

16.10.21 使用 udf 函数提示"Invalid function"

问题现象

在 Hive客户端中使用Spark创建UDF函数时，报出"ERROR 10011","invalid function"的异常，如下：

```
Error: Error while compiling statement: FAILED: SemanticException [Error 10011]: Line 1:7 Invalid function 'test_udf' (state=42000,code=10011)
```

在多个HiveServer之间使用UDF也存在上述问题。例如，在HiveServer1中使用HiverServer2创建的UDF，如果不及时同步元数据信息，连接HiveServer1的客户端也会提示上述错误信息。

可能原因

多个HiveServer之间或者Hive与Spark之间共用的元数据未同步，导致不同HiveServer实例内存数据不一致，造成UDF不生效。

解决方案

需要将新建的UDF信息同步到HiveServer中，执行reload function操作即可。

16.10.22 Hive 服务状态为 Unknown 总结

可能原因

Hive服务停止。

解决方案

重启Hive服务。

16.10.23 Hiveserver 或者 Metastore 实例的健康状态为 unknown

问题现象

hiveserver或者metastore实例的健康状态为unknown。

可能原因

hiveserver或者metastore实例被停止。

解决方案

重启hiveserver或者metastore实例。

16.10.24 Hiveserver 或者 Metastore 实例的健康状态为 Concerning

问题现象

Hiveserver或者Metastore实例的健康状态为Concerning。

可能原因

hiveserver或者metastore实例在启动的时候发生异常，无法正常启动。如，当修改MetaStore/HiveServer GC参数时，可通过查看对应进程的启动日志，如hiveserver.out(hadoop-omm-jar-192-168-1-18.out)文件排查。 如下异常：

Error: Could not find or load main class Xmx2048M

说明java虚拟机启动时，将Xmx2048M 作为java进程的启动参数而不是JVM的启动参数了，如下将符号‘-’误删掉。

```
METASTORE_GC_OPTS=Xms1024M Xmx2048M -DIgnoreReplayReqDetect
-XX\:CMSFullGCsBeforeCompaction\=1 -XX\:+UseConcMarkSweepGC
-XX\:+CMSParallelRemarkEnabled -XX\:+UseCMSCompactAtFullCollection
-XX\:+ExplicitGCInvokesConcurrent -server -XX\:MetaspaceSize\=128M
-XX\:MaxMetaspaceSize\=256M
```

解决方案

因此遇到此类异常应该检查最近的变更项，以确认是否配置有误。

```
METASTORE_GC_OPTS=Xms1024M -Xmx2048M -DIgnoreReplayReqDetect
-XX\:CMSFullGCsBeforeCompaction\=1 -XX\:+UseConcMarkSweepGC
-XX\:+CMSParallelRemarkEnabled -XX\:+UseCMSCompactAtFullCollection
-XX\:+ExplicitGCInvokesConcurrent -server -XX\:MetaspaceSize\=128M
-XX\:MaxMetaspaceSize\=256M
```

16.10.25 TEXTFILE 类型文件使用 ARC4 压缩时 select 结果乱码

问题现象

Hive查询结果表做压缩存储（ARC4），对结果表做select * 查询时返回结果为乱码。

可能原因

Hive默认压缩格式不是ARC4格式或者未开启输出压缩。

解决方案

步骤1 在select结果乱码时，在beeline中进行如下设置。

set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.encryption.arc4.ARC4BlockCodec;

set hive.exec.compress.output=true;

步骤2 使用块解压的方式先将表导入一个新表中。

insert overwrite table tbl_result select * from tbl_source;

步骤3 再进行查询。

select * from tbl_result;

----结束

16.10.26 hive 任务运行过程中失败，重试成功

问题现象

当hive任务在正常运行时失败，在客户端报出错误，类似的错误打印：

```
Error:Invalid OperationHandler:OperationHander [opType=EXECUTE_STATEMENT,getHandleIdentifier()=XXX] (state=,code=0)
```

而此任务提交到yarn上的mapreduce任务运行成功。

```
0: jdbc:hive2://189.120.204.104:21066/> select count(*) from test1;
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1484563934624_0003
INFO  : Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 7 for admin)
INFO  : Kind: HIVE_DELEGATION_TOKEN, Service: HiveServer2ImpersonationToken, Ident: 00 05 61 64 6d 69 6e 05 61 64 6d 69 6e 21 68 69 76 65 2f 68 61 64 6f 6f 70 2e 68
 85 ce e4 8a 01 59 ce 92 52 e4 8e 07 d8 0c
INFO  : The url to track the job: https://189-120-204-104:26001/proxy/application_1484563934624_0003/
INFO  : Starting Job = job_1484563934624_0003, Tracking URL = https://189-120-204-104:26001/proxy/application_1484563934624_0003/
INFO  : Kill Command = /opt/huawei/Bigdata/FusionInsight-Hive-1.1.0/hadoop/bin/hadoop job  -kill job_1484563934624_0003
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2017-01-17 11:46:12,579 Stage-1 map = 0%,  reduce = 0%
INFO  : 2017-01-17 11:46:23,243 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.32 sec
Error: Invalid OperationHandle: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=386323de-dfla-4299-826e-96368d4baf80] (state=,code=0)
0: jdbc:hive2://189.120.204.215:21066/>
```

原因分析

出错的集群有两个hiveserver实例，首先查看其中一个hiveserver日志发现里面的报错与客户端中的错误一样均是Error:Invalid OperationHandler，查看另一个hiveserver发现在出错的时间段此实例有如下类似START_UP的打印，说明那段时间进程被停止过，后来又启动成功，提交的任务本来连接的是重启过的hiveserver实例，当这个实例被停止后，任务进程连接到另一个健康的hiveserver上导致报错。

```
2017-02-15 14:40:11,309 | INFO  | main | STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting HiveServer2
STARTUP_MSG:   host = XXX-120-85-154/XXX.120.85.154
STARTUP_MSG:   args = []
STARTUP_MSG:   version = 1.3.0
```

解决办法

重新提交一次任务即可，保证在任务执行期间不手动重启hiveserver进程。

16.10.27 执行 select 语句报错

问题现象

执行语句select count(*) from XXX;时客户端报错：Error:Error while processing statement :FAILED:Execution Error,return code 2 from ...

这个报错return code2说明是在执行mapreduce任务期间报错导致任务失败。

```
0: jdbc:hive2://134.160.37.21:21066/> select count(*) from src.gn_data_info_gz where day_id='18' and timenap='10';
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:496
INFO  : Submitting tokens for job: job_1482323187492_57815
INFO  : Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 1083948 for boncusermm)
INFO  : Kind: HIVE_DELEGATION_TOKEN, Service: HiveServer2ImpersonationToken, Ident: 00 0a 62 6f 6e 63 75 73 65 72 6d 6d 0a 62 6f 6e 63 75 73 65 72 6d 6d 21 68 69
74 55 8a 01 59 d4 b5 f8 55 8d 02 50 ae 8e 03 65
INFO  : The url to track the job: https://hnocno3:26001/proxy/application_1482323187492_57815/
INFO  : Starting Job = job_1482323187492_57815, Tracking URL = https://hnocno3:26001/proxy/application_1482323187492_57815/
INFO  : Kill Command = /opt/huawei/Bigdata/FusionInsight_V100R002C60U10/FusionInsight-Hive-1.3.0/hive-1.3.0/bin/../../hadoop/bin/hadoop job  -kill job_148232318
INFO  : Hadoop job information for Stage-1: number of mappers: 496; number of reducers: 1
INFO  : 2017-01-18 16:21:00,906 Stage-1 map = 0%,  reduce = 0%
INFO  : 2017-01-18 16:21:18,957 Stage-1 map = 1%,  reduce = 0%, Cumulative CPU 50.53 sec
INFO  : 2017-01-18 16:21:32,526 Stage-1 map = 2%,  reduce = 0%, Cumulative CPU 416.23 sec
INFO  : 2017-01-18 16:21:35,035 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 1421.09 sec
INFO  : 2017-01-18 16:21:36,331 Stage-1 map = 7%,  reduce = 0%, Cumulative CPU 2159.35 sec
INFO  : 2017-01-18 16:21:37,810 Stage-1 map = 9%,  reduce = 0%, Cumulative CPU 2548.77 sec
INFO  : 2017-01-18 16:21:39,126 Stage-1 map = 15%,  reduce = 0%, Cumulative CPU 3264.95 sec
INFO  : 2017-01-18 16:21:40,509 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 3621.79 sec
INFO  : 2017-01-18 16:21:41,710 Stage-1 map = 26%,  reduce = 0%, Cumulative CPU 3913.79 sec
INFO  : 2017-01-18 16:21:42,890 Stage-1 map = 32%,  reduce = 0%, Cumulative CPU 4202.18 sec
INFO  : 2017-01-18 16:21:44,037 Stage-1 map = 41%,  reduce = 0%, Cumulative CPU 4595.63 sec
INFO  : 2017-01-18 16:21:45,119 Stage-1 map = 49%,  reduce = 0%, Cumulative CPU 4822.15 sec
INFO  : 2017-01-18 16:21:46,213 Stage-1 map = 57%,  reduce = 0%, Cumulative CPU 5107.44 sec
INFO  : 2017-01-18 16:21:47,300 Stage-1 map = 68%,  reduce = 0%, Cumulative CPU 5405.71 sec
INFO  : 2017-01-18 16:21:48,407 Stage-1 map = 76%,  reduce = 0%, Cumulative CPU 5611.73 sec
INFO  : 2017-01-18 16:21:49,483 Stage-1 map = 85%,  reduce = 0%, Cumulative CPU 5804.64 sec
INFO  : 2017-01-18 16:21:50,565 Stage-1 map = 92%,  reduce = 0%, Cumulative CPU 5958.81 sec
INFO  : 2017-01-18 16:21:51,641 Stage-1 map = 96%,  reduce = 0%, Cumulative CPU 6041.06 sec
INFO  : 2017-01-18 16:21:52,744 Stage-1 map = 98%,  reduce = 0%, Cumulative CPU 6073.82 sec
INFO  : 2017-01-18 16:22:08,352 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6078.4 sec
INFO  : MapReduce Total cumulative CPU time: 0 days 1 hours 41 minutes 18 seconds 400 msec
ERROR : Ended Job = job_1482323187492_57815 with errors
Error: Error while processing statement: FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask (state=08S01,code=2)
0: jdbc:hive2://134.160.37.21:21066/>
```

原因分析

1. 进入yarn原生页面查看mapreduce任务的日志看到报错是无法识别到压缩方式导致错误，看文件后缀是gzip压缩，堆栈却报出是zlib方式。

```
2017-01-18 16:22:07,566 INFO [main] org.apache.hadoop.hive.ql.exec.MapOperator: 4 Close done
2017-01-18 16:22:07,572 WARN [main] org.apache.hadoop.mapred.YarnChild: Exception running child : java.io.IOException: java.io.IOException: unknown compression method
        at org.apache.hadoop.hive.io.HiveIOExceptionHandlerChain.handleRecordReaderNextException(HiveIOExceptionHandlerChain.java:121)
        at org.apache.hadoop.hive.io.HiveIOExceptionHandlerUtil.handleRecordReaderNextException(HiveIOExceptionHandlerUtil.java:77)
        at org.apache.hadoop.hive.ql.io.HiveContextAwareRecordReader.doNext(HiveContextAwareRecordReader.java:355)
        at org.apache.hadoop.hive.ql.io.HiveRecordReader.doNext(HiveRecordReader.java:79)
        at org.apache.hadoop.hive.ql.io.HiveRecordReader.doNext(HiveRecordReader.java:33)
        at org.apache.hadoop.hive.ql.io.HiveContextAwareRecordReader.next(HiveContextAwareRecordReader.java:116)
        at org.apache.hadoop.mapred.MapTask$TrackedRecordReader.moveToNext(MapTask.java:199)
        at org.apache.hadoop.mapred.MapTask$TrackedRecordReader.next(MapTask.java:185)
        at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:52)
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:453)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:180)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1726)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:174)
Caused by: java.io.IOException: unknown compression method
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.inflateBytesDirect(Native Method)
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.decompress(ZlibDecompressor.java:225)
        at org.apache.hadoop.io.compress.DecompressorStream.decompress(DecompressorStream.java:91)
        at org.apache.hadoop.io.compress.DecompressorStream.read(DecompressorStream.java:85)
        at java.io.InputStream.read(InputStream.java:101)
        at org.apache.hadoop.util.LineReader.fillBuffer(LineReader.java:180)
        at org.apache.hadoop.util.LineReader.readDefaultLine(LineReader.java:216)
        at org.apache.hadoop.util.LineReader.readLine(LineReader.java:174)
        at org.apache.hadoop.mapred.LineRecordReader.next(LineRecordReader.java:248)
        at org.apache.hadoop.mapred.LineRecordReader.next(LineRecordReader.java:48)
        at org.apache.hadoop.hive.ql.io.HiveContextAwareRecordReader.doNext(HiveContextAwareRecordReader.java:350)
        ... 13 more

2017-01-18 16:22:07,576 INFO [main] org.apache.hadoop.mapred.Task: Runnning cleanup for the task
```

2. 因此怀疑此语句查询的表对应的HDFS上的文件有问题，map日志中打印出了解析的对应的文件名，将其从HDFS上下载到本地，看到是gz结尾的文件，使用**tar**命令解压报错，格式不正确无法解压。使用file命令查看文件属性发现此文件来自于FAT系统的压缩而非unix。

```
[root@hn0dn01 ~]# ls -l *.txt.gz
-rw-r--r-- 1 root root 101966463 Jan 18 20:13 20170118095958920074010l.txt.gz
-rw-r--r-- 1 root root  90448283 Jan 18 19:55 20170118104000000074002O.txt.gz
[root@hn0dn01 ~]# file 20170118095958920074010l.txt.gz
20170118095958920074010l.txt.gz: gzip compressed data, was "20170118095958920074010l.txt", from Unix, last modified: Wed Jan 18 09:59:52 2017
[root@hn0dn01 ~]# file  20170118104000000074002O.txt.gz
20170118104000000074002O.txt.gz: gzip compressed data, from FAT filesystem (MS-DOS, OS/2, NT)
[root@hn0dn01 ~]# tar -zxvf  20170118104000000074002O.txt.gz
tar: This does not look like a tar archive
tar: Skipping to next header

gzip: stdin: decompression OK, trailing garbage ignored
tar: Child returned status 2
tar: Error is not recoverable: exiting now
[root@hn0dn01 ~]#
```

解决办法

将格式不正确的文件移除hdfs目录或者替换为正确的格式的文件。

16.10.28 drop partition 操作，有大量分区时操作失败

问题背景与现象

执行drop partitions 操作，执行异常：

```
MetaStoreClient lost connection. Attempting to reconnect. |
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.invoke(RetryingMetaStoreClient.java:187)
org.apache.thrift.transport.TTransportException
at org.apache.thrift.transport.TIOStreamTransport.read(TIOStreamTransport.java:132)
at org.apache.thrift.transport.TTransport.xxx(TTransport.java:86)
at org.apache.thrift.transport.TSaslTransport.readLength(TSaslTransport.java:376)
at org.apache.thrift.transport.TSaslTransport.readFrame(TSaslTransport.java:453)
at org.apache.thrift.transport.TSaslTransport.read(TSaslTransport.java:435)
...
```

查看对应metaStore日志，有StackOverFlow异常

```
2017-04-22 01:00:58,834 | ERROR | pool-6-thread-208 | java.lang.StackOverflowError
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:330)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
```

原因分析

drop partition的处理逻辑是将找到所有满足条件的分区，将其拼接起来，最后统一删除。由于分区数过多，拼删元数据堆栈较深，出现StackOverFlow异常。

解决办法

分批次删除分区。

16.10.29 localtask 启动失败

问题背景与现象

1. 执行join等操作，数据量较小时，会启动localtask执行，执行过程会报错：

   ```
   jdbc:hive2://10.*.*.*:21066/> select a.name ,b.sex from student a join student1 b on (a.name = b.name);
   ERROR : Execution failed with exit status: 1
   ERROR : Obtaining error information
   ERROR :
   Task failed!
   Task ID:
     Stage-4
   ...
   Error: Error while processing statement: FAILED: Execution Error, return code 1 from
   org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask (state=08S01,code=1)
   ...
   ```

2. 查看对应hiveserver日志，发现是启动localtask失败

   ```
   2018-04-25 16:37:19,296 | ERROR | HiveServer2-Background-Pool: Thread-79 | Execution failed with
   exit status: 1 | org.apache.hadoop.hive.ql.session.SessionState
   $LogHelper.printError(SessionState.java:1016)
   2018-04-25 16:37:19,296 | ERROR | HiveServer2-Background-Pool: Thread-79 | Obtaining error
   information | org.apache.hadoop.hive.ql.session.SessionState
   $LogHelper.printError(SessionState.java:1016)
   2018-04-25 16:37:19,297 | ERROR | HiveServer2-Background-Pool: Thread-79 |
   Task failed!
   Task ID:
     Stage-4
   Logs:
   ```

```
 | org.apache.hadoop.hive.ql.session.SessionState$LogHelper.printError(SessionState.java:1016)
2018-04-25 16:37:19,297 | ERROR | HiveServer2-Background-Pool: Thread-79 | /var/log/Bigdata/hive/
hiveserver/hive.log | org.apache.hadoop.hive.ql.session.SessionState
$LogHelper.printError(SessionState.java:1016)
2018-04-25 16:37:19,297 | ERROR | HiveServer2-Background-Pool: Thread-79 | Execution failed with
exit status: 1 |
org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask.executeInChildVM(MapredLocalTask.java:342)
2018-04-25 16:37:19,309 | ERROR | HiveServer2-Background-Pool: Thread-79 | FAILED: Execution
Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask |
org.apache.hadoop.hive.ql.session.SessionState$LogHelper.printError(SessionState.java:1016)
...
2018-04-25 16:37:36,438 | ERROR | HiveServer2-Background-Pool: Thread-88 | Error running hive
query:  | org.apache.hive.service.cli.operation.SQLOperation$1$1.run(SQLOperation.java:248)
org.apache.hive.service.cli.HiveSQLException: Error while processing statement: FAILED: Execution
Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask
        at org.apache.hive.service.cli.operation.Operation.toSQLException(Operation.java:339)
        at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:169)
        at org.apache.hive.service.cli.operation.SQLOperation.access$200(SQLOperation.java:75)
        at org.apache.hive.service.cli.operation.SQLOperation$1$1.run(SQLOperation.java:245)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1710)
        at org.apache.hive.service.cli.operation.SQLOperation$1.run(SQLOperation.java:258)
        at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)
```

3. 查看对应hiveserver日志目录(/var/log/Bigdata/hive/hiveserver)下，hs_err_pid_*****.log，发现有内存不够的错误

```
# There is insufficient memory for the Java Runtime Environment to continue.
# Native memory allocation (mmap) failed to map 20776943616 bytes for committing reserved
memory.
 ...
```

原因分析

Hive在执行join操作，数据量小时会生成MapJoin，执行MapJoin时会生成localtask任务，localtask启动的jvm内存继承了父进程的内存。

当有多个join执行的时候，启动多个localtask，如果机器内存不够，就会导致启动localtask失败。

解决办法

步骤1 搜索“hive.auto.convert.join”参数并修改hive的配置hive.auto.convert.join 为false，保存配置并重启服务。

该参数修改后会对业务性能有一定影响。继续执行后续步骤可不影响业务性能。

步骤2 搜索“HIVE_GC_OPTS”参数并修改 hive的HIVE_GC_OPTS，把Xms调小，具体要根据业务评估，最小设置为Xmx的一半，修改完后保存配置并重启服务。

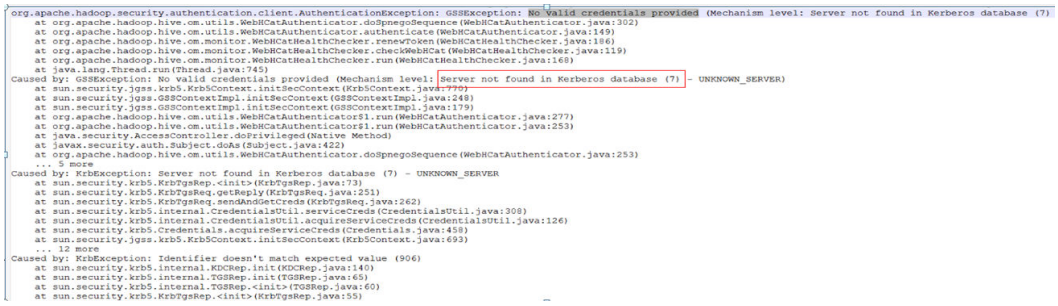
----结束

16.10.30 WebHCat 启动失败

问题背景与现象

用户修改hostname导致WebHCat启动失败。

查看对应节点WebHCat启动日志（ /var/log/Bigdata/hive/webhcat/hive.log ），发现报如下错误：



原因分析

1. MRSwebhcat角色的服务端账户中涉及到hostname，如果安装完后再修改hostname，就会导致启动失败。
2. /etc/hosts中配置了一对多或者多对一的主机名和IP对应关系，导致在执行**hostname**和**hostname -i**获取不到正确的IP和hostname。

解决办法

步骤1 将修改了节点的hostname全部修改为集群安装前的hostname。

步骤2 排查WebHCat所在节点的/etc/hosts是否配置正确。

步骤3 重启WebHCat。

----结束

16.10.31 切域后 Hive 二次开发样例代码报错

问题背景与现象

hive的二次开发代码样例运行报No rules applied to ****的错误：

原因分析

1. hive的二次开发样例代码会加载core-site.xml，此文件默认是通过classload加载，所以使用的时候要把此配置文件放到启动程序的classpath路径下面。
2. 如果修改了集群的域名，那么core-site.xml将发生变化，需要下载最新的core-site.xml并放入到打包hive二次开发样例代码进程的classpath路径下面。

解决办法

步骤1 下载集群Hive最新的客户端，获取最新的core-site.xml。

步骤2 将core-site.xml放入到打包hive二次开发样例代码进程的classpath路径下面。

----结束

16.10.32 DBService 超过最大连接数，导致 metastore 异常

问题背景与现象

DBService默认最大连接数是300，如果当业务量比较大，导致连接DBService的最大连接数超过300时，metastore会出现异常，并报slots are reserved for non-replication superuser connections的错误：

```
2018-04-26 14:58:55,657 | ERROR | BoneCP-pool-watch-thread | Failed to acquire connection to jdbc:postgresql://10.*.*.*:20051/hivemeta?socketTimeout=60. Sleeping for 1000 ms. Attempts left: 9 | com.jolbox.bonecp.BoneCP.obtainInternalConnection(BoneCP.java:292)
org.postgresql.util.PSQLException: FATAL: remaining connection slots are reserved for non-replication superuser connections
    at org.postgresql.core.v3.ConnectionFactoryImpl.readStartupMessages(ConnectionFactoryImpl.java:643)
    at org.postgresql.core.v3.ConnectionFactoryImpl.openConnectionImpl(ConnectionFactoryImpl.java:184)
    at org.postgresql.core.ConnectionFactory.openConnection(ConnectionFactory.java:64)
    at org.postgresql.jdbc2.AbstractJdbc2Connection.<init>(AbstractJdbc2Connection.java:124)
    at org.postgresql.jdbc3.AbstractJdbc3Connection.<init>(AbstractJdbc3Connection.java:28)
    at org.postgresql.jdbc3g.AbstractJdbc3gConnection.<init>(AbstractJdbc3gConnection.java:20)
    at org.postgresql.jdbc4.AbstractJdbc4Connection.<init>(AbstractJdbc4Connection.java:30)
    at org.postgresql.jdbc4.Jdbc4Connection.<init>(Jdbc4Connection.java:22)
    at org.postgresql.Driver.makeConnection(Driver.java:392)
    at org.postgresql.Driver.connect(Driver.java:266)
    at java.sql.DriverManager.getConnection(DriverManager.java:664)
    at java.sql.DriverManager.getConnection(DriverManager.java:208)
    at com.jolbox.bonecp.BoneCP.obtainRawInternalConnection(BoneCP.java:361)
    at com.jolbox.bonecp.BoneCP.obtainInternalConnection(BoneCP.java:269)
    at com.jolbox.bonecp.ConnectionHandle.<init>(ConnectionHandle.java:242)
    at com.jolbox.bonecp.PoolWatchThread.fillConnections(PoolWatchThread.java:115)
    at com.jolbox.bonecp.PoolWatchThread.run(PoolWatchThread.java:82)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)
```

原因分析

业务量大导致连接DBService的最大连接数超过了300，需要修改DBService的最大连接数。

解决办法

步骤1 搜索dbservice.database.max.connections配置项，并修改dbservice.database.max.connections配置的值到合适值，不能超过1000。

步骤2 保存配置，并重启受影响的服务或者实例。

步骤3 如果调整完还报超过最大连接数，需要排查业务代码，是否有连接泄露。

----结束

16.10.33 beeline 报 Failed to execute session hooks: over max connections 错误

问题背景与现象

HiveServer连接的最大连接数默认为200，当超过200时，beeline会报Failed to execute session hooks: over max connections

```
beeline> [root@172-27-16-38 c70client]# beeline
Connecting to
jdbc:hive2://129.188.82.38:24002,129.188.82.36:24002,129.188.82.35:24002/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2;sasl.qop=auth-conf;auth=KERBEROS;principal=hive/hadoop.hadoop.com@HADOOP.COM
Debug is  true storeKey false useTicketCache true useKeyTab false doNotPrompt false ticketCache is null isInitiator true KeyTab is null refreshKrb5Config is false principal is null tryFirstPass is false useFirstPass is false storePass is false clearPass is false
Acquire TGT from Cache
Principal is xxx@HADOOP.COM
Commit Succeeded

Error: Failed to execute session hooks: over max connections. (state=,code=0)
Beeline version 1.2.1 by Apache Hive
```

查看hiveserver日志(/var/log/Bigdata/hive/hiveserver/hive.log)报over max connections错误

```
2018-05-03 04:31:56,728 | WARN  | HiveServer2-Handler-Pool: Thread-137 | Error opening session:  | org.apache.hive.service.cli.thrift.ThriftCLIService.OpenSession(ThriftCLIService.java:542)
org.apache.hive.service.cli.HiveSQLException: Failed to execute session hooks: over max connections.
    at org.apache.hive.service.cli.session.SessionManager.openSession(SessionManager.java:322)
    at org.apache.hive.service.cli.CLIService.openSessionWithImpersonation(CLIService.java:189)
    at org.apache.hive.service.cli.thrift.ThriftCLIService.getSessionHandle(ThriftCLIService.java:663)
    at org.apache.hive.service.cli.thrift.ThriftCLIService.OpenSession(ThriftCLIService.java:527)
    at org.apache.hive.service.cli.thrift.TCLIService$Processor$OpenSession.getResult(TCLIService.java:1257)
    at org.apache.hive.service.cli.thrift.TCLIService$Processor$OpenSession.getResult(TCLIService.java:1242)
    at org.apache.thrift.ProcessFunction.process(ProcessFunction.java:39)
    at org.apache.thrift.TBaseProcessor.process(TBaseProcessor.java:39)
    at org.apache.hadoop.hive.thrift.HadoopThriftAuthBridge$Server$TUGIAssumingProcessor.process(HadoopThriftAuthBridge.java:710)
    at org.apache.thrift.server.TThreadPoolServer$WorkerProcess.run(TThreadPoolServer.java:286)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)
Caused by: org.apache.hive.service.cli.HiveSQLException: over max connections.
    at org.apache.hadoop.hive.transporthook.SessionControllerTsaslTransportHook.checkTotalSessionNumber(SessionControllerTsaslTransportHook.java:208)
    at org.apache.hadoop.hive.transporthook.SessionControllerTsaslTransportHook.postOpen(SessionControllerTsaslTransportHook.java:163)
    at org.apache.hadoop.hive.transporthook.SessionControllerTsaslTransportHook.run(SessionControllerTsaslTransportHook.java:134)
    at org.apache.hive.service.cli.session.SessionManager.executeSessionHooks(SessionManager.java:432)
    at org.apache.hive.service.cli.session.SessionManager.openSession(SessionManager.java:314)
    ... 12 more
```

原因分析

业务量大导致连接HiveServer单个节点最大连接数超过了200，需要调大连接HiveServer实例的最大连接数。

解决办法

步骤1 搜索hive.server.session.control.maxconnections配置项，并修改hive.server.session.control.maxconnections配置的值到合适值，不能超过1000。

步骤2 保存配置并重启受影响的服务或者实例。

----结束

16.10.34 beeline 报 OutOfMemoryError 错误

问题背景与现象

beeline客户端查询大量数据时，报OutOFMemoryError：Java heap space，具体报错信息如下：

```
org.apache.thrift.TException: Error in calling method FetchResults
    at org.apache.hive.jdbc.HiveConnection$SynchronizedHandler.invoke(HiveConnection.java:1514)
    at com.sun.proxy.$Proxy4.FetchResults(Unknown Source)
    at org.apache.hive.jdbc.HiveQueryResultSet.next(HiveQueryResultSet.java:358)
    at org.apache.hive.beeline.BufferedRows.<init>(BufferedRows.java:42)
    at org.apache.hive.beeline.BeeLine.print(BeeLine.java:1856)
    at org.apache.hive.beeline.Commands.execute(Commands.java:873)
    at org.apache.hive.beeline.Commands.sql(Commands.java:714)
    at org.apache.hive.beeline.BeeLine.dispatch(BeeLine.java:1035)
    at org.apache.hive.beeline.BeeLine.execute(BeeLine.java:821)
    at org.apache.hive.beeline.BeeLine.begin(BeeLine.java:778)
    at org.apache.hive.beeline.BeeLine.mainWithInputRedirection(BeeLine.java:486)
    at org.apache.hive.beeline.BeeLine.main(BeeLine.java:469)
Caused by: java.lang.OutOfMemoryError: Java heap space
    at com.sun.crypto.provider.CipherCore.doFinal(CipherCore.java:959)
    at com.sun.crypto.provider.CipherCore.doFinal(CipherCore.java:824)
    at com.sun.crypto.provider.AESCipher.engineDoFinal(AESCipher.java:436)
    at javax.crypto.Cipher.doFinal(Cipher.java:2223)
    at sun.security.krb5.internal.crypto.dk.AesDkCrypto.decryptCTS(AesDkCrypto.java:414)
    at sun.security.krb5.internal.crypto.dk.AesDkCrypto.decryptRaw(AesDkCrypto.java:291)
    at sun.security.krb5.internal.crypto.Aes256.decryptRaw(Aes256.java:86)
    at sun.security.jgss.krb5.CipherHelper.aes256Decrypt(CipherHelper.java:1397)
    at sun.security.jgss.krb5.CipherHelper.decryptData(CipherHelper.java:576)
    at sun.security.jgss.krb5.WrapToken_v2.getData(WrapToken_v2.java:130)
    at sun.security.jgss.krb5.WrapToken_v2.getData(WrapToken_v2.java:105)
    at sun.security.jgss.krb5.Krb5Context.unwrap(Krb5Context.java:1058)
    at sun.security.jgss.GSSContextImpl.unwrap(GSSContextImpl.java:403)
    at com.sun.security.sasl.gsskerb.GssKrb5Base.unwrap(GssKrb5Base.java:77)
    at org.apache.thrift.transport.TSaslTransport$SaslParticipant.unwrap(TSaslTransport.java:559)
    at org.apache.thrift.transport.TSaslTransport.readFrame(TSaslTransport.java:462)
    at org.apache.thrift.transport.TSaslTransport.read(TSaslTransport.java:435)
    at org.apache.thrift.transport.TSaslClientTransport.read(TSaslClientTransport.java:37)
    at org.apache.thrift.transport.TTransport.xxx(TTransport.java:86)
    at org.apache.hadoop.hive.thrift.TFilterTransport.xxx(TFilterTransport.java:62)
    at org.apache.thrift.protocol.TBinaryProtocol.xxx(TBinaryProtocol.java:429)
    at org.apache.thrift.protocol.TBinaryProtocol.readI32(TBinaryProtocol.java:318)
    at org.apache.thrift.protocol.TBinaryProtocol.readMessageBegin(TBinaryProtocol.java:219)
    at org.apache.thrift.TServiceClient.receiveBase(TServiceClient.java:77)
    at org.apache.hive.service.cli.thrift.TCLIService$Client.recv_FetchResults(TCLIService.java:505)
    at org.apache.hive.service.cli.thrift.TCLIService$Client.FetchResults(TCLIService.java:492)
    at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.apache.hive.jdbc.HiveConnection$SynchronizedHandler.invoke(HiveConnection.java:1506)
    at com.sun.proxy.$Proxy4.FetchResults(Unknown Source)
    at org.apache.hive.jdbc.HiveQueryResultSet.next(HiveQueryResultSet.java:358)
Error: Error retrieving next row (state=,code=0)
```

原因分析

- 客户查询大量数据，数据量过大。
- 客户在检索数据时使用**select * from table_name;**，进行全表查询，表内数据过多。
- beeline默认启动内存128M，查询时返回结果集过大，导致beeline无法承载导致。

解决办法

步骤1 执行**select count(*) from table_name;**前确认需要查询的数据量大小，确认是否需要在beeline中显示如此数量级的数据。

步骤2 如数量在一定范围内需要显示，请调整hive客户端的jvm参数， 在hive客户端目录/Hive下的component_env中添加export HIVE_OPTS=-Xmx1024M(具体数值请根据业务调整)，并重新执行**source** *客户端目录*/**bigdata_env**配置环境变量。

----结束

16.10.35 输入文件数超出设置限制导致任务执行失败

问题背景与现象

Hive执行查询操作时报Job Submission failed with exception 'java.lang.RuntimeException(input file number exceeded the limits in the conf;input file num is: 2380435,max heap memory is: 16892035072,the limit conf is: 500000/4)'，此报错中具体数值根据实际情况会发生变化，具体报错信息如下:

```
ERROR : Job Submission failed with exception 'java.lang.RuntimeException(input file numbers exceeded the limits in the conf;
 input file num is: 2380435 ,
 max heap memory is: 16892035072 ,
 the limit conf is: 500000/4)'
java.lang.RuntimeException: input file numbers exceeded the limits in the conf;
 input file num is: 2380435 ,
 max heap memory is: 16892035072 ,
 the limit conf is: 500000/4
    at org.apache.hadoop.hive.ql.exec.mr.ExecDriver.checkFileNum(ExecDriver.java:545)
    at org.apache.hadoop.hive.ql.exec.mr.ExecDriver.execute(ExecDriver.java:430)
    at org.apache.hadoop.hive.ql.exec.mr.MapRedTask.execute(MapRedTask.java:137)
    at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:158)
    at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:101)
    at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:1965)
    at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:1723)
    at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1475)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1283)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1278)
    at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:167)
    at org.apache.hive.service.cli.operation.SQLOperation.access$200(SQLOperation.java:75)
    at org.apache.hive.service.cli.operation.SQLOperation$1$1.run(SQLOperation.java:245)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1710)
    at org.apache.hive.service.cli.operation.SQLOperation$1.run(SQLOperation.java:258)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)

Error: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask (state=08S01,code=1)
```

原因分析

MapReduce 任务提交前对输入文件数的检查策略：在提交的MapReduce 任务中，允许的最大输入文件数和HiveServer最大堆内存的比值，例如500000/4（默认值），表示每4GB堆内存最大允许500000个输入文件。在输入的文件数超出此限制时则会发生此错误。

解决办法

步骤1 搜索hive.mapreduce.input.files2memory配置项，并修改hive.mapreduce.input.files2memory配置的值到合适值，根据实际内存和任务情况对此值进行调整。

步骤2 保存配置并重启受影响的服务或者实例。

步骤3 如调整后问题仍未解决，请根据业务情况调整HiveServer的GC参数至合理的值。

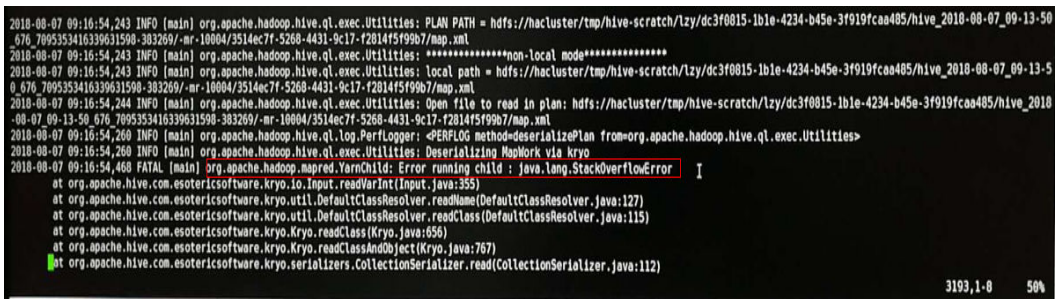
----结束

16.10.36 任务执行中报栈内存溢出导致任务执行失败

问题背景与现象

Hive执行查询操作时报错**Error running child : java.lang.StackOverflowError**，具体报错信息如下：

```
FATAL [main] org.apache.hadoop.mapred.YarnChild: Error running child : java.lang.StackOverflowError
at org.apache.hive.come.esotericsoftware.kryo.io.Input.readVarInt(Input.java:355)
at
org.apache.hive.come.esotericsoftware.kryo.util.DefautClassResolver.readName(DefautClassResolver.java:127)
at
org.apache.hive.come.esotericsoftware.kryo.util.DefautClassResolver.readClass(DefautClassResolver.java:115)
at org.apache.hive.come.esotericsoftware.kryo.Kryo.readClass(Kryo.java.656)
at org.apache.hive.come.esotericsoftware.kryo.kryo.readClassAnd0bject(Kryo.java:767)
at
org.apache.hive.come.esotericsoftware.kryo.serializers.collectionSerializer.read(CollectionSerializer.java:112)
```



原因分析

java.lang.StackOverflowError这是内存溢出错误的一种，即线程栈的溢出，方法调用层次过多（比如存在无限递归调用）或线程栈太小都会导致此报错。

解决办法

通过调整mapreduce阶段的map和reduce子进程JVM参数中的栈内存解决此问题，主要涉及参数为mapreduce.map.java.opts（调整map的栈内存）和

mapreduce.reduce.java.opts（调整reduce的栈内存），调整方法如下（以mapreduce.map.java.opts参数为例）。

- 临时增加map内存（只针对此次beeline生效）：

  在beeline中执行如下命令**set mapreduce.map.java.opts=-Xss8G;**（具体数值请结合实际业务情况进行调整）。

- 永久增加map内存mapreduce.map.memory.mb和mapreduce.map.java.opts的值：

  a. 在hiveserver自定义参数界面添加自定义参数mapreduce.map.java.opts及相应的值。

  b. 保存配置并重启受影响的服务或者实例。

  修改配置后需要保存，请注意参数在HiveServer自定义参数处修改，保存重启后生效（重启期间hive服务不可用），请注意执行时间窗口。

16.10.37 对同一张表或分区并发写数据导致任务失败

问题背景与现象

Hive执行插入语句时，报错HDFS上文件或目录已存在或被清除，具体报错如下：

```
2019-03-18 14:34:23,016 | WARN  | HiveServer2-Background-Pool: Thread-1179606 | Failed to move to trash: hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f/000000_0; Force to delete it. | org.apache.hadoop.hive.common.FileUtils.moveToTrash(FileUtils.java:651)
2019-03-18 14:34:23,017 | INFO  | HiveServer2-Background-Pool: Thread-1179604 | Moved to trash: hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp6_f/000000_0 | org.apache.hadoop.hive.common.FileUtils.moveToTrash(FileUtils.java:644)
2019-03-18 14:34:23,017 | ERROR | HiveServer2-Background-Pool: Thread-1179606 | Failed to delete hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f/000000_0 | org.apache.hadoop.hive.common.FileUtils.moveToTrash(FileUtils.java:660)
2019-03-18 14:34:23,017 | ERROR | HiveServer2-Background-Pool: Thread-1179606 | Failed with exception Destination directory hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f has not be cleaned up.
org.apache.hadoop.hive.ql.metadata.HiveException: Destination directory hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f has not be cleaned up.
at org.apache.hadoop.hive.ql.metadata.Hive.replaceFiles(Hive.java:2974)
at org.apache.hadoop.hive.ql.metadata.Hive.loadTable(Hive.java:1664)
at org.apache.hadoop.hive.ql.exec.MoveTask.execute(MoveTask.java:374)
at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:158)
at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:101)
```

原因分析

1. 根据HiveServer的审计日志，确认该任务的开始时间和结束时间。
2. 在上述时间区间内，查找是否有对同一张表或分区进行插入数据的操作。
3. Hive不支持对同一张表或分区进行并发数据插入，这样会导致多个任务操作同一个数据临时目录，一个任务将另一个任务的数据移走，导致任务失败。

解决办法

修改业务逻辑，单线程插入数据到同一张表或分区。

16.10.38 Hive 任务失败，报没有 HDFS 目录的权限

问题背景与现象

Hive任务报错，提示执行用户没有HDFS目录权限

```
2019-04-09 17:49:19,845 | ERROR | HiveServer2-Background-Pool: Thread-3160445 | Job Submission failed with exception 'org.apache.hadoop.security.AccessControlException(Permission denied: user=hive_quanxian, access=READ_EXECUTE, inode="/user/hive/warehouse/bigdata.db/gd_ga_wa_swryswjl":zhongao:hive:drwx------
at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkAccessAcl(FSPermissionChecker.java:426)
at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:329)
at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkSubAccess(FSPermissionChecker.java:300)
```

```
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:24
1)
at
com.xxx.hadoop.adapter.hdfs.plugin.HWAccessControlEnforce.checkPermission(HWAccessControlEnforce.java:
69)
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:19
0)
at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1910)
at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1894)
at
org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getContentSummary(FSDirStatAndListingO
p.java:135)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getContentSummary(FSNamesystem.java:3983)
at
org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getContentSummary(NameNodeRpcServer.ja
va:1342)
at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getContentSummary(Cli
entNamenodeProtocolServerSideTranslatorPB.java:925)
at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol
$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:616)
at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:973)
at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2260)
at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2256)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1781)
at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2254)
)'
```

原因分析

1. 根据堆栈信息，可以看出在检查子目录的权限时失败。
   ```
   org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkSubAccess(FSPermissionChecker.ja
   va:300)
   ```
2. 检查HDFS上表目录下所有文件目录的权限，发现有一个目录权限为700（只有文件属主能够访问），确认存在异常目录。

解决办法

1. 确认该文件是否为手动异常导入，如不是数据文件或目录，删除该文件或目录。
2. 当无法删除时，建议修改文件或目录权限为770。

16.10.39 Load 数据到 Hive 表失败

问题背景与现象

用户在建表成功后，通过Load命令往此表导入数据，但导入操作中遇到如下问题：

```
.....
> LOAD DATA INPATH '/user/tester1/hive-data/data.txt' INTO TABLE employees_info;
Error: Error while compiling statement: FAILED: SemanticException Unable to load data to destination table.
Error: The file that you are trying to load does not match the file format of the destination table.
(state=42000,code=40000)
.....
```

原因分析

1. 经分析，发现在建表时没有指定存储格式，所以采用了缺省存储格式RCFile。
2. 在导入数据时，被导入数据格式是TEXTFILE格式，最终导致此问题。

解决办法

属于应用侧问题，解决办法有多种。只要保证表所指定存储格式和被导入数据格式是一致的，可以根据实际情况采用合适方法。

- 方法1：

  可以使用具有Hive表操作权限的用户在建表时指定存储格式，例如：

  CREATE TABLE IF NOT EXISTS employees_info(name STRING,age INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

  指定被导入数据格式为TEXTFILE。
- 方法2：

  被导入数据存储格式不能为TEXTFILE，而应为RCFile。

16.10.40 HiveServer 和 HiveHCat 进程故障

用户问题

客户集群HiveServer和WebHCat进程状态均为故障。

问题现象

客户MRS集群Master2节点上的HiveServer和WebHCat进程状态显示为故障，重启之后仍为故障状态。

原因分析

在Manager界面单独启动故障的hiveserver进程，登录后台查找hiveserver.out日志中对应时间点的报错，报错信息为：error parsing conf mapred-site.xml 和 Premature end of file。然后重启webhcat也发现同样报错，原因即为解析mapred-site.xml文件错误。

处理步骤

1. 以root用户登录Master2节点。
2. 执行**find / -name 'mapred-site.xml'**命令获取mapred-site.xml文件所在位置。
   - hiveserver对应路径为/opt/Bigdata/*集群版本*/1_13_HiveServer/etc/mapred-site.xml，
   - webhcat对应路径为/opt/Bigdata/*集群版本*/1_13_WebHCat/etc/mapred-site.xml。
3. 确认mapred-site.xml文件是否有异常，该案例中该配置文件内容为空导致解析失败。
4. 修复mapred-site.xml文件，将Master1节点上对应目录下的配置文件用scp命令拷贝到Master2节点对应目录替换原文件.

5. 执行**chown omm:wheel mapred-site.xml**命令更改所属组和用户。
6. 在Manager界面重启故障的HiveServer和WebHCat进程，恢复正常。

16.10.41 Hive 执行 insert into 语句报错，命令界面报错信息不明

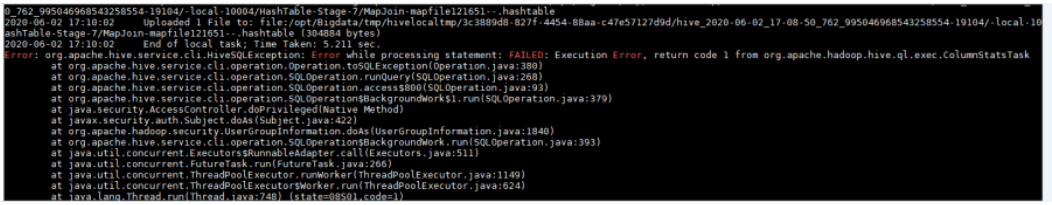
用户问题

客户使用MRS Hive执行一条SQL报错。

问题现象

客户使用MRS Hive执行一条SQL，有如下报错：

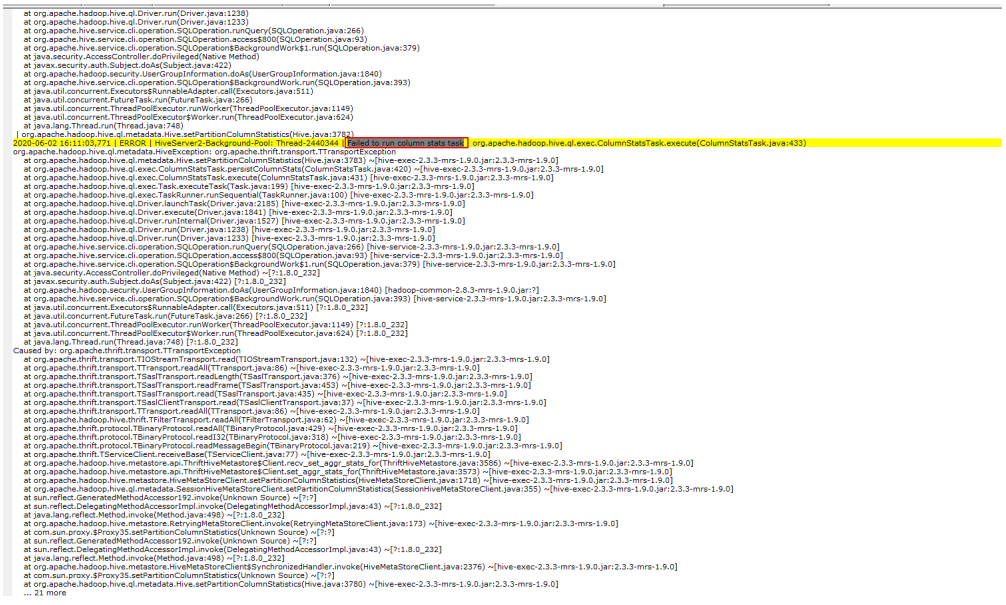
图 16-38 使用 MRS Hive 执行 SQL 报错



原因分析

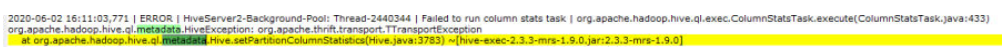
1. 查看Hiveserver日志，在对应时间点，有如下的报错信息。

   图 16-39 Hiveserver 日志

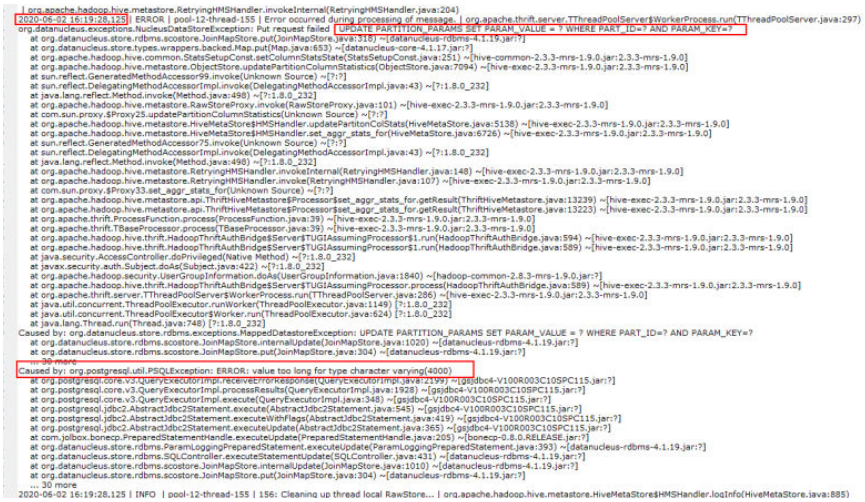
2. 在如上报错信息中未发现重要信息，但从堆栈中发现metadata字样，怀疑报错是和metastore有关。

   图 16-40 堆栈中 metadata 字样

3. 查看metastore日志，发现如下报错。

图 16-41 metastore 日志



查看如上错误的上下文，确定是本次执行SQL的报错，在报错信息里面发现如下内容：

```
Caused by: org.postgresql.util.PSQLException: ERROR: value too long for type character varying(4000)
```

确认是客户该条SQL对表的操作，所有列的字节长度超过4000的限制，导致SQL执行失败，需要修改该限制。

处理步骤

步骤1 以root用户登录集群任意一个Master节点，并执行**su - omm**命令切换到omm用户。

步骤2 执行如下命令登录高斯DB。

gsql -p 20051 -d hivemeta -U *username* -W *password*

步骤3 执行如下命令修改限制。

alter table PARTITION_PARAMS alter column PARAM_VALUE type varchar(6000);

----结束

16.10.42 增加 Hive 表字段超时

用户问题

增加Hive表字段报错。

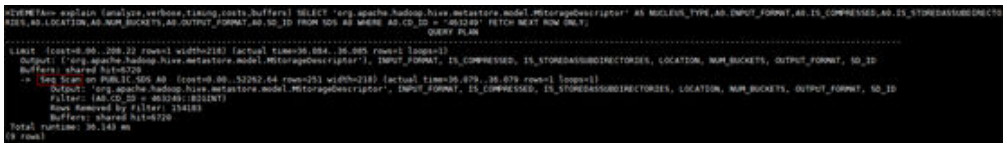
问题现象

Hive对包含10000+分区的表执行**ALTER TABLE table_name ADD COLUMNS(column_name string) CASCADE;**，报错如下：

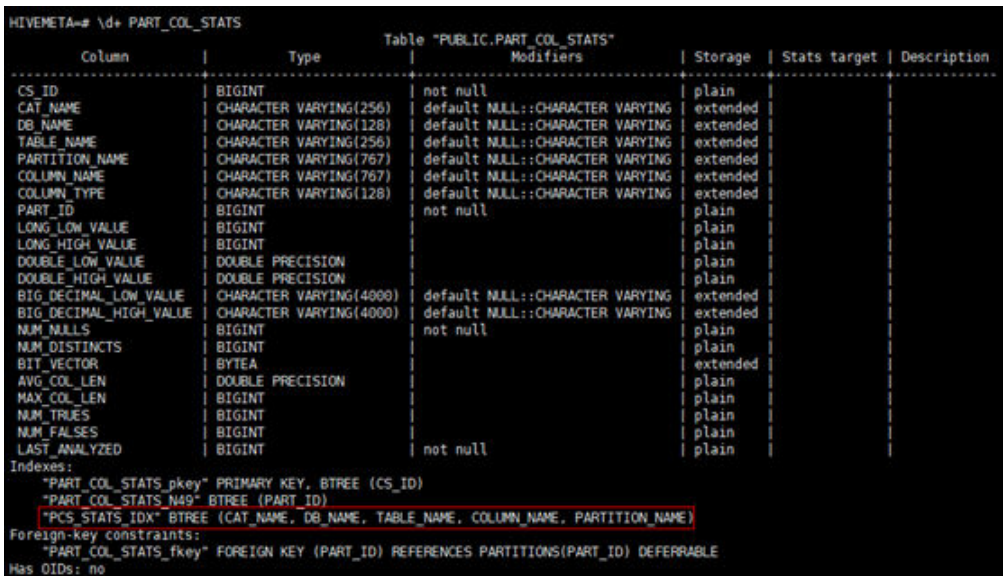
```
Timeout when executing method: alter_table_with_environment_context; 600525ms exceeds 600000ms
```

原因分析

1. MetaStore客户端连接超时，MRS默认MetaStore客户端和服务端连接的超时时间是600s，在Manager页面调大hive.metastore.client.socket.timeout为3600s。
2. 出现另一个报错：
   ```
   Error: org.apache.hive.service.cli.HiveSQLException: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. Unable to alter table. java.net.SocketTimeoutException: Read timed out
   ```
   Metastore元数据JDBC连接超时，默认60ms。
3. 调大javax.jdo.option.ConnectionURL中socketTimeout=60000，仍然产生最初的报错：
   ```
   Timeout when executing method: alter_table_with_environment_context;3600556ms exceeds 3600000ms
   ```
4. 尝试调大hive.metastore.batch.retrieve.max、hive.metastore.batch.retrieve.table.partition.max、dbservice.database.max.connections等参数均未能解决。
5. 怀疑是GaussDB的问题，因为增加字段会遍历每个分区执行getPartitionColumnStatistics和alterPartition。
6. 使用omm用户执行**gsql -p 20051 -U omm -W dbserverAdmin@123 -d hivemeta**登录hive元数据库。
7. 执行**select * from pg_locks;**没有发现锁等待。
8. 执行**select * from pg_stat_activity;**发现进程执行时间较长。
   ```
   SELECT 'org.apache.hadoop.hive.metastore.model.MPartitionColumnStatistics'AS NUCLEUS_TYPE,A0.AVG_COL_LEN,A0."COLUMN_NAME",A0.COLUMN_TYPE,A0.DB_NAME,A0.BIG_DECIMAL_HIGH_VALUE,A0.BIG_DECIMAL_LOW_VALUE,A0.DOUBLE_HIGH_VALUE,A0.DOUBLE_LOW_VALUE,A0.LAST_ANALYZED,A0.LONG_HIGH_VALUE,A0.LONG_LOW_VALUE,A0.MAX_COL_LEN,A0.NUM_DISTINCTS,A0.NUM_FALSES,A0.NUM_NULLS,A0.NUM_TRUES,A0.PARTITION_NAME,A0."TABLE_NAME",A0.CS_ID,A0.PARTITION_NAMEAS NUCORDER0 FROM PART_COL_STATS A0 WHERE A0."TABLE_NAME" = '$1' ANDA0.DB_NAME = '$2' AND A0.PARTITION_NAME = '$3' AND((((((A0."COLUMN_NAME" = '$4') OR (A0."COLUMN_NAME" ='$5')) OR (A0."COLUMN_NAME" = '$6')) OR (A0."COLUMN_NAME" ='$7')) OR (A0."COLUMN_NAME" = '$8')) OR (A0."COLUMN_NAME" ='$9')) ORDER BY NUCORDER0;
   ```
   
9. 执行**gs_guc reload -c log_min_duration_statement=100 -D /srv/BigData/dbdata_service/data/**开启sql录制，发现**8.执行select * from pg_sta...**中语句执行时长700ms，而且因为有10000+分区，会触发执行10000+次命令。
10. 在sql前加explain (analyze,verbose,timing,costs,buffers)分析执行计划，发现执行时需要全表扫描。



11. 查看索引，发现不满足最左匹配原则。



```
HIVEMETA=# \d+ PART_COL_STATS
                                    Table "PUBLIC.PART_COL_STATS"
        Column         |          Type          |           Modifiers           | Storage  | Stats target | Description
-----------------------+------------------------+-------------------------------+----------+--------------+-------------
 CS_ID                 | BIGINT                 | not null                      | plain    |              |
 CAT_NAME              | CHARACTER VARYING(256) | default NULL::CHARACTER VARYING | extended |              |
 DB_NAME               | CHARACTER VARYING(128) | default NULL::CHARACTER VARYING | extended |              |
 TABLE_NAME            | CHARACTER VARYING(256) | default NULL::CHARACTER VARYING | extended |              |
 PARTITION_NAME        | CHARACTER VARYING(767) | default NULL::CHARACTER VARYING | extended |              |
 COLUMN_NAME           | CHARACTER VARYING(767) | default NULL::CHARACTER VARYING | extended |              |
 COLUMN_TYPE           | CHARACTER VARYING(128) | default NULL::CHARACTER VARYING | extended |              |
 PART_ID               | BIGINT                 | not null                      | plain    |              |
 LONG_LOW_VALUE        | BIGINT                 |                               | plain    |              |
 LONG_HIGH_VALUE       | BIGINT                 |                               | plain    |              |
 DOUBLE_LOW_VALUE      | DOUBLE PRECISION       |                               | plain    |              |
 DOUBLE_HIGH_VALUE     | DOUBLE PRECISION       |                               | plain    |              |
 BIG_DECIMAL_LOW_VALUE | CHARACTER VARYING(4000) | default NULL::CHARACTER VARYING | extended |              |
 BIG_DECIMAL_HIGH_VALUE | CHARACTER VARYING(4000) | default NULL::CHARACTER VARYING | extended |              |
 NUM_NULLS             | BIGINT                 | not null                      | plain    |              |
 NUM_DISTINCTS         | BIGINT                 |                               | plain    |              |
 BIT_VECTOR            | BYTEA                  |                               | extended |              |
 AVG_COL_LEN           | DOUBLE PRECISION       |                               | plain    |              |
 MAX_COL_LEN           | BIGINT                 |                               | plain    |              |
 NUM_TRUES             | BIGINT                 |                               | plain    |              |
 NUM_FALSES            | BIGINT                 |                               | plain    |              |
 LAST_ANALYZED         | BIGINT                 | not null                      | plain    |              |
Indexes:
    "PART_COL_STATS_pkey" PRIMARY KEY, BTREE (CS_ID)
    "PART_COL_STATS_N49" BTREE (PART_ID)
    "PCS_STATS_IDX" BTREE (CAT_NAME, DB_NAME, TABLE_NAME, COLUMN_NAME, PARTITION_NAME)
Foreign-key constraints:
    "PART_COL_STATS_fkey" FOREIGN KEY (PART_ID) REFERENCES PARTITIONS(PART_ID) DEFERRABLE
Has OIDs: no
```

处理步骤

1. 重建索引。

```
su – omm
gsql -p 20051 -U omm -W dbserverAdmin@123 -d hivemeta
DROP INDEX PCS_STATS_IDX;
CREATE INDEX PCS_STATS_IDX ON PART_COL_STATS(DB_NAME, TABLE_NAME, COLUMN_NAME, PARTITION_NAME);
CREATE INDEX SDS_N50 ON SDS(CD_ID);
```

2. 重新查看执行计划，发现语句已经可以索引查询，且5ms执行完成（原来是700ms）。重新执行hive表字段增加，已经可以添加成功。

16.10.43 Hive 服务重启失败

用户问题

修改Hive服务配置后，保存配置失败，Manager页面Hive服务的配置状态为配置失败。

问题现象

用户A在MRS节点后台上打开了Hive相关配置文件且未关闭，此时用户B在MRS Manager页面的“服务管理”中修改Hive配置项，保存配置并重启Hive服务，此时保存配置失败，并且Hive服务启动失败。

原因分析

由于用户B在MRS Manager页面修改配置时，配置文件被用户A在MRS节点后台打开，导致该配置文件不能被替换，最终导致Hive服务启动失败。

处理步骤

步骤1 用户需要首先手动关闭集群节点后台打开的Hive配置文件。

步骤2 在MRS Manager页面重新修改Hive的配置并保存配置。

步骤3 重启Hive服务。

----结束

16.10.44 hive 执行删除表失败

用户问题

hive表删除失败

问题现象

hive创建的二级分区表有两万多个分区，导致用户在执行**truncate table ${TableName},drop table ${TableName}**时失败。

原因分析

删除文件操作是单线程串行执行的，hive分区数过多导致在元数据数据库会保存大量元数据信息，在执行删表语句时删除元数据就要用很长时间，最终在超时时间内删除不完，就会导致操作失败。

说明

超时时间可通过登录FusionInsight Manager，选择“集群 > 服务 > Hive > 配置 >全部配置 > MetaStore（角色） > 服务初始化”查看，“hive.metastore.client.socket.timeout”对应的值即为超时时间时长，在“描述”列可查看默认值。

处理步骤

步骤1 如果是内部表可以先通过**alter table ${TableName} set TBLPROPERTIES('EXTERNAL'='true')**来将内部表转成外部表，这样hive删除的时候只删除元数据省去了删除hdfs数据的时间。

步骤2 如果要用相同的表名可以先将表结构用**show create table ${TableName}**来导出表结构，再用**ALTER TABLE ${TableName} RENAME TO ${new_table_name};**来将表重命名。这样就可以新建一个和原来一样表。

步骤3 执行**hdfs dfs –rm –r –f ${hdfs_path}**在hdfs上删除表数据。

步骤4 在hive中用**alter table ${Table_Name} drop partition (${PartitionName}<’XXXX’, ${PartitionName}>’XXXX’);**删除分区(具体删除条件可灵活处理),减少文件数。

步骤5 删除分区少于一千个后，直接用**drop table ${TableName}**删掉表即可。

----结束

建议与总结

hive分区虽然可以提高查询效率,但要避免分区不合理导致出现大量小文件的问题,要提前规划好分区策略.

16.10.45 Hive 执行 msck repair table table_name 报错

现象描述

Hive执行msck repair table table_name报错：FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask (state=08S01,code=1)。

可能原因

查看HiveServer日志/var/log/Bigdata/hive/hiveserver/hive.log，发现目录名不符合分区格式。

```
2020-07-15 15:38:10,427 | WARN  | HiveServer2-Background-Pool: Thread-10905216 | Failed to run metacheck:  | org.apache.hadoop.hive.ql.exec.DDLTask.msck(DDLTask.java:2023)
org.apache.hadoop.hive.ql.metadata.HiveException: Repair: Cannot add partition ads_marketing.t_marketing_telemarketing_order_list:dtime=2020-04-24 17%3A55%3A00 due to invalid characters in the name
    at org.apache.hadoop.hive.ql.exec.DDLTask.msck(DDLTask.java:1986) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.exec.DDLTask.execute(DDLTask.java:424) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:199) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:100) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:2185) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:1841) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1527) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1238) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1233) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:266) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hive.service.cli.operation.SQLOperation.access$800(SQLOperation.java:93) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork$1.run(SQLOperation.java:379) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at java.security.AccessController.doPrivileged(Native Method) ~[?:1.8.0_232]
    at javax.security.auth.Subject.doAs(Subject.java:422) [?:1.8.0_232]
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1840) [hadoop-common-2.8.3-mrs-1.9.0.jar:?]
    at org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork.run(SQLOperation.java:393) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511) [?:1.8.0_232]
    at java.util.concurrent.FutureTask.run(FutureTask.java:266) [?:1.8.0_232]
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149) [?:1.8.0_232]
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624) [?:1.8.0_232]
    at java.lang.Thread.run(Thread.java:748) [?:1.8.0_232]
```

处理步骤

- 方法一：删除错误的文件或目录。
- 方法二：执行set hive.msck.path.validation=skip，跳过无效的目录。

16.10.46 在 Hive 中 drop 表后，如何完全释放磁盘空间

用户问题

在Hive命令行执行**drop**表的操作后，通过命令**hdfs dfsadmin -report**查看磁盘空间，发现表没有删除。

原因分析

在Hive命令行执行**drop**表只删除了外部表的表结构，并没有删除该表存储在HDFS上的表数据。

处理步骤

步骤1 使用**root**用户登录安装客户端的节点，并认证用户。

cd *客户端安装目录*

source bigdata_env

kinit *组件业务用户*（未开启Kerberos认证的集群跳过此操作）

步骤2 执行以下命令删除存储在HDFS上的表。

hadoop fs -rm hdfs://hacluster/*表所在的具体路径*

----结束

16.10.47 客户端执行 SQL 报错连接超时

现象描述

客户端执行SQL失败，报错：Timed out waiting for a free available connection。

可能原因

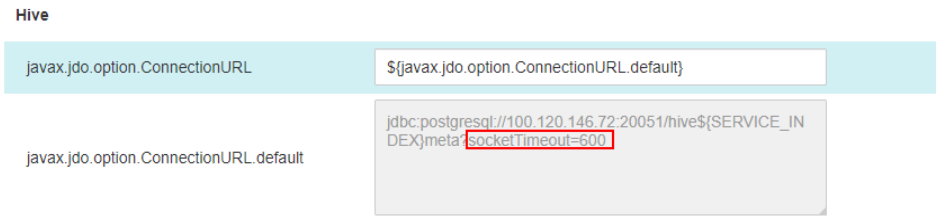
DBService连接较多，获取连接超时。

操作步骤

步骤1 客户端是否使用Spark-SQL客户端执行SQL。

- 是，检查连接的URL中超时参数，将其修改为600，执行**步骤7**。
- 否，执行**步骤2**。

步骤2 登录Manager页面，选择“集群 > 服务 > Hive > 配置 > 全部配置”，搜索“javax.jdo.option.ConnectionURL”，检查超时参数是否小于600。



说明

Hive、HiveServer、MetaStore、WebHCat中均有该参数，请确保它们的参数值一致。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 检查参数“javax.jdo.option.ConnectionURL”的值是否为“${javax.jdo.option.ConnectionURL.default}”。

- 是，执行**步骤4**。
- 否，修改URL中超时参数为600，单击“保存”，执行**步骤7**

步骤4 单击“实例”，选择任意HiveServer实例，并使用**root**用户登录实例节点。

步骤5 打开配置文件“${BIGDATA_HOME}/FusionInsight_Current/*HiveServer/etc/hivemetastore-site.xml”，查找配置项“javax.jdo.option.ConnectionURL”，复制配置项值。

```
<property>
<name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:postgresql://100.120.146.72:20051/hivemeta?socketTimeout=60</value>
</property>
<property>
```

步骤6 登录Manager页面，选择“集群 > 服务 > Hive > 配置 > 全部配置”，搜索“javax.jdo.option.ConnectionURL”，修改配置为**步骤5**中复制的URL，并修改超时参数为600，单击“保存”。

说明

Hive、HiveServer、MetaStore、WebHCat中均有该参数，请确保它们的参数值一致。

步骤7 选择“集群 > 服务 > Hive > 配置 > 全部配置”，搜索“maxConnectionsPerPartition”，检查是否小于100。

- 是，修改参数为100，单击“保存”，执行**步骤8**。
- 否，执行**步骤8**。

步骤8 若以上步骤有修改参数，选择“集群 > 服务 > Hive > 概览”，选择“更多 > 滚动重启服务”，未修改则无需执行此步骤。

----结束

16.10.48 WebHCat 健康状态异常导致启动失败

用户问题

WebHCat实例启动失败。

问题现象

在Manager页面上查看到WebHCat实例的健康状态为“故障”，并上报“ALM-12007 进程故障”告警，该告警的服务名称为“Hive”，实例名称为“WebHCat”。且重启Hive服务报错。

查看WebCat实例的日志“/var/log/Bigdata/hive/webhcat/webhcat.log”报错“Service not found in Kerberos database”和“Address already in use”。

处理步骤

步骤1 依次登录WebHCat实例所在节点检查“/etc/hosts”文件中的IP及主机名称映射关系是否正确。且“/etc/hostname”和“/etc/HOSTNAME”文件的WebHCat配置需与“/etc/hosts”保持一致，若不一致则需手动修改。

说明

WebHCat实例的IP地址及主机名称映射关系可登录FusionInsight Manager界面，选择“集群 > 服务 > Hive > 实例”查看。

步骤2 登录WebHCat实例所在节的任一节点，执行以下命令切换到**omm**用户。

su - omm

步骤3 执行以下命令查看是否存在WebHCat进程。

ps -ef|grep webhcat|grep -v grep

若存在，则需执行以下命令结束WebHCat进程：

kill -9 ${webhcat_pid}

步骤4 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 实例”，勾选所有WebHCat实例，选择“更多 > 重启实例”，等待WebHCat重启成功即可。

----结束

16.10.49 mapred-default.xml 文件解析异常导致 WebHCat 启动失败

用户问题

MRS的Hive服务故障，重新启动后，Master2节点上的HiveServer和WebHCat进程启动失败，Master1节点进程正常。

原因分析

登录Master2节点，查看“/var/log/Bigdata/hive/hiveserver/hive.log”日志，发现HiveServer一直加载“/opt/Bigdata/*/*_HiveServer/etc/hive-site.xml”；查看HiveServer退出时的“/var/log/Bigdata/hive/hiveserver/hiveserver.out”日志，发现解析“mapred-default.xml”文件异常。

处理步骤

步骤1 登录Master2节点，使用以下命令查找“mapred-default.xml”所在路径：

find /opt/ -name 'mapred-default.xml'

查询到该配置文件在“/opt/Bigdata/*/*_WebHCat/etc/”目录下面，且该文件内容为空。

步骤2 登录到Master1节点，将“/opt/Bigdata/*/*_WebHCat/etc/mapred-default.xml”文件拷贝到Master2节点，并修改文件的属组为“omm:wheel”。

步骤3 登录Manager，重启异常的HiveServer和WebHCat实例。

----结束

16.11 使用 Hue

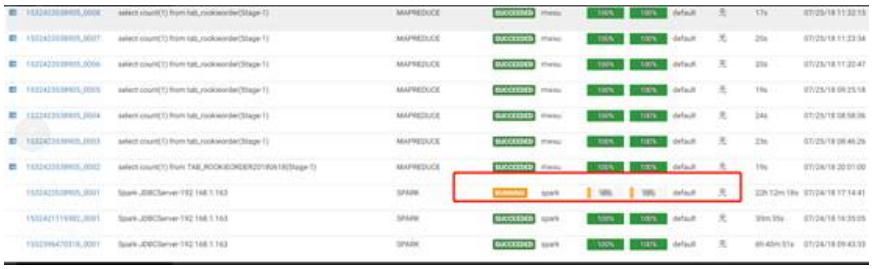
16.11.1 Hue 上有 job 在运行

用户问题

客户查到Hue上有job在运行。

问题现象

客户的MRS装好后， Hue上查到有Job在运行 ，并且目前在运行的job并不是客户操作的。



原因分析

此Job为Spark启动之后，系统自身连接jdbc的一个任务，是常驻的。

处理步骤

非问题，无需处理。

16.11.2 使用 IE 浏览器在 Hue 中执行 HQL 失败

问题背景与现象

使用IE浏览器在Hue中访问Hive Editor并执行所有HQL失败，界面提示“There was an error with your query.”。

原因分析

IE浏览器存在功能问题，不支持在307重定向中处理含有form data的AJAX POST请求，建议更换兼容的浏览器。

解决办法

使用Google Chrome浏览器21及以上版本。

16.11.3 Hue（主）无法打开 web 网页

问题背景与现象

访问Hue（主）的WebUI界面提示如下：

Service Unavailable
The server is temporarily unable to service your request due to maintenance downtime or capacity problems. Please try again later.

原因分析

- Hue配置过期。
- 单Master节点集群中，Hue服务需要手动修改配置。

解决办法

- Hue配置过期，重启Hue服务即可。
- 单Master节点的集群Hue服务需要手动修改配置。

a. 登录Master节点。
b. 执行**hostname -i**获取本机IP。
c. 执行如下命令获取“HUE_FLOAT_IP”的地址：

   grep "HUE_FLOAT_IP" ${BIGDATA_HOME}/*MRS*_Current/1_*/etc*/ENV_VARS，

   其中*MRS*以实际文件名为准。
d. 比较本机IP和“HUE_FLOAT_IP”的值是否相同，若不相同，请修改“HUE_FLOAT_IP”的值为本机IP。
e. 重启Hue服务。

16.11.4 Hue WebUI 访问失败

用户问题

访问Hue WebUI跳转到错误的页面。

问题现象

查看hue web ui报错如下：

```
503 Service Unavailable
The server is temporarily unable to service your requster due to maintenance downtime or capacity problems.Please try again later.
```

原因分析

- hue配置过期。
- 单Master节点集群中，Hue服务需要手动修改配置。

处理步骤

步骤1 登录Master节点。

步骤2 执行**hostname -i**获取本机IP。

步骤3 执行如下命令获取“HUE_FLOAT_IP”的地址：

grep "HUE_FLOAT_IP" ${BIGDATA_HOME}/*MRS*_Current/1_*/etc*/ENV_VARS,其中*MRS*以实际文件名为准。

步骤4 比较本机IP和“HUE_FLOAT_IP”的值是否相同，若不相同，请修改“HUE_FLOAT_IP”的值为本机IP。

步骤5 重启Hue服务。

----结束

16.11.5 Hue 界面无法加载 HBase 表

用户问题

用户在Hue界面将hive数据导入hbase后，报检测不到hbase表的错误。

问题现象

Kerberos集群中，IAM子账户权限不足导致无法加载hbase表。

原因分析

IAM子账户权限不足。

处理步骤

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统管理 > 用户管理”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加supergroup组。

步骤5 单击“确定”完成修改操作。

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行单击“修改”。

步骤4 为用户添加supergroup组。

步骤5 单击“确定”完成修改操作。

----结束

建议与总结

如果是开启Kerberos认证的集群，页面出现 No data available优先排查权限问题。

16.12 使用 Impala

16.12.1 用户连接 impala-shell 失败

用户问题

用户连接impala-shell失败。

问题现象

用户在“组件管理”页面修改任意组件的配置并重启服务后，连接impala-shell，会出现连接失败，报错no such file/directory。

```
[root@node-master1emdj etc]# pwd
/opt/Bigdata/MRS_2.1.0/1_7_KuduMaster/etc
[root@node-master1emdj etc]# impala-shell -i 192.168.0.73
shell-init: error retrieving current directory: getcwd: cannot access parent directories: No such file or directory
chdir: error retrieving current directory: getcwd: cannot access parent directories: No such file or directory
Traceback (most recent call last):
  File "/opt/client/Impala/impala/shell/impala_shell.py", line 38, in <module>
    from impala_client import (ImpalaClient, DisconnectedException, QueryStateException,
  File "/opt/client/Impala/impala/shell/lib/impala_client.py", line 20, in <module>
    import sasl
  File "build/bdist.linux-x86_64/egg/sasl/__init__.py", line 1, in <module>

  File "build/bdist.linux-x86_64/egg/sasl/saslwrapper.py", line 7, in <module>
  File "build/bdist.linux-x86_64/egg/_saslwrapper.py", line 7, in <module>
  File "build/bdist.linux-x86_64/egg/_saslwrapper.py", line 3, in __bootstrap__
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 2594, in <module>
    for comparator, version in req.specs:
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 425, in __init__

  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 440, in add_entry
    `req`.  But, if there is an active distribution for the project and it
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 1688, in find_on_path
    return ()
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 1835, in _normalize_cached

  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 1829, in normalize_path
    register_namespace_handler(object,null_ns_handler)
  File "/usr/lib64/python2.7/posixpath.py", line 368, in realpath
    return abspath(path)
  File "/usr/lib64/python2.7/posixpath.py", line 356, in abspath
    cwd = os.getcwd()
OSError: [Errno 2] No such file or directory
```

原因分析

修改服务配置并重启服务后，部分服务的目录结构会删除并重新创建，如服务的etc目录等。如果重启服务前所在的目录为etc或者其子目录，由于重启后目录重建，仍在原来目录执行impala-shell时会产生某些系统变量或者参数无法找到的情况，所以连接impala-shell连接失败。

处理步骤

任意切换到存在的目录，重新连接impala-shell即可。

16.12.2 创建 Kudu 表报错

用户问题

创建Kudu表报错。

问题现象

新建了集群，在创建表时，报错“[Cloudera]ImpalaJDBCDriver ERROR processing query/statement. Error Code: 0, SQL state: TStatus(statusCode:ERROR_STATUS, sqlState:HY000, errorMessage:AnalysisException: Table property 'kudu.master_addresses' is required when the impalad startup flag -kudu_master_hosts is not used.”

原因分析

客户未在impala sql中指定kudu.master_addresses地址导致报错：Table property 'kudu.master_addresses' is required when the impalad startup flag -kudu_master_hosts is not used.

处理步骤

在创建Kudu表时指定“kudu.master_addresses”地址。

16.12.3 Impala 客户端登录失败

用户问题

运行Impala client会报类似如下错误信息：

```
[root@node-master1avIy ~]# impala-shell -i 192.168.128.49:21000
  File "/opt/client/Impala/impala/shell/impala_shell.py", line 1675
    except Exception, e:
                    ^
SyntaxError: invalid syntax
[root@node-master1avIy ~]#
```

原因分析

由于最新的MRS集群使用的是Euler2.9及以上版本的操作系统，系统自带只python3版本，而Impala client是基于python2实现的，和python3部分语法不兼容，运行Impala client会报错误信息，所以需要手动安装python2以解决Impala client运行问题。

处理步骤

步骤1 使用root用户登录Impala所在节点，执行如下命令，确认当前系统上安装的python版本：

python --version

```
[root@node-master2JgOY ~]# python --version
Python 3.7.4
```

步骤2 执行命令**yum install make**，查看yum是否可用。

- 如果yum install报如下错误，说明yum设置有问题，执行**步骤3**。

  ```
  [root@node-master2JgOY ~]# yum install make
  Error: There are no enabled repositories in "/etc/yum.repos.d", "/etc/yum/repos.d", "/etc/distro.repos.d".
  ```

- 如果没有报错，执行**步骤4**。

步骤3 执行命令**cat /etc/yum.repos.d/EulerOS-base.repo**，查看yum源和系统版本信息不匹配是否匹配，如果不匹配则修改，如下所示：

修改前：

```
[root@node-master1avIy ~]# cat /etc/yum.repos.d/EulerOS-base.repo
[base]
name=EulerOS-2.0SP2 base
baseurl=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.2/os/x86_64/
enabled=1
gpgcheck=1
gpgkey=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.2/os/RPM-GPG-KEY-EulerOS
[root@node-master1avIy ~]# uname -a
Linux node-master1avIy.mrs-mq7v.com 4.18.0-147.5.1.6.h541.eulerosv2r9.x86_64 #1 SMP Wed Aug 4 02:30:13 UTC
 x86_64 GNU/Linux
```

修改后：

```
[base]
name=EulerOS-2.0SP9 base
baseurl=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.9/os/x86_64/
enabled=1
gpgcheck=1
gpgkey=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.9/os/RPM-GPG-KEY-EulerOS
```

步骤4 执行如下命令，查看yum源上python2开头的软件。

yum list python2*

```
[root@node-master2JgOY ~]# yum list python2*
Last metadata expiration check: 0:02:36 ago on Thu 16 Dec 2021 10:05:52 AM CST.
Available Packages
python2.x86_64                                          2.7.16-16.eulerosv2r9
python2-debug.x86_64                                    2.7.16-16.eulerosv2r9
python2-devel.x86_64                                    2.7.16-16.eulerosv2r9
python2-help.noarch                                     2.7.16-16.eulerosv2r9
python2-pip.noarch                                      18.0-13.h2.eulerosv2r9
python2-setuptools.noarch                               40.4.3-4.h1.eulerosv2r9
python2-tkinter.x86_64                                  2.7.16-16.eulerosv2r9
python2-tools.x86_64                                    2.7.16-16.eulerosv2r9
```

步骤5 执行如下命令，安装python2。

yum install python2

```
[root@node-master2JgOY ~]# yum install python2
Last metadata expiration check: 0:00:48 ago on Thu 16 Dec 2021 10:05:52 AM CST.
Error:
 Problem: problem with installed package python3-unversioned-command-3.7.4-7.h29.eulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h29.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h11.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h13.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h15.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h18.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h33.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h38.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - conflicting requests
(try to add '--allowerasing' to command line to replace conflicting packages or '--skip-broken' to skip uninstallable package
s or '--nobest' to use not only best candidate packages)
```

因为当前系统上已安装python3，所有直接安装python2会有上面的冲突提示。

可以选择--allowerasing或--skip-broken安装，例如：

yum install python2 --skip-broken

```
[root@node-master2JgOY ~]# yum install python2 --skip-broken
Last metadata expiration check: 0:34:08 ago on Thu 16 Dec 2021 10:05:52 AM CST.
Dependencies resolved.
================================================================================================================================
 Package                         Architecture         Version                                 Repository          Size
================================================================================================================================
Installing:
 python2                         x86_64               2.7.16-16.eulerosv2r9                   base               6.4 M
Installing dependencies:
 libXft                          x86_64               2.3.2-13.eulerosv2r9                    base                41 k
```

安装完成后，会自动将python版本修改为python2，如下所示：

```
Installed:
  libXft-2.3.2-13.eulerosv2r9.x86_64                        libXrender-0.9.10-10.eulerosv2r9.x86_64
  python2-2.7.16-16.eulerosv2r9.x86_64                      python2-debug-2.7.16-16.eulerosv2r9.x86_64
  python2-devel-2.7.16-16.eulerosv2r9.x86_64                python2-help-2.7.16-16.eulerosv2r9.noarch
  python2-setuptools-40.4.3-4.h1.eulerosv2r9.noarch         python2-tkinter-2.7.16-16.eulerosv2r9.x86_64
  python2-tools-2.7.16-16.eulerosv2r9.x86_64                python3-rpm-generators-9-1.eulerosv2r9.noarch
  tk-1:8.6.8-5.eulerosv2r9.x86_64

Complete!
[root@node-master2JgOY ~]# python --version
Python 2.7.16
```

如果python2安装成功，但是显示的python版本不对，需要执行以下命令手动给“/usr/bin/python2”创建软链接“/usr/bin/python”。

ln -s /usr/bin/python2 /usr/bin/python

步骤6 验证Impala client是否可用。

```
[root@node-master1avIy ~]# impala-shell -i 192.168.128.49:21000
Starting Impala Shell without Kerberos authentication
Opened TCP connection to 192.168.128.49:21000
Connected to 192.168.128.49:21000
Server version: impalad version 3.4.0-RELEASE RELEASE (build eebadd34c1563cbf5825a4e4d361e7b3601f9827)
***********************************************************************************
Welcome to the Impala shell.
(Impala Shell v3.4.0-RELEASE (eebadd3) built on Thu Nov  4 11:29:54 CST 2021)

After running a query, type SUMMARY to see a summary of where time was spent.
***********************************************************************************
[192.168.128.49:21000] default> show databases;
Query: show databases
+------------------+----------------------------------------------+
| name             | comment                                      |
+------------------+----------------------------------------------+
| _impala_builtins | System database for Impala builtin functions |
| default          | Default Hive database                        |
+------------------+----------------------------------------------+
Fetched 2 row(s) in 0.16s
[192.168.128.49:21000] default>
```

----结束

16.13 使用 Kafka

16.13.1 运行 Kafka 获取 topic 报错

用户问题

客户运行Kafka获取topic报错。

问题现象

运行Kafka获取topic时报错，报错内容如下：

```
ERROR org.apache.kafka.common.errors.InvalidReplicationFactorException: Replication factor: 2 larger than available brokers: 0.
```

原因分析

由特殊字符导致获取zookeeper地址的变量错误。

处理步骤

步骤1 登录任意一个Master节点。

步骤2 执行**cat /opt/client/Kafka/kafka/config/server.properties |grep '^zookeeper.connect ='**命令，查看zookeeper地址的变量。

步骤3 重新运行Kafka获取topic，其中从**步骤2**中获取的变量不要添加任何字符。

----结束

16.13.2 Flume 可以正常连接 Kafka，但是发送消息失败。

问题现象

使用MRS版本安装集群，主要安装ZooKeeper、Flume、Kafka。

在使用Flume向Kafka发送数据功能时，发现Flume发送数据到Kafka失败。

可能原因

1. Kafka服务异常。
2. Flume连接Kafka地址错误，导致发送失败。
3. Flume发送超过Kafka大小限制的消息，导致发送失败。

原因分析

Flume发送数据到Kafka失败，可能原因是Flume侧问题或者Kafka侧问题。

1. Manager界面查看当前Kafka状态及监控指标。
   - MRS Manager界面操作：登录MRS Manager，选择 “服务管理 > Kafka ”，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Flume日志，发现打印MessageSizeTooLargeException异常信息，如下所示：

   ```
   2016-02-26 14:55:19,126 | WARN  | [SinkRunner-PollingRunner-DefaultSinkProcessor] |  Produce request with correlation id 349829 failed due to [LOG,7]: kafka.common.MessageSizeTooLargeException  | kafka.utils.Logging$class.warn(Logging.scala:83)
   ```

   通过异常信息，发现当前Flume向Kafka写入的数据超过了Kafka服务端定义的消息的最大值。
3. 通过Manager查看Kafka服务端定义的消息的最大值。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

   进入Kafka配置页面，参数类别选择全部配置，显示所有Kafka相关配置，在“搜索”中输入message.max.bytes进行检索。

   MRS中Kafka服务端默认可以接收的消息最大为1000012 bytes =977KB。

解决办法

与用户确认，当前Flume发送数据确实存在超过1M的消息。因此，为了确保当前这些消息能够写入Kafka，需要调整Kafka服务端相关参数。

步骤1 修改message.max.bytes，使得message.max.bytes的值大于当前业务中消息最大值，使得Kafka服务端可以接收全部消息。

步骤2 修改replica.fetch.max.bytes，使得**replica.fetch.max.bytes >= message.max.bytes**，使得不同Broker上的Partition的Replica可以同步到全部消息。

- MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
- FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

进入Kafka配置页面，参数类别选择全部配置， 显示所有Kafka相关配置，在“搜索”中输入replica.fetch.max.bytes进行检索。

步骤3 单击“保存”，并重启Kafka服务，使得Kafka相关配置生效。

步骤4 修改消费者业务应用中fetch.message.max.bytes，使得fetch.message.max.bytes >= message.max.bytes，确保消费者可以消费到全部消息。

----结束

16.13.3 Producer 发送数据失败，抛出 NullPointerException

问题现象

使用MRS安装集群，主要安装ZooKeeper、Kafka。

在使用Producer向Kafka发送数据功能时，发现客户端抛出NullPointerException。

可能原因

1. Kafka服务异常。
2. 客户端Producer侧配置jaas和keytab文件错误。

原因分析

Producer发送数据到Kafka失败，可能原因客户端Producer侧问题或者Kafka侧问题。

1. 通过Manager页面查看kafka服务状态及监控指标。
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

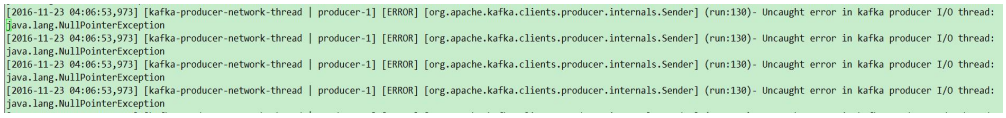
     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Producer客户端日志，发现打印NullPointerException异常信息，如图16-42所示。

图 16-42 Producer 客户端日志

```
[2016-12-06 02:04:05,906]-[schedule-C50D0717-4643-4D4E-9D6E-B940E4BD7159]-[kafka-producer-network-thread |
SZX1000161910-10.21.219.222-bigdata-producer-5]-[1005]-[org.apache.kafka.clients.producer.internals.Sender.run
thread:
java.lang.NullPointerException
    at org.apache.kafka.common.network.Selector.pollSelectionKeys(Selector.java:302)
    at org.apache.kafka.common.network.Selector.poll(Selector.java:283)
    at org.apache.kafka.clients.NetworkClient.poll(NetworkClient.java:260)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:229)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:134)
    at java.lang.Thread.run(Thread.java:745)
[2016-12-06 02:04:05,921]-[schedule-C50D0717-4643-4D4E-9D6E-B940E4BD7159]-[kafka-producer-network-thread |
SZX1000161910-10.21.219.222-bigdata-producer-3]-[1005]-[org.apache.kafka.clients.producer.internals.Sender.run
thread:
java.lang.NullPointerException
    at org.apache.kafka.common.network.Selector.pollSelectionKeys(Selector.java:302)
    at org.apache.kafka.common.network.Selector.poll(Selector.java:283)
    at org.apache.kafka.clients.NetworkClient.poll(NetworkClient.java:260)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:229)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:134)
    at java.lang.Thread.run(Thread.java:745)
```

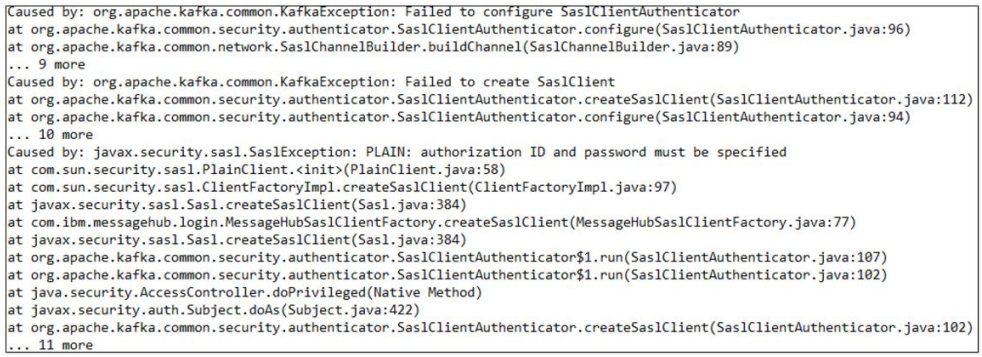
或者日志中只有异常信息没有堆栈信息（只有NullPointerException无堆栈信息，出现这个问题是jdk的自我保护，相同堆栈打印太多，就会触发这个保护开关，后续不再打印堆栈），如图16-43所示。

图 16-43 异常信息



3. 查看Producer客户端日志，发现打印Failed to configure SaslClientAuthenticator异常信息，如图16-44所示。

图 16-44 异常日志信息



4. 认证失败导致创建KafkaChannel失败，导致通过channel(key)方法获取的KafkaChannel为空，以至于疯狂打印NullPointerException，上述日志可以发现，认证失败的原因是用户密码不正确，密码不正确的原因可能是用户名不匹配导致。
5. 检查Jaas文件和Keytab文件，发现Jaas文件中配置使用的pricipal为stream。

图 16-45 检查 Jaas 文件

```
KafkaClient {
com.sun.security.auth.module.Krb5LoginModule required
debug=false
keyTab="/opt/client/user.keytab"
useTicketCache=false
storeKey=true
principal="stream@HADOOP.COM"
useKeyTab=true;
};
```

查看user.keytab文件，发现principal为zmk_kafka。

图 16-46 查看 user.keytab 文件

```
[root@8-5-148-6 client]# klist -kt user.keytab
Keytab name: FILE:user.keytab
KVNO Timestamp         Principal
---- ----------------- --------------------------------------------------------
   1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
   1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
```

发现jaas文件和user.keytab文件中principal不对应。

该情况是由于应用程序自动定时更新Jaas文件，但是有两个不同的进程在进行更新，一个进程写入正确的Principal而另一个却写入了错误的Principal，以至于程序时而正常，时而异常。

解决办法

步骤1 修改Jaas文件，确保使用的Principal在Keytab文件中存在。

----结束

16.13.4 Producer 发送数据失败，抛出 TOPIC_AUTHORIZATION_FAILED

问题现象

使用MRS安装集群，主要安装ZooKeeper、Kafka。

在使用Producer向Kafka发送数据功能时，发现客户端抛出TOPIC_AUTHORIZATION_FAILED。

可能原因

1. Kafka服务异常。
2. 客户端Producer侧采用非安全访问，服务端配置禁止访问。
3. 客户端Producer侧采用非安全访问，Kafka Topic设置ACL。

原因分析

Producer发送数据到Kafka失败，可能原因客户端Producer侧问题或者Kafka侧问题。

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Producer客户端日志，发现打印TOPIC_AUTHORIZATION_FAILED异常信息。

   ```
   [root@10-10-144-2 client]# kafka-console-producer.sh --broker-list 10.5.144.2:9092 --topic test
   1
   [2017-01-24 16:58:36,671] WARN Error while fetching metadata with correlation id 0 :
   {test=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
   [2017-01-24 16:58:36,672] ERROR Error when sending message to topic test with key: null, value: 1
   bytes with error: Not authorized to access topics: [test]
   (org.apache.kafka.clients.producer.internals.ErrorLoggingCallback)
   ```

   Producer采用9092端口来访问Kafka，9092为非安全端口。
3. 通过Manager页面，查看当前Kafka集群配置，发现未设置自定义配置“allow.everyone.if.no.acl.found”=“false”。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

4. 当acl设置为false不允许采用9092来进行访问。
5. 查看Producer客户端日志，发现打印TOPIC_AUTHORIZATION_FAILED异常信息。

```
[root@10-10-144-2 client]# kafka-console-producer.sh --broker-list 10.5.144.2:21005 --topic test_acl
1
[2017-01-25 11:09:40,012] WARN Error while fetching metadata with correlation id 0 :
{test_acl=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
[2017-01-25 11:09:40,013] ERROR Error when sending message to topic test_acl with key: null, value:
1 bytes with error: Not authorized to access topics: [test_acl]
(org.apache.kafka.clients.producer.internals.ErrorLoggingCallback)
[2017-01-25 11:14:40,010] WARN Error while fetching metadata with correlation id 1 :
{test_acl=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
```

   Producer采用21005端口来访问Kafka，21005为非安全端口。
6. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:24002/
kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
User:test_user has Allow permission for operations: Write from hosts: *
```

   Topic设置acl，则不允许采用9092来访问。
7. 查看Producer客户端日志，发现打印TOPIC_AUTHORIZATION_FAILED异常信息。

```
[root@10-10-144-2 client]# kafka-console-producer.sh --broker-list 10.5.144.2:21007 --topic topic_acl
--producer.config /opt/client/Kafka/kafka/config/producer.properties
1
[2017-01-25 12:43:58,506] WARN Error while fetching metadata with correlation id 0 :
{topic_acl=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
[2017-01-25 12:43:58,507] ERROR Error when sending message to topic topic_acl with key: null,
value: 1 bytes with error: Not authorized to access topics: [topic_acl]
(org.apache.kafka.clients.producer.internals.ErrorLoggingCallback)
```

   Producer采用21007端口来访问Kafka。
8. 通过客户端命令klist查询当前认证用户。

```
[root@10-10-144-2 client]# klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: test@HADOOP.COM

Valid starting     Expires            Service principal
01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
```

   当前认证用户为**test**。
9. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/
kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
User:test_user has Allow permission for operations: Write from hosts: *
```

   Topic设置acl，用户test_user具有producer权限。test无权限进行producer操作。

   解决方法参考步骤2。
10. 通过SSH登录Kafka Broker：

    通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

    查看kafka-authorizer.log发现如下日志提示用户不属于kafka或者kafkaadmin组。

```
2017-01-25 13:26:33,648 | INFO  | [kafka-request-handler-0] | The principal is test, belongs to Group :
[hadoop, ficommon] | kafka.authorizer.logger (SimpleAclAuthorizer.scala:169)
2017-01-25 13:26:33,648 | WARN  | [kafka-request-handler-0] | The user is not belongs to kafka or
kafkaadmin group, authorize failed! | kafka.authorizer.logger (SimpleAclAuthorizer.scala:170)
```

    解决方法参考步骤3。

解决办法

步骤1 配置自定义配置“allow.everyone.if.no.acl.found”参数为“true”，重启Kafka服务。

步骤2 采用具有权限用户登录。

例如：

kinit test_user

或者赋予用户相关权限。

须知

需要使用Kafka管理员用户（属于kafkaadmin组）操作。

例如：

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --topic topic_acl --producer --add --allow-principal User:test

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=8.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
 User:test_user has Allow permission for operations: Write from hosts: *
  User:test has Allow permission for operations: Describe from hosts: *
  User:test has Allow permission for operations: Write from hosts: *
```

步骤3 用户加入**Kafka**组或者**Kafkaadmin**组。

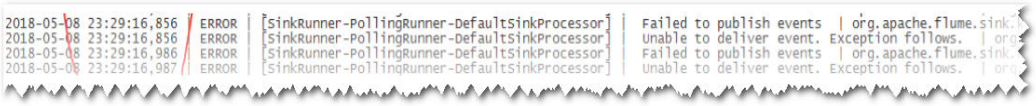
----结束

16.13.5 Producer 偶现发送数据失败，日志提示 Too many open files in system

问题背景与现象

在使用Producer向Kafka发送数据功能时，发现客户端发送失败。

图 16-47 Producer 发送数据失败



可能原因

1. Kafka服务异常。
2. 网络异常。
3. Kafka Topic异常。

原因分析

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

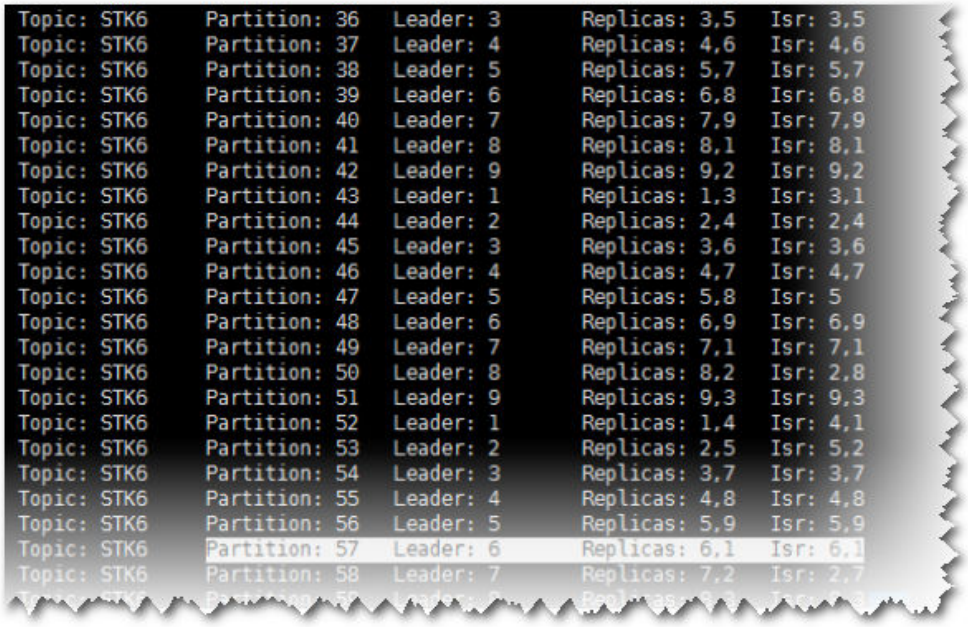
     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看SparkStreaming日志中提示错误的Topic信息。

   执行Kafka相关命令，获取Topic分布信息和副本同步信息，观察返回结果。

   kafka-topics.sh --describe --zookeeper <zk_host:port/chroot>

   如图16-48所示，发现对应Topic状态正常。所有Partition均存在正常Leader信息。

   图 16-48 Topic 状态

3. 通过telnet命令，查看是否可以连接Kafka。

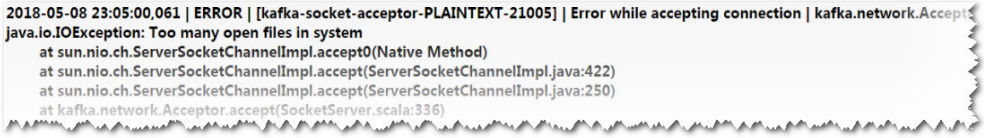
   telnet kafka业务ip kafka业务port

   如果无法telnet成功，请检查网络安全组与ACL。
4. 通过SSH登录Kafka Broker。

   通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

   查看server.log发现如下日志抛出java.io.IOException: Too many open files in system。

   图 16-49 日志异常

5. 通过lsof命令查看当前节点Kafka进程句柄使用情况，发现占用的句柄数达到了47万。

   图 16-50 句柄数

6. 排查业务代码，不停地new新的producer对象，未正常关闭。

解决办法

步骤1 停止当前应用，保证服务端句柄不再疯狂增加影响服务正常运行。

步骤2 优化应用代码，解决句柄泄露问题。

建议：全局尽量使用一个Producer对象，在使用完成之后主动调用close接口进行句柄关闭。

----结束

16.13.6 Consumer 初始化成功，但是无法从 Kafka 中获取指定 Topic 消息

问题背景与现象

使用MRS安装集群，主要安装ZooKeeper、Flume、Kafka、Storm、Spark。

使用Storm、Spark、Flume或者自己编写consumer代码来消费Kafka中指定Topic的消息时，发现消费不到任何数据。

可能原因

1. Kafka服务异常。
2. Consumer中ZooKeeper相关连接地址配置错误，导致无法消费。
3. Consumer发生ConsumerRebalanceFailedException异常，导致无法消费。
4. Consumer发生网络导致的ClosedChannelException异常，导致无法消费。

原因分析

Storm、Spark、Flume或者自定义Consumer代码可以都称为Consumer。

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。

2. 通过Kafka Client，判断是否可以正常消费数据。

   假设客户端已经安装在/opt/client目录，test为需要消费的Topic名称，192.168.234.231为ZooKeeper的IP地址。

   ```
   cd /opt/client
   source bigdata_env
   kinit admin
   kafka-topics.sh --zookeeper 192.168.234.231:2181/kafka --describe --topic testkafka-console-consumer.sh --topic test --zookeeper 192.168.234.231:2181/kafka --from-beginning
   ```

   当可以消费到数据时，表示集群服务正常。

3. 查看Consumer相关配置，发现ZooKeeper连接地址错误。

   - Flume

     ```
     server.sources.Source02.type=org.apache.flume.source.kafka.KafkaSource

     server.sources.Source02.zookeeperConnect=192.168.234.231:2181
     server.sources.Source02.topic = test
     server.sources.Source02.groupId = test_01
     ```

   - Spark

     ```
     val zkQuorum = "192.168.234.231:2181"
     ```

   - Storm

     ```
     BrokerHosts brokerHosts = new ZKHosts("192.168.234.231:2181");
     ```

   - Consumer API

     ```
     zookeeper.connect="192.168.234.231:2181"
     ```

   MRS中Kafka在ZooKeeper存储的ZNode是以/kafka为根路径，有别于开源。Kafka对应的ZooKeeper连接配置为192.168.234.231:2181/kafka。

   Consumer中配置为ZooKeeper连接配置为192.168.234.231:2181，导致无法正确获取Kafka中Topic相关信息。

   解决方法参考**步骤1**。

4. 查看Consumer相关日志，发现打印ConsumerRebalanceFailedException异常。

   ```
   2016-02-03 15:55:32,557 | ERROR | [ZkClient-EventThread-75- 192.168.234.231:2181/kafka] | Error handling event ZkEvent[New session event sent to kafka.consumer.ZookeeperConsumerConnector$ZKSessionExpireListener@34b41dfe] | org.I0Itec.zkclient.ZkEventThread.run(ZkEventThread.java:77)
   kafka.common.ConsumerRebalanceFailedException: pc-zjqbetl86-1454482884879-2ec95ed3 can't rebalance after 4 retries
   at kafka.consumer.ZookeeperConsumerConnector$ZKRebalancerListener.syncedRebalance(ZookeeperConsumerConnector.scala:633)
   at kafka.consumer.ZookeeperConsumerConnector$ZKSessionExpireListener.handleNewSession(ZookeeperConsumerConnector.scala:487)
   at org.I0Itec.zkclient.ZkClient$4.run(ZkClient.java:472)
   at org.I0Itec.zkclient.ZkEventThread.run(ZkEventThread.java:71)
   ```

   通过异常信息，发现当前Consumer没有在指定的重试次数内完成Rebalance，使得Consumer没有被分配Kafka Topic-Partition，则无法消费消息。

   解决方法参考**步骤3**。

5. 查看Consumer相关日志，发现打印Fetching topic metadata with correlation id 0 for topics [Set(test)] from broker [id:26,host:192-168-234-231,port:9092] failed错误信息和ClosedChannelException异常。

   ```
   [2016-03-04 03:33:53,047] INFO Fetching metadata from broker id:26,host:192-168-234-231,port:9092 with correlation id 0 for 1 topic(s) Set(test) (kafka.client.ClientUtils$)
   [2016-03-04 03:33:55,614] INFO Connected to 192-168-234-231:21005 for producing (kafka.producer.SyncProducer)
   [2016-03-04 03:33:55,614] INFO Disconnecting from 192-168-234-231:21005 (kafka.producer.SyncProducer)
   [2016-03-04 03:33:55,615] WARN Fetching topic metadata with correlation id 0 for topics [Set(test)] from broker [id:26,host: 192-168-234-231,port:21005] failed (kafka.client.ClientUtils$)
   java.nio.channels.ClosedChannelException
   at kafka.network.BlockingChannel.send(BlockingChannel.scala:100)
   ```

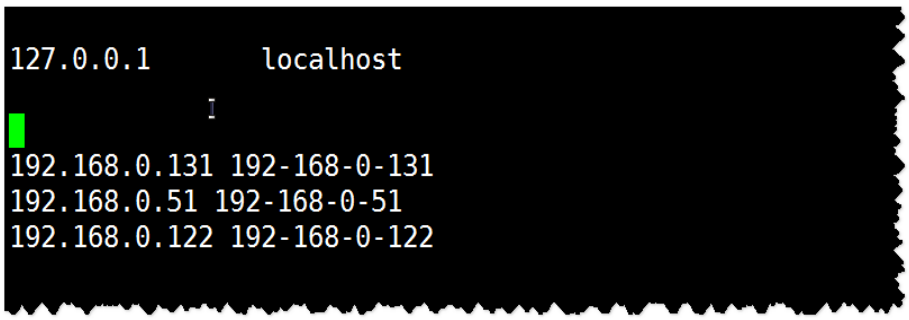
```
at kafka.producer.SyncProducer.liftedTree1$1(SyncProducer.scala:73)
at kafka.producer.SyncProducer.kafka$producer$SyncProducer$$doSend(SyncProducer.scala:72)
at kafka.producer.SyncProducer.send(SyncProducer.scala:113)
at kafka.client.ClientUtils$.fetchTopicMetadata(ClientUtils.scala:58)
at kafka.client.ClientUtils$.fetchTopicMetadata(ClientUtils.scala:93)
at kafka.consumer.ConsumerFetcherManager
$LeaderFinderThread.doWork(ConsumerFetcherManager.scala:66)
at kafka.utils.ShutdownableThread.run(ShutdownableThread.scala:60)
[2016-03-04 03:33:55,615] INFO Disconnecting from 192-168-234-231:21005
(kafka.producer.SyncProducer)
```

通过异常信息，发现当前Consumer无法从Kafka Broker 192-168-234-231节点获取元数据，导致无法连接正确的Broker获取消息。

6. 检查网络是否存在问题，如果网络没有问题，检查是否配置主机和IP的对应关系
   - Linux

     执行**cat /etc/hosts**命令。

     图 16-51 示例 1

   - Windows

     打开“C:\Windows\System32\drivers\etc\hosts”。

     图 16-52 示例 2

     ```
     # For example:
     #
     #      192.168.94.97     rhino.acme.com          # source server
     #      192.168.63.10     x.acme.com              # x client host

     # localhost name resolution is handled within DNS itself.
     #       127.0.0.1       localhost
     #       ::1             localhost
     10.82.129.120 rms.huawei.com # modified by IrmTool at 2015-01-18 17:55:13
     ```

     解决方法参考**步骤4**。

解决办法

步骤1 ZooKeeper连接地址配置错误。

步骤2 修改Consumer配置中的ZooKeeper连接地址信息，保证和MRS相一致。

- Flume

  ```
  server.sources.Source02.type=org.apache.flume.source.kafka.KafkaSource
  server.sources.Source02.zookeeperConnect=192.168.234.231:2181/kafka
  server.sources.Source02.topic = test
  server.sources.Source02.groupId = test_01
  ```

- Spark

  ```
  val zkQuorum = "192.168.234.231:2181/kafka"
  ```

- Storm
  ```
  BrokerHosts brokerHosts = new ZKHosts("192.168.234.231:2181/kafka");
  ```
- Consumer API
  ```
  zookeeper.connect="192.168.234.231:2181/kafka"
  ```

步骤3 Rebalance异常。

同一个消费者组(consumer group)有多个consumer先后启动，就是一个消费者组内有多个consumer同时消费多个partition数据，consumer端也会有负载均衡（consumer个数小于partitions数量时）。

consumer实际上是靠存储在zk中的临时节点来表明针对哪个topic的那个partition拥有读权限的。所在路径为：/consumers/consumer-group-xxx/owners/topic-xxx/x。

当触发负载均衡后，原来的consumer会重新计算并释放已占用的partitions，此过程需要一定的处理时间，新来的consumer抢占该partitions时很有可能会失败。

表 16-3 参数说明

名称	作用	默认值
rebalance.max.retries	Rebalance最大重试次数	4
rebalance.backoff.ms	Rebalance每次重试间隔	2000
zookeeper.session.timeout.ms	Zookeeper连接会话超时时间	15000

可以适当调大上述三个参数，可以参考如下数值：

```
zookeeper.session.timeout.ms = 45000
rebalance.max.retries = 10
rebalance.backoff.ms = 5000
```

参数设置应遵循：

rebalance.max.retries * rebalance.backoff.ms > zookeeper.session.timeout.ms

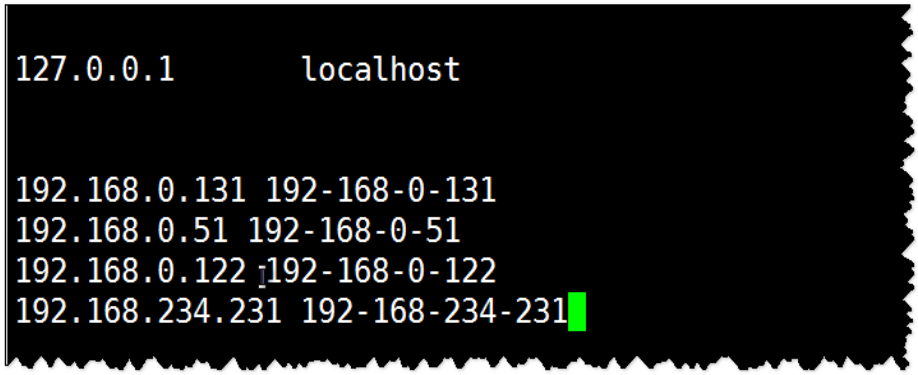
步骤4 网络异常。

在hosts文件中没有配置主机名和IP的对应关系，导致使用主机名进行访问时，无法获取信息。

步骤5 在hosts文件中添加对应的主机名和IP的对应关系。

- Linux

图 16-53 示例 3



- Windows

  图 16-54 示例 4

  ```
  # For example:
  #
  #      192.168.94.97     rhino.acme.com          # source server
  #      192.168.63.10     x.acme.com              # x client host

  # localhost name resolution is handled within DNS itself.
  #       127.0.0.1       localhost
  #       ::1             localhost
  10.82.129.120 rms.huawei.com # modified by IrmTool at 2015-01-18 17:55:13
  192.168.234.231 192-168-234-231
  ```

----结束

16.13.7 Consumer 消费数据失败，Consumer 一直处于等待状态

问题现象

使用MRS服务安装集群，主要安装ZooKeeper、Kafka。

在使用Consumer从Kafka消费数据时，发现客户端一直处于等待状态。

可能原因

1. Kafka服务异常。
2. 客户端Consumer侧采用非安全访问，服务端配置禁止访问。
3. 客户端Consumer侧采用非安全访问，Kafak Topic设置ACL。

原因分析

Consumer向Kafka消费数据失败，可能原因客户端Consumer侧问题或者Kafka侧问题。

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Consumer客户端日志发现频繁连接Broker节点和断链打印信息，如下所示。

   ```
   [root@10-10-144-2 client]# kafka-console-consumer.sh --topic test --zookeeper 10.5.144.2:2181/kafka
   --from-beginning


   [2017-03-07 09:22:00,658] INFO Fetching metadata from broker BrokerEndPoint(1,10.5.144.2,9092)
   with correlation id 26 for 1 topic(s) Set(test) (kafka.client.ClientUtils$)
   [2017-03-07 09:22:00,659] INFO Connected to 10.5.144.2:9092 for producing
   (kafka.producer.SyncProducer)
   [2017-03-07 09:22:00,659] INFO Disconnecting from 10.5.144.2:9092 (kafka.producer.SyncProducer)
   ```

   Consumer采用9092端口来访问Kafka，9092为非安全端口。
3. 通过Manager页面查看当前Kafka集群配置，发现未配置自定义参数“allow.everyone.if.no.acl.found”=“false”。

- MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
- FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择 “集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

4. 当acl设置为false不允许采用9092来进行访问。
5. 查看Consumer客户端日志发现频繁连接Broker节点和断链打印信息，如下所示。

```
[root@10-10-144-2 client]# kafka-console-consumer.sh --topic test_acl --zookeeper 10.5.144.2:2181/kafka --from-beginning


[2017-03-07 09:49:16,992] INFO Fetching metadata from broker BrokerEndPoint(2,10.5.144.3,9092) with correlation id 16 for 1 topic(s) Set(topic_acl) (kafka.client.ClientUtils$)
[2017-03-07 09:49:16,993] INFO Connected to 10.5.144.3:9092 for producing (kafka.producer.SyncProducer)
[2017-03-07 09:49:16,994] INFO Disconnecting from 10.5.144.3:9092 (kafka.producer.SyncProducer)
```

   Consumer采用21005端口来访问Kafka，21005为非安全端口。

6. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
User:test_user has Allow permission for operations: Write from hosts: *
```

   Topic设置acl，则不允许采用9092来访问。

7. 查看Consumer客户端日志发现打印信息，如下所示。

```
[root@10-10-144-2 client]# kafka-console-consumer.sh --topic topic_acl --bootstrap-server 10.5.144.2:21007 --consumer.config /opt/client/Kafka/kafka/config/consumer.properties --from-beginning --new-consumer

[2017-03-07 10:19:18,478] INFO Kafka version : 0.9.0.0 (org.apache.kafka.common.utils.AppInfoParser)
[2017-03-07 10:19:18,478] INFO Kafka commitId : unknown (org.apache.kafka.common.utils.AppInfoParser)
```

   Consumer采用21007端口来访问Kafka。

8. 通过客户端命令klist查询当前认证用户：

```
[root@10-10-144-2 client]# klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: test@HADOOP.COM

Valid starting     Expires            Service principal
01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
```

   当前认证用户为**test**。

9. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:24002/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
 User:test_user has Allow permission for operations: Write from hosts: *
 User:ttest_user has Allow permission for operations: Read from hosts: *
```

   Topic设置acl，test无权限进行consumer操作。

   解决方法参考**步骤2**。

10. 通过SSH登录Kafka Broker：

    通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

    查看kafka-authorizer.log发现如下日志提示用户不属于kafka或者kafkaadmin组。

```
2017-01-25 13:26:33,648 | INFO  | [kafka-request-handler-0] | The principal is test, belongs to Group : [hadoop, ficommon] | kafka.authorizer.logger (SimpleAclAuthorizer.scala:169)
2017-01-25 13:26:33,648 | WARN  | [kafka-request-handler-0] | The user is not belongs to kafka or kafkaadmin group, authorize failed! | kafka.authorizer.logger (SimpleAclAuthorizer.scala:170)
```

解决方法参考**步骤3**。

解决办法

步骤1 配置自定义参数“allow.everyone.if.no.acl.found”参数为“true”，重启Kafka服务。

步骤2 采用具有权限用户登录。

例如：

kinit test_user

或者赋予用户相关权限。

须知

需要使用Kafka管理员用户（属于kafkaadmin组）操作。

例如：

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --topic topic_acl --consumer --add --allow-principal User:test --group test

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=8.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
 User:test_user has Allow permission for operations: Write from hosts: *
  User:test has Allow permission for operations: Describe from hosts: *
  User:test has Allow permission for operations: Write from hosts: *
  User:test has Allow permission for operations: Read from hosts: *
```

步骤3 用户加入**Kafka**组或者**Kafkaadmin**组。

----结束

16.13.8 SparkStreaming 消费 Kafka 消息失败，提示 Error getting partition metadata

问题现象

使用SparkStreaming来消费Kafka中指定Topic的消息时，发现无法从Kafka中获取到数据。提示如下错误： Error getting partition metadata。

```
Exception in thread "main" org.apache.spark.SparkException:  Error getting partition metadata for 'testtopic'. Does the topic exist?
org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
scala.util.Either.fold(Either.scala:97)
org.apache.spark.streaming.kafka.KafkaCluster$.checkErrors(KafkaCluster.scala:365)
org.apache.spark.streaming.kafka.KafkaUtils$.createDirectStream(KafkaUtils.scala:422)
com.xxxxxx.bigdata.spark.examples.FemaleInfoCollectionPrint$.main(FemaleInfoCollectionPrint.scala:45)
com.xxxxxx.bigdata.spark.examples.FemaleInfoCollectionPrint.main(FemaleInfoCollectionPrint.scala)
sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
java.lang.reflect.Method.invoke(Method.java:498)
org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$
$runMain(SparkSubmit.scala:762)
```

```
org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:183)
org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:208)
org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:123)
org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
```

可能原因

1. Kafka服务异常。
2. 客户端Consumer侧采用非安全访问，服务端配置禁止访问。
3. 客户端Consumer侧采用非安全访问，Kafak Topic设置ACL。

原因分析

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 通过Manager页面，查看当前Kafka集群配置，发现未配置“allow.everyone.if.no.acl.found”或配置为“false”。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka > 配置"。
3. 当acl设置为false不允许采用Kafka非安全端口21005来进行访问。
4. 通过客户端命令查看topic的acl权限设置信息：

   ```
   [root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --list --topic topic_acl
   Current ACLs for resource `Topic:topic_acl`:
    User:test_user has Allow permission for operations: Describe from hosts: *
   User:test_user has Allow permission for operations: Write from hosts: *
   ```

   Topic设置acl，则不允许采用Kafka非安全端口21005来访问。

解决办法

步骤1 修改或者添加自定义配置“allow.everyone.if.no.acl.found”参数为“true”，重启Kafka服务。

步骤2 删除Topic设置的ACL。

例如：

kinit test_user

须知

需要使用Kafka管理员用户（属于kafkaadmin组）操作。

例如：

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --remove --allow-principal User:test_user --producer --topic topic_acl

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --remove --allow-principal User:test_user --consumer --topic topic_acl --group test

----结束

16.13.9 新建集群 Consumer 消费数据失败，提示 GROUP_COORDINATOR_NOT_AVAILABLE

问题背景与现象

新建Kafka集群，部署Broker节点数为2，使用Kafka客户端可以正常生产，但是无法正常消费。Consumer消费数据失败，提示GROUP_COORDINATOR_NOT_AVAILABLE，关键日志如下：

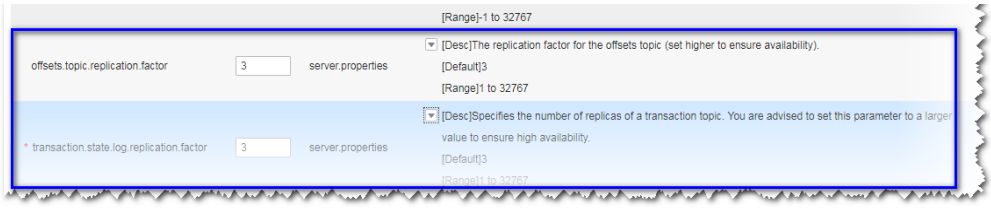
```
2018-05-12 10:58:42,561 | INFO | [kafka-request-handler-3] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 118 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,562 | INFO | [executor-Heartbeat] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
```

可能原因

__consumer_offsets无法创建。

原因分析

1. 查看日志，发现大量__consumer_offset创建异常。
2. 查看集群发现当前Broker数量为2。
3. 查看__cosumer_offset topic要求副本为3，因此创建失败。



解决办法

可以将扩容至至少3个流式core节点，或参考如下步骤修改服务配置参数。

步骤1 进入服务参数配置界面。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > Kafka > 服务配置”，“参数类别”设置为“全部配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager。选择“集群 > 服务 > Kafka”，单击“配置”，选择“全部配置”。

步骤2 搜索并修改offsets.topic.replication.factor和transaction.state.log.replication.factor的值为2。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

16.13.10 SparkStreaming 消费 Kafka 消息失败，提示 Couldn't find leader offsets

问题背景与现象

使用SparkStreaming来消费Kafka中指定Topic的消息时，发现无法从Kafka中获取到数据。提示如下错误： Couldn't find leader offsets。

```
2018-05-30 12:01:17,816 | INFO  | [Driver] | Reconnect due to socket error: java.net.SocketTimeoutException | kafka.utils.Logging$class.info(Logging.scala:68)
2018-05-30 12:01:47,859 | ERROR | [Driver] | User class threw exception: org.apache.spark.SparkException: java.net.SocketTimeoutException
org.apache.spark.SparkException: Couldn't find leader offsets for Set([STK5,57], [STK5,21]) | org.apache.spark.Logging$class.logError(Logging.scala:96)
org.apache.spark.SparkException: java.net.SocketTimeoutException
org.apache.spark.SparkException: Couldn't find leader offsets for Set([STK5,57], [STK5,21])
        at org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
        at org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
        at scala.util.Either.fold(Either.scala:97)
        at org.apache.spark.streaming.kafka.KafkaCluster$.checkErrors(KafkaCluster.scala:365)
        at org.apache.spark.streaming.kafka.KafkaUtils$.createDirectStream(KafkaUtils.scala:422)
        at org.apache.spark.streaming.kafka.KafkaUtils$.createDirectStream(KafkaUtils.scala:532)
        at org.apache.spark.streaming.kafka.KafkaUtils.createDirectStream(KafkaUtils.scala)
        at com.stk.bigdata.sparkstreaming.notify.SparkAlarmControlwz.main(SparkAlarmControlwz.java:194)
        at com.stk.bigdata.sparkstreaming.submit.SparkNotifyA.main(SparkNotifyA.java:14)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.spark.deploy.yarn.ApplicationMaster$$anon$2.run(ApplicationMaster.scala:540)
2018-05-30 12:01:47,863 | INFO  | [Driver] | Final app status: FAILED, exitCode: 15, (reason: User class threw exception: org.apache.spark.SparkException: java.n
org.apache.spark.SparkException: Couldn't find leader offsets for Set([STK5,57], [STK5,21])) | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
2018-05-30 12:01:47,866 | INFO  | [pool-1-thread-1] | Invoking stop() from shutdown hook | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
```

可能原因

- Kafka服务异常。
- 网络异常。
- Kafka Topic异常。

原因分析

步骤1 通过Manager页面，查看Kafka集群当前状态，发现状态为“良好”，且监控指标内容显示正确。

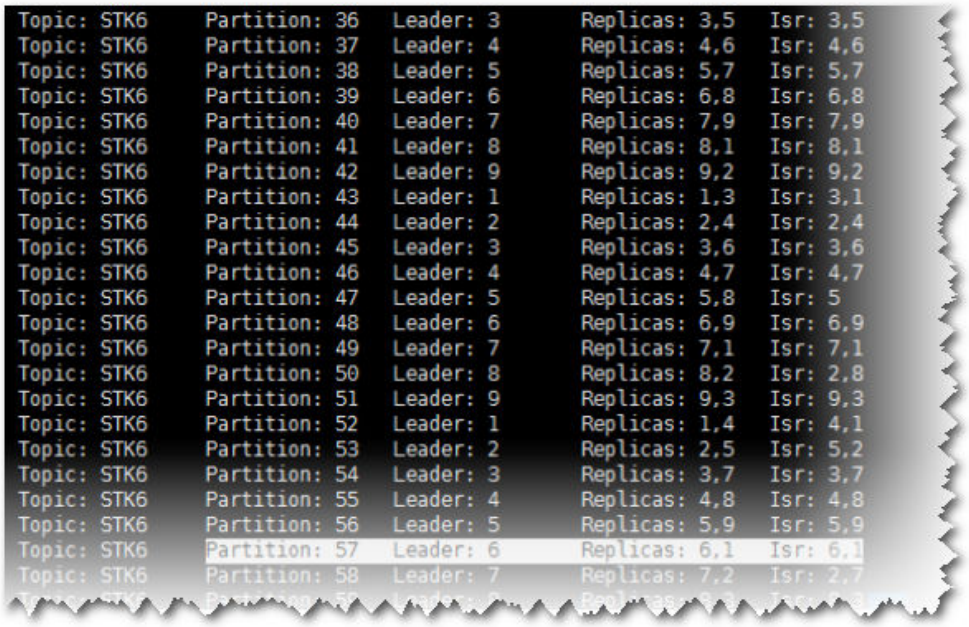
步骤2 查看SparkStreaming日志中提示错误的Topic信息。

执行Kafka相关命令，获取Topic分布信息和副本同步信息，观察返回结果。

kafka-topics.sh --describe --zookeeper <zk_host:port/chroot> --topic <topic name>

如下所示，发现对应Topic状态正常。所有Partition均存在正常Leader信息。

图 16-55 Topic 分布信息和副本同步信息



步骤3 检查客户端与Kafka集群网络是否连通，若网络不通协调网络组进行处理。

步骤4 通过SSH登录Kafka Broker。

通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

查看server.log发现如下日志抛出java.lang.OutOfMemoryError: Direct buffer memory。

```
2018-05-30 12:02:00,246 | ERROR | [kafka-network-thread-6-PLAINTEXT-3] | Processor got uncaught exception. | kafka.network.Processor (Logging.scala:103)

java.lang.OutOfMemoryError: Direct buffer memory
 at java.nio.Bits.reserveMemory(Bits.java:694)
 at java.nio.DirectByteBuffer.<init>(DirectByteBuffer.java:123)
 at java.nio.ByteBuffer.allocateDirect(ByteBuffer.java:311)
 at sun.nio.ch.Util.getTemporaryDirectBuffer(Util.java:241)
 at sun.nio.ch.IOUtil.read(IOUtil.java:195)
 at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:380)
```

at org.apache.kafka.common.network.PlaintextTransportLayer.read(PlaintextTransportLayer.java:110)

步骤5 通过Manager页面，查看当前Kafka集群配置。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > Kafka > 服务配置”，“参数类别”设置为“全部配置”，发现“KAFKA_JVM_PERFORMANCE_OPTS”的中“-XX:MaxDirectMemorySize”值为“1G”。
- FusionInsight Manager界面操作：登录FusionInsight Manager。选择“集群 > 服务 > Kafka”，单击“配置”，选择“全部配置”，发现“KAFKA_JVM_PERFORMANCE_OPTS”的中“-XX:MaxDirectMemorySize”值为“1G”。

步骤6 直接内存配置过小导致报错，而且一旦直接内存溢出，该节点将无法处理新请求，会导致其他节点或者客户端访问超时失败。

----结束

解决办法

步骤1 登录到Manager，进入 Kafka 配置页面。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > Kafka > 服务配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager。选择“集群 > 服务 > Kafka”，单击“配置”。

步骤2 选择“全部配置”，搜索并修改KAFKA_JVM_PERFORMANCE_OPTS的值。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

16.13.11 Consumer 消费数据失败，提示 SchemaException: Error reading field 'brokers'

问题背景与现象

Consumer来消费Kafka中指定Topic的消息时，发现无法从Kafka中获取到数据。提示如下错误： org.apache.kafka.common.protocol.types.SchemaException: Error reading field 'brokers': Error reading field 'host': Error reading string of length 28271, only 593 bytes available。

```
Exception in thread "Thread-0" org.apache.kafka.common.protocol.types.SchemaException: Error reading field 'brokers': Error reading field 'host': Error reading string of length 28271, only 593 bytes available
at org.apache.kafka.common.protocol.types.Schema.read(Schema.java:73)
at org.apache.kafka.clients.NetworkClient.parseResponse(NetworkClient.java:380)
at org.apache.kafka.clients.NetworkClient.handleCompletedReceives(NetworkClient.java:449)
at org.apache.kafka.clients.NetworkClient.poll(NetworkClient.java:269)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.clientPoll(ConsumerNetworkClient.java:360)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.poll(ConsumerNetworkClient.java:224)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.poll(ConsumerNetworkClient.java:192)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.poll(ConsumerNetworkClient.java:163)
atorg.apache.kafka.clients.consumer.internals.AbstractCoordinator.ensureCoordinatorReady(AbstractCoordinator.java:179)
at org.apache.kafka.clients.consumer.KafkaConsumer.pollOnce(KafkaConsumer.java:973)
at org.apache.kafka.clients.consumer.KafkaConsumer.poll(KafkaConsumer.java:937)
at KafkaNew.Consumer$ConsumerThread.run(Consumer.java:40)
```

可能原因

客户端和服务端Jar版本不一致。

解决办法

修改Consumer应用程序中kafka jar，确保和服务端保持一致。

16.13.12 Consumer 消费数据是否丢失排查

问题背景与现象

客户将消费完的数据存入数据库，发现数据与生产数据不一致，怀疑Kafka消费丢数据

可能原因

- 客户代码原因
- Kafka生产数据写入异常
- Kafka消费数据异常

解决办法

Kafka排查：

步骤1 通过consumer-groups.sh来观察写入和消费的offerset的变化情况（生产一定数量的消息，客户端进行消费，观察offerset的变化）。

步骤2 新建一个消费组，用客户端进行消费，然后查看消费的消息。

new-consumer:

kafka-console-consumer.sh --topic <topic name> --bootstrap-server <IP1:PORT, IP2:PORT,...> --new-consumer --consumer.config <config file>

----结束

客户代码排查：

步骤1 查看客户端里有没有提交offerset的报错。

步骤2 如果没有报错把消费的API里加上打印消息，打印少量数据（只打印key即可），查看丢失的数据。

----结束

16.13.13 帐号锁定导致启动组件失败

问题背景与现象

新安装集群，启动Kafka失败。显示认证失败，导致启动失败。

```
/home/omm/kerberos/bin/kinit -k -t /opt/xxxxxx/Bigdata/etc/2_15_ Broker /kafka.keytab kafka/
hadoop.hadoop.com -c /opt/xxxxxx/Bigdata/etc/2_15_ Broker /11846 failed.
export key tab file for kafka/hadoop.hadoop.com failed.export and check keytab file failed, errMsg=]}] for
Broker #192.168.1.92@192-168-1-92.
[2015-07-11 02:34:33] RoleInstance started failure for ROLE[name: Broker].
[2015-07-11 02:34:34] Failed to complete the instances start operation. Current operation entities: [Broker
#192.168.1.92@192-168-1-92], Failure entites : [Broker #192.168.1.92@192-168-1-92].Operation
```

```
Failed.Failed to complete the instances start operation. Current operation entities:
[Broker#192.168.1.92@192-168-1-92], Failure entites: [Broker #192.168.1.92@192-168-1-92].
```

原因分析

查看Kerberos日志，/var/log/Bigdata/kerberos/krb5kdc.log，发现有集群外的IP使用kafka用户连接，导致多次认证失败，最终导致Kafka帐号被锁定。

```
Jul 11 02:49:16 192-168-1-91 krb5kdc[1863](info): AS_REQ (2 etypes {18 17}) 192.168.1.93:
NEEDED_PREAUTH: kafka/hadoop.hadoop.com@HADOOP.COM for krbtgt/HADOOP.COM@HADOOP.COM,
Additional pre-authentication required
Jul 11 02:49:16 192-168-1-91 krb5kdc[1863](info): preauth (encrypted_timestamp) verify failure: Decrypt
integrity check failed
Jul 11 02:49:16 192-168-1-91 krb5kdc[1863](info): AS_REQ (2 etypes {18 17}) 192.168.1.93:
PREAUTH_FAILED: kafka/hadoop.hadoop.com@HADOOP.COM for krbtgt/HADOOP.COM@HADOOP.COM,
Decrypt integrity check failed
```

解决办法

进入集群外的节点（如原因分析示例中的192.168.1.93），断开其对Kafka的认证。等待5分钟，此帐号就会被解锁。

16.13.14 Kafka Broker 上报进程异常，日志提示 IllegalArgumentException

问题背景与现象

使用Manager提示进程故障告警，查看告警进程为Kafka Broker。

可能原因

Broker配置异常。

原因分析

1. 在Manager页面，在告警页面得到主机信息。
2. 通过SSH登录Kafka Broker，执行**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

   查看server.log发现如下日志抛出IllegalArgumentException异常，提示java.lang.IllegalArgumentException: requirement failed: replica.fetch.max.bytes should be equal or greater than message.max.bytes。

   ```
   2017-01-25 09:09:14,930 | FATAL | [main] | | kafka.Kafka$ (Logging.scala:113)
   java.lang.IllegalArgumentException: requirement failed: replica.fetch.max.bytes should be equal or
   greater than message.max.bytes
       at scala.Predef$.require(Predef.scala:233)
       at kafka.server.KafkaConfig.validateValues(KafkaConfig.scala:959)
       at kafka.server.KafkaConfig.<init>(KafkaConfig.scala:944)
       at kafka.server.KafkaConfig$.fromProps(KafkaConfig.scala:701)
       at kafka.server.KafkaConfig$.fromProps(KafkaConfig.scala:698)
       at kafka.server.KafkaServerStartable$.fromProps(KafkaServerStartable.scala:28)
       at kafka.Kafka$.main(Kafka.scala:60)
       at kafka.Kafka.main(Kafka.scala)
   ```

   Kafka要求replica.fetch.max.bytes 需要大于等于message.max.bytes。
3. 进入Kafka配置页面，选择“全部配置”， 显示所有Kafka相关配置，分别搜索message.max.bytes、

   replica.fetch.max.bytes进行检索，发现replica.fetch.max.bytes小于message.max.bytes。

解决办法

步骤1 登录Manager界面，进入Kafka配置页面。

- MRS 3.x之前的版本：登录MRS Manager，选择“服务管理 > Kafka > 配置 > 全部配置”。
- MRS 3.x及后续版本，登录FusionInsight Manager，选择“集群 > 服务 > Kafka > 配置 > 全部配置”。

步骤2 搜索并修改replica.fetch.max.bytes参数，使得replica.fetch.max.bytes的值大于等于message.max.bytes，使得不同Broker上的Partition的Replica可以同步到全部消息。

步骤3 保存配置，查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。

步骤4 修改消费者业务应用中fetch.message.max.bytes，使得fetch.message.max.bytes的值大于等于message.max.bytes，确保消费者可以消费到全部消息。

----结束

16.13.15 执行 Kafka Topic 删除操作，发现无法删除

问题背景与现象

在使用Kafka客户端命令删除Topic时，发现Topic无法被删除。

```
kafka-topics.sh --delete --topic test --zookeeper 192.168.234.231:2181/kafka
```

可能原因

- 客户端命令连接ZooKeeper地址错误。
- Kafka服务异常Kafka部分节点处于停止状态。
- Kafka服务端配置禁止删除。
- Kafka配置自动创建，且Producer未停止。

原因分析

1. 客户端命令，打印ZkTimeoutException异常。

   ```
   [2016-03-09 10:41:45,773] WARN Can not get the principle name from server 192.168.234.231 (org.apache.zookeeper.ClientCnxn)
   Exception in thread "main" org.I0Itec.zkclient.exception.ZkTimeoutException: Unable to connect to zookeeper server within timeout: 30000
   at org.I0Itec.zkclient.ZkClient.connect(ZkClient.java:880)
   at org.I0Itec.zkclient.ZkClient.<init>(ZkClient.java:98)
   at org.I0Itec.zkclient.ZkClient.<init>(ZkClient.java:84)
   at kafka.admin.TopicCommand$.main(TopicCommand.scala:51)
   at kafka.admin.TopicCommand.main(TopicCommand.scala)
   ```

   解决方法参考**步骤1**。

2. 客户端查询命令。

   kafka-topics.sh --list --zookeeper 192.168.0.122:2181/kafka

   ```
   test - marked for deletion
   ```

   通过Manager查看Kafka Broker实例的运行状态。

   通过**cd /var/log/Bigdata/kafka/broker**命令进入RunningAsController节点日志目录，在controller.log发现ineligible for deletion: test。

```
2016-03-09 11:11:26,228 | INFO  | [main] | [Controller 1]: List of topics to be deleted:  |
kafka.controller.KafkaController (Logging.scala:68)
2016-03-09 11:11:26,229 | INFO  | [main] | [Controller 1]: List of topics ineligible for deletion: test |
kafka.controller.KafkaController (Logging.scala:68)
```

3. 通过Manager查询Broker删除Topic相关配置。

   解决方法参考**步骤2**

4. 客户端查询命令：

   ```
   kafka-topics.sh --describe –topic test --zookeeper 192.168.0.122:2181/kafka
   ```

   进入RunningAsController节点日志目录，在controller.log发现marked ineligible for deletion。

   ```
   2016-03-10 11:11:17,989 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Handling
   deletion for topics test | kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
   2016-03-10 11:11:17,990 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Not retrying
   deletion of topic test at this time since it is marked ineligible for deletion |
   kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
   ```

5. 通过Manager查询Broker状态。

   其中一个Broker处于停止或者故障状态。Topic进行删除必须保证该Topic的所有Partition所在的Broker必须处于正常状态。

   解决方法参考**步骤3**。

6. 进入RunningAsController节点日志目录，在controller.log发现Deletion successfully，然后又出现New topics: [Set(test)]，表明被再次创建。

   ```
   2016-03-10 11:33:35,208 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Deletion of topic
   test successfully completed | kafka.controller.TopicDeletionManager$DeleteTopicsThread
   (Logging.scala:68)

   2016-03-10 11:33:38,501 | INFO  | [ZkClient-
   EventThread-19-192.168.0.122:2181,160.172.0.52:2181,160.172.0.51:2181/kafka] |
   [TopicChangeListener on Controller 3]: New topics: [Set(test)], deleted topics: [Set()], new partition
   replica assignment
   ```

7. 通过Manager查询Broker创建Topic相关配置。

   经确认，对该Topic操作的应用没有停止。

   解决方法参考**步骤4**。

解决办法

步骤1 ZooKeeper连接失败导致。

Kafka客户端连接ZooKeeper服务超时。检查客户端到ZooKeeper的网络连通性。

网络连接失败，通过Manager界面查看Zookeeper服务信息。

配置错误，修改客户端命令中ZooKeeper地址。

步骤2 Kafka服务端配置禁止删除。

通过Manager界面修改delete.topic.enable为true。保存配置并重启服务。

客户端查询命令，无Topic:test。

```
kafka-topics.sh --list --zookeeper 192.168.0.122:24002/kafka
```

进入RunningAsController节点日志目录，在controller.log发现Deletion of topic test successfully。

```
2016-03-10 10:39:40,665 | INFO  | [delete-topics-thread-3] | [Partition state machine on Controller 3]:
Invoking state change to OfflinePartition for partitions [test,2],[test,15],[test,6],[test,16],[test,12],[test,7],
[test,10],[test,13],[test,9],[test,19],[test,3],[test,5],[test,1],[test,0],[test,17],[test,8],[test,4],[test,11],[test,14],
[test,18] | kafka.controller.PartitionStateMachine (Logging.scala:68)
```

```
2016-03-10 10:39:40,668 | INFO  | [delete-topics-thread-3] | [Partition state machine on Controller 3]: Invoking state change to NonExistentPartition for partitions [test,2],[test,15],[test,6],[test,16],[test,12],[test,7],[test,10],[test,13],[test,9],[test,19],[test,3],[test,5],[test,1],[test,0],[test,17],[test,8],[test,4],[test,11],[test,14],[test,18] | kafka.controller.PartitionStateMachine (Logging.scala:68)
2016-03-10 10:39:40,977 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Deletion of topic test successfully completed | kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
```

步骤3 Kafka部分节点处于停止或者故障状态。

启动停止的Broker实例。

客户端查询命令，无Topic:test。

```
kafka-topics.sh --list --zookeeper 192.168.0.122:24002/kafka
```

进入RunningAsController节点日志目录，在controller.log发现Deletion of topic test successfully。

```
2016-03-10 11:17:56,463 | INFO  | [delete-topics-thread-3] | [Partition state machine on Controller 3]: Invoking state change to NonExistentPartition for partitions [test,4],[test,1],[test,8],[test,2],[test,5],[test,9],[test,7],[test,6],[test,0],[test,3] | kafka.controller.PartitionStateMachine (Logging.scala:68)
2016-03-10 11:17:56,726 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Deletion of topic test successfully completed | kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
```

步骤4 Kafka配置自动创建，且Producer未停止。

停止相关应用，通过Manager界面修改“auto.create.topics.enable”为“false”，保存配置并重启服务。

步骤5 再次执行delete操作。

----结束

16.13.16 执行 Kafka Topic 删除操作，提示 AdminOperationException

问题背景与现象

在使用Kafka客户端命令设置Topic ACL权限时，发现Topic无法被设置。

```
kafka-topics.sh --delete --topic test4 --zookeeper 10.5.144.2:2181/kafka
```

提示错误ERROR kafka.admin.AdminOperationException: Error while deleting topic test4。

具体如下：

```
Error while executing topic command : Error while deleting topic test4
[2017-01-25 14:00:20,750] ERROR kafka.admin.AdminOperationException: Error while deleting topic test4
at kafka.admin.TopicCommand$$anonfun$deleteTopic$1.apply(TopicCommand.scala:177)
at kafka.admin.TopicCommand$$anonfun$deleteTopic$1.apply(TopicCommand.scala:162)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
at kafka.admin.TopicCommand$.deleteTopic(TopicCommand.scala:162)
at kafka.admin.TopicCommand$.main(TopicCommand.scala:68)
at kafka.admin.TopicCommand.main(TopicCommand.scala)
 (kafka.admin.TopicCommand$)
```

可能原因

用户不属于**kafkaadmin**组，Kafka提供安全访问接口，kafkaamdin组用户才可以进行topic删除操作。

原因分析

1. 使用客户端命令，打印AdminOperationException异常。
2. 通过客户端命令**klist**查询当前认证用户：

   ```
   [root@10-10-144-2 client]# klist
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: test@HADOOP.COM

   Valid starting     Expires            Service principal
   01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   如上例中当前认证用户为test。
3. 通过命令**id**查询用户组信息

   ```
   [root@10-10-144-2 client]# id test
   uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10003(kafka)
   ```

解决办法

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 用户管理”。

步骤3 在操作用户对应的“操作”列，单击“修改”。

步骤4 为用户加入**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

16.13.17 执行 Kafka Topic 创建操作，发现无法创建提示 NoAuthException

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现Topic无法被创建。

```
kafka-topics.sh --create --zookeeper 192.168.234.231:2181/kafka --replication-factor 1 --partitions 2 --
topic test
```

提示错误NoAuthException，KeeperErrorCode = NoAuth for /config/topics。

具体如下：

```
Error while executing topic command org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /config/topics
org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /config/topics
 at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
 at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:685)
 at org.I0Itec.zkclient.ZkClient.create(ZkClient.java:304)
 at org.I0Itec.zkclient.ZkClient.createPersistent(ZkClient.java:213)
 at kafka.utils.ZkUtils$.createParentPath(ZkUtils.scala:215)
 at kafka.utils.ZkUtils$.updatePersistentPath(ZkUtils.scala:338)
 at kafka.admin.AdminUtils$.writeTopicConfig(AdminUtils.scala:247)
```

可能原因

用户不属于**kafkaadmin**组，Kafka提供安全访问接口，kafkaamdin组用户才可以进行topic删除操作。

原因分析

1. 使用客户端命令，打印NoAuthException异常。

   ```
   Error while executing topic command org.apache.zookeeper.KeeperException$NoAuthException:
   KeeperErrorCode = NoAuth for /config/topics
   org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException:
   KeeperErrorCode = NoAuth for /config/topics
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:685)
    at org.I0Itec.zkclient.ZkClient.create(ZkClient.java:304)
    at org.I0Itec.zkclient.ZkClient.createPersistent(ZkClient.java:213)
    at kafka.utils.ZkUtils$.createParentPath(ZkUtils.scala:215)
    at kafka.utils.ZkUtils$.updatePersistentPath(ZkUtils.scala:338)
    at kafka.admin.AdminUtils$.writeTopicConfig(AdminUtils.scala:247)
   ```

2. 通过客户端命令**klist**查询当前认证用户：

   ```
   [root@10-10-144-2 client]# klist
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: test@HADOOP.COM

   Valid starting     Expires            Service principal
   01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   如上例中当前认证用户为**test**。

3. 通过命令**id**查询用户组信息。

   ```
   [root@10-10-144-2 client]# id test
   uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10003(kafka)
   ```

解决办法

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 用户管理”。

步骤3 在操作用户对应的“操作”列，单击“修改”。

步骤4 为用户加入**kafkaadmin**组。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

16.13.18 执行 Kafka Topic 设置 ACL 操作失败，提示 NoAuthException

问题背景与现象

在使用Kafka客户端命令设置Topic ACL权限时，发现Topic无法被设置。

```
kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka  --topic topic_acl --producer
--add --allow-principal User:test_acl
```

提示错误NoAuthException: KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002。

具体如下：

```
Error while executing ACL command: org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002
org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002
at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:995)
at org.I0Itec.zkclient.ZkClient.delete(ZkClient.java:1038)
at kafka.utils.ZkUtils.deletePath(ZkUtils.scala:499)
at kafka.common.ZkNodeChangeNotificationListener$$anonfun$purgeObsoleteNotifications
$1.apply(ZkNodeChangeNotificationListener.scala:118)
at kafka.common.ZkNodeChangeNotificationListener$$anonfun$purgeObsoleteNotifications
$1.apply(ZkNodeChangeNotificationListener.scala:112)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
at
kafka.common.ZkNodeChangeNotificationListener.purgeObsoleteNotifications(ZkNodeChangeNotificationLis
tener.scala:112)
at kafka.common.ZkNodeChangeNotificationListener.kafka$common$ZkNodeChangeNotificationListener$
$processNotifications(ZkNodeChangeNotificationListener.scala:97)
at
kafka.common.ZkNodeChangeNotificationListener.processAllNotifications(ZkNodeChangeNotificationListener
.scala:77)
at kafka.common.ZkNodeChangeNotificationListener.init(ZkNodeChangeNotificationListener.scala:65)
at kafka.security.auth.SimpleAclAuthorizer.configure(SimpleAclAuthorizer.scala:136)
at kafka.admin.AclCommand$.withAuthorizer(AclCommand.scala:73)
at kafka.admin.AclCommand$.addAcl(AclCommand.scala:80)
at kafka.admin.AclCommand$.main(AclCommand.scala:48)
```

```
at kafka.admin.AclCommand.main(AclCommand.scala)
Caused by: org.apache.zookeeper.KeeperException$NoAuthException: KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002
at org.apache.zookeeper.KeeperException.create(KeeperException.java:117)
at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
at org.apache.zookeeper.ZooKeeper.delete(ZooKeeper.java:1416)
at org.I0Itec.zkclient.ZkConnection.delete(ZkConnection.java:104)
at org.I0Itec.zkclient.ZkClient$11.call(ZkClient.java:1042)
at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:985)
```

可能原因

用户不属于**kafkaadmin**组，Kafka提供安全访问接口，**kafkaamdin**组用户才可以进行设置操作。

原因分析

1. 使用客户端命令，打印NoAuthException异常。
2. 通过客户端命令**klist**查询当前认证用户：

   ```
   [root@10-10-144-2 client]# klist
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: test@HADOOP.COM

   Valid starting     Expires            Service principal
   01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   如上例中当前认证用户为**test**。
3. 通过命令**id**查询用户组信息。

   ```
   [root@10-10-144-2 client]# id test
   uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10003(kafka)
   ```

解决办法

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 用户管理”。

步骤3 在操作用户对应的“操作”列，单击“修改”。

步骤4 为用户加入**kafkaadmin**组。

步骤5 通过命令**id**查询用户组信息。

```
[root@host1 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),10003(kafka)
```

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

16.13.19 执行 Kafka Topic 创建操作，发现无法创建提示 NoNode for /brokers/ids

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现Topic无法被创建。

```
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper
192.168.234.231:2181
```

提示错误NoNodeException: KeeperErrorCode = NoNode for /brokers/ids。

具体如下：

```
Error while executing topic command : org.apache.zookeeper.KeeperException$NoNodeException:
KeeperErrorCode = NoNode for /brokers/ids
[2017-09-17 16:35:28,520] ERROR org.I0Itec.zkclient.exception.ZkNoNodeException:
org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /brokers/ids
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:47)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:995)
    at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:675)
    at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:671)
    at kafka.utils.ZkUtils.getChildren(ZkUtils.scala:541)
    at kafka.utils.ZkUtils.getSortedBrokerList(ZkUtils.scala:176)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:235)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:105)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:60)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
Caused by: org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /
brokers/ids
    at org.apache.zookeeper.KeeperException.create(KeeperException.java:115)
    at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
    at org.apache.zookeeper.ZooKeeper.getChildren(ZooKeeper.java:2256)
    at org.apache.zookeeper.ZooKeeper.getChildren(ZooKeeper.java:2284)
    at org.I0Itec.zkclient.ZkConnection.getChildren(ZkConnection.java:114)
    at org.I0Itec.zkclient.ZkClient$4.call(ZkClient.java:678)
    at org.I0Itec.zkclient.ZkClient$4.call(ZkClient.java:675)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:985)
    ... 8 more
 (kafka.admin.TopicCommand$)
```

可能原因

- Kafka服务处于停止状态。
- 客户端命令中zookeeper地址参数配置错误。

原因分析

1. 使用客户端命令，打印NoNodeException异常。

   ```
   Error while executing topic command : org.apache.zookeeper.KeeperException$NoNodeException:
   KeeperErrorCode = NoNode for /brokers/ids
   [2017-09-17 16:35:28,520] ERROR org.I0Itec.zkclient.exception.ZkNoNodeException:
   org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /brokers/ids
       at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:47)
       at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:995)
       at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:675)
       at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:671)
   ```

```
    at kafka.utils.ZkUtils.getChildren(ZkUtils.scala:541)
    at kafka.utils.ZkUtils.getSortedBrokerList(ZkUtils.scala:176)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:235)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:105)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:60)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
```

2. 通过Manager查看Kafka服务是否处于正常状态。
3. 检查客户端命令中ZooKeeper地址是否正确，访问ZooKeeper上所存放的Kafka信息，其路径（Znode）应该加上/kafka，发现配置中缺少/kafka：

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper
192.168.234.231:2181
```

解决办法

步骤1 保证Kafka服务处于正常状态。

步骤2 创建命令中ZooKeeper地址信息需要添加/kafka。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper
192.168.234.231:2181/kafka
```

----结束

16.13.20 执行 Kakfa Topic 创建操作，发现无法创建提示 replication factor larger than available brokers

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现Topic无法被创建。

```
kafka-topics.sh --create  --replication-factor 2 --partitions 2 --topic test --zookeeper
192.168.234.231:2181
```

提示错误replication factor larger than available brokers。

具体如下：

```
Error while executing topic command : replication factor: 2 larger than available brokers: 0
[2017-09-17 16:44:12,396] ERROR kafka.admin.AdminOperationException: replication factor: 2 larger than
available brokers: 0
    at kafka.admin.AdminUtils$.assignReplicasToBrokers(AdminUtils.scala:117)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:403)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:110)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:61)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
 (kafka.admin.TopicCommand$)
```

可能原因

- Kafka服务处于停止状态。
- Kafka服务当前可用Broker小于设置的replication-factor。
- 客户端命令中Zookeeper地址参数配置错误。

原因分析

1. 使用客户端命令，打印replication factor larger than available brokers异常。

```
Error while executing topic command : replication factor: 2 larger than available brokers: 0
[2017-09-17 16:44:12,396] ERROR kafka.admin.AdminOperationException: replication factor: 2 larger
```

```
than available brokers: 0
    at kafka.admin.AdminUtils$.assignReplicasToBrokers(AdminUtils.scala:117)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:403)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:110)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:61)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
 (kafka.admin.TopicCommand$)
```

2. 通过Manager参看Kafka服务是否处于正常状态，当前可用Broker是否小于设置的replication-factor。
3. 检查客户端命令中ZooKeeper地址是否正确，访问ZooKeeper上所存放的Kafka信息，其路径（Znode）应该加上/kafka，发现配置中缺少/kafka。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 2 --partitions 2 --topic test --zookeeper
192.168.234.231:2181
```

解决办法

步骤1 保证Kafka服务处于正常状态，且可用Broker不小于设置的replication-factor。

步骤2 创建命令中ZooKeeper地址信息需要添加/kafka。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper
192.168.234.231:2181/kafka
```

----结束

16.13.21 Consumer 消费数据存在重复消费现象

问题背景与现象

当数据量较大时会频繁的发生rebalance导致出现重复消费的情况，关键日志如下：

```
2018-05-12 10:58:42,561 | INFO | [kafka-request-handler-3] | [GroupCoordinator 2]: Preparing to restabilize
group DemoConsumer with old generation 118 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:58:43,245 | INFO | [kafka-request-handler-5] | [GroupCoordinator 2]: Stabilized group
DemoConsumer generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:58:43,560 | INFO | [kafka-request-handler-7] | [GroupCoordinator 2]: Assignment received
from leader for group DemoConsumer for generation 119 | kafka.coordinator.GroupCoordinator
(Logging.scala:68)
2018-05-12 10:59:13,562 | INFO | [executor-Heartbeat] | [GroupCoordinator 2]: Preparing to restabilize
group DemoConsumer with old generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,790 | INFO | [kafka-request-handler-3] | [GroupCoordinator 2]: Stabilized group
DemoConsumer generation 120 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,791 | INFO | [kafka-request-handler-0] | [GroupCoordinator 2]: Assignment received
from leader for group DemoConsumer for generation 120 | kafka.coordinator.GroupCoordinator
(Logging.scala:68)
2018-05-12 10:59:43,802 | INFO | [kafka-request-handler-2] | Rolled new log segment for
'__consumer_offsets-17' in 2 ms. | kafka.log.Log (Logging.scala:68)
2018-05-12 10:59:52,456 | INFO | [group-metadata-manager-0] | [Group Metadata Manager on Broker 2]:
Removed 0 expired offsets in 0 milliseconds. | kafka.coordinator.GroupMetadataManager (Logging.scala:68)
2018-05-12 11:00:49,772 | INFO | [kafka-scheduler-6] | Deleting segment 0 from log __consumer_offsets-17.
| kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:49,773 | INFO | [kafka-scheduler-6] | Deleting index /srv/BigData/kafka/data4/kafka-logs/
__consumer_offsets-17/00000000000000000000.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:00:49,773 | INFO | [kafka-scheduler-2] | Deleting segment 2147948547 from log
__consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:49,773 | INFO | [kafka-scheduler-4] | Deleting segment 4282404355 from log
__consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:49,775 | INFO | [kafka-scheduler-2] | Deleting index /srv/BigData/kafka/data4/kafka-logs/
__consumer_offsets-17/00000000002147948547.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:00:49,775 | INFO | [kafka-scheduler-4] | Deleting index /srv/BigData/kafka/data4/kafka-logs/
__consumer_offsets-17/00000000004282404355.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:00:50,533 | INFO | [kafka-scheduler-6] | Deleting segment 4283544095 from log
__consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
```

```
2018-05-12 11:00:50,569 | INFO | [kafka-scheduler-6] | Deleting index /srv/BigData/kafka/data4/kafka-logs/__consumer_offsets-17/00000000004283544095.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:02:21,178 | INFO | [kafka-request-handler-2] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 120 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 11:02:22,839 | INFO | [kafka-request-handler-4] | [GroupCoordinator 2]: Stabilized group DemoConsumer generation 121 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 11:02:23,169 | INFO | [kafka-request-handler-1] | [GroupCoordinator 2]: Assignment received from leader for group DemoConsumer for generation 121 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 11:02:49,913 | INFO | [kafka-request-handler-6] | Rolled new log segment for '__consumer_offsets-17' in 2 ms. | kafka.log.Log (Logging.scala:68)
```

其中Preparing to restabilize group DemoConsumer with old generation表示正在发生rebalance。

可能原因

参数设置不合理。

原因分析

原因：由于参数设置不当，数据量大时数据处理时间过长，导致频繁发生balance，此时offset无法正常提交，导致重复消费数据。

原理：每次poll的数据处理完后才提交offset，如果poll数据后的处理时长超出了session.timeout.ms的设置时长，此时发生rebalance导致本次消费失败，已经消费数据的offset无法正常提交，所以下次重新消费时还是在旧的offset消费数据，从而导致消费数据重复。

解决办法

建议用户在Manager页面调整以下服务参数：

request.timeout.ms=100000

session.timeout.ms=90000

max.poll.records=50

heartbeat.interval.ms=3000

其中：

request.timeout.ms要比session.timeout.ms大10s。

session.timeout.ms的大小设置要在服务端参数group.min.session.timeout.ms和group.max.session.timeout.ms之间。

以上参数可以根据实际情况进行适当的调整，特别是max.poll.records，这个参数是为了控制每次poll数据的records量，保证每次的处理时长尽量保持稳定。目的是为了保证poll数据以后的处理时间不要超过session.timeout.ms的时间。

参考信息

- poll之后的数据处理效率要高，不要阻塞下一次poll
- poll方法和数据处理建议异步处理

16.13.22 执行 Kafka Topic 创建操作，发现 Partition 的 Leader 显示为 none

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现创建Topic Partition的Leader显示为none。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper 10.6.92.36:2181/kafka

Created topic "test".
```

```
[root@10-10-144-2 client]#
kafka-topics.sh --describe --zookeeper 10.6.92.36:2181/kafka

Topic:test      PartitionCount:2        ReplicationFactor:2     Configs:
        Topic: test     Partition: 0    Leader: none        Replicas: 2,3   Isr:
        Topic: test     Partition: 1    Leader: none        Replicas: 3,1   Isr:
```

可能原因

- Kafka服务处于停止状态。
- 找不到用户组信息。

原因分析

1. 查看kafka服务状态及监控指标。
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 在Kafka概览页面获取Controller节点信息。
3. 登录Controller所在节点，通过**cd /var/log/Bigdata/kafka/broker**命令进入节点日志目录，在state-change.log发现存在ZooKeeper权限异常，提示NoAuthException。

   ```
   2018-05-31 09:20:42,436 | ERROR | [ZkClient-EventThread-34-10.6.92.36:24002,10.6.92.37:24002,10.6.92.38:24002/kafka] | Controller 4 epoch 6 initiated state change for partition [test,1] from NewPartition to OnlinePartition failed | state.change.logger (Logging.scala:103)

   org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException: KeeperErrorCode = NoAuth for /brokers/topics/test/partitions
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:1000)
    at org.I0Itec.zkclient.ZkClient.create(ZkClient.java:527)
    at org.I0Itec.zkclient.ZkClient.createPersistent(ZkClient.java:293)
   ```

4. 查看对应时间段ZooKeeper审计日志，权限异常。

   ```
   2018-05-31 09:20:42,421 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18 user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode    target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/0/state    result=failure
   2018-05-31 09:20:42,423 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18 user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode    target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/0    result=failure
   ```

```
2018-05-31 09:20:42,435 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions    result=failure
2018-05-31 09:20:42,439 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/1/state    result=failure
2018-05-31 09:20:42,441 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/1    result=failure
2018-05-31 09:20:42,453 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions    result=failure
```

5. 在ZooKeeper各个实例节点上执行**id -Gn kafka**命令，发现有一个节点无法查询用户组信息。

```
[root @bdpsit3ap03 ~]# id -Gn kafka
id:  kafka: No such user
[root @bdpsit3ap03 ~]#
```

6. MRS集群中的用户管理由LDAP服务管理提供，又依赖于操作系统的sssd（red hat），nscd（suse）服务，用户的建立到同步到sssd服务需要一定时间，如果此时用户没有生效，或者sssd版本存在bug的情况下，某些情况下在ZooKeeper节点会出现用户无效的情况，导致创建Topic异常。

解决办法

步骤1 重启sssd/nscd服务。

- Red Hat

```
service sssd restart
```

- SUSE

```
sevice nscd restart
```

步骤2 重启相关服务后，在节点通过**id username**命令查看相应用户信息是否已有效。

----结束

16.13.23 Kafka 安全使用说明

Kafka API 简单说明

- 新Producer API

  指org.apache.kafka.clients.producer.KafkaProducer中定义的接口，在使用“kafka-console-producer.sh”时，默认使用此API。

- 旧Producer API

  指kafka.producer.Producer中定义的接口，在使用“kafka-console-producer.sh”时，加“--old-producer”参数会调用此API。

- 新Consumer API

  指org.apache.kafka.clients.consumer.KafkaConsumer中定义的接口，在使用“kafka-console-consumer.sh”时，加“--new-consumer”参数会调用此API。

- 旧Consumer API

  指kafka.consumer.ConsumerConnector中定义的接口，在使用“kafka-console-consumer.sh”时，默认使用此API。

说明

新Producer API和新Consumer API，在下文中统称为新API。

Kafka 访问协议说明

Kafka当前支持四种协议类型的访问：PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL。

Kafka服务启动时，默认会启动PLAINTEXT和SASL_PLAINTEXT两种协议类型的访问监听。可通过设置Kafka服务配置参数“ssl.mode.enable”为“true”，来启动SSL和SASL_SSL两种协议类型的访问监听。

下表是四中协议类型的简单说明：

协议类型	说明	支持的API	默认端口
PLAINTEXT	支持无认证的明文访问	新API和旧API	9092
SASL_PLAINTEXT	支持Kerberos认证的明文访问	新API	21007
SSL	支持无认证的SSL加密访问	新API	9093
SASL_SSL	支持Kerberos认证的SSL加密访问	新API	21009

Topic 的 ACL 设置

Kafka支持安全访问，因此可以针对Topic进行ACL设置，从而控制不同的用户可以访问不同的Topic。Topic的权限信息，需要在Linux客户端上，使用“kafka-acls.sh”脚本进行查看和设置。

- 操作场景

  该任务指导Kafka管理员根据业务需求，为其他使用Kafka的系统用户授予相关Topic的特定权限。

  Kafka默认用户组信息表所示。

用户组名	描述
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限，才能访问对应Topic。

- 前提条件

a. 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组）。

b. 已安装Kafka客户端。

- 操作步骤

a. 以客户端安装用户，登录安装Kafka客户端的节点。

b. 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。

cd /opt/kafkaclient

c. 执行以下命令，配置环境变量。

source bigdata_env

d. 执行以下命令，进行用户认证。（普通集群跳过此步骤）

kinit *组件业务用户*

e. 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

f. 使用“kafka-acl.sh”进行用户授权常用命令如下：

- 查看某Topic权限控制列表：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --list --topic <Topic名称>

- 添加给某用户Producer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --add --allow-principal User:<用户名> --producer --topic <Topic名称>

- 删除某用户Producer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --remove --allow-principal User:<用户名> --producer --topic <Topic名称>

- 添加给某用户Consumer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --add --allow-principal User:<用户名> --consumer --topic <Topic名称> --group <消费者组名称>

- 删除某用户Consumer权限：

  ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --remove --allow-principal User:<用户名> --consumer --topic <Topic名称> --group <消费者组名称>

针对不同的 Topic 访问场景，Kafka 新旧 API 使用说明

- 场景一：访问设置了ACL的Topic

使用的API	用户属组	客户端参数	服务端参数	访问的端口
新API	用户需满足以下条件之一即可： • 属于系统管理员组 • 属于kafkaadmin组 • 属于kafkasuperuser组 • 被授权的kafka组的用户	security.protocol=SASL_PLAINTEXT sasl.kerberos.service.name = kafka	-	sasl.port（默认21007）
		security.protocol=SASL_SSL sasl.kerberos.service.name = kafka	ssl.mode.enable配置为true	sasl-ssl.port（默认21009）
旧API	不涉及	不涉及	不涉及	不涉及

- 场景二：访问未设置ACL的Topic

使用的API	用户属组	客户端参数	服务端参数	访问的端口
新API	用户需满足以下条件之一： • 属于系统管理员组 • 属于kafkaadmin组 • 属于kafkasuperuser组	security.protocol=SASL_PLAINTEXT sasl.kerberos.service.name = kafka	-	sasl.port（默认21007）
	用户属于kafka组		allow.everyone.if.no.acl.found配置为true	sasl.port（默认21007）
	用户需满足以下条件之一： • 属于系统管理员组 • 属于kafkaadmin组 • kafkasuperuser组用户	security.protocol=SASL_SSLsasl.kerberos.service.name = kafka	ssl-enable配置为“true”	sasl-ssl.port（默认21009）

使用的API	用户属组	客户端参数	服务端参数	访问的端口
	用户属于kafka组		allow.everyone.if.no.acl.found配置为“true” ssl-enable配置为“true”	sasl-ssl.port（默认21009）
	-	security.protocol=PLAINTEXT	allow.everyone.if.no.acl.found配置为“true”	port（默认21005）
	-	security.protocol=SSL	allow.everyone.if.no.acl.found配置为“true” ssl-enable配置为“true”	ssl.port（默认21008）
旧Producer	-	-	allow.everyone.if.no.acl.found配置为“true”	port（默认21005）
旧Consumer	-	-	allow.everyone.if.no.acl.found配置为“true”	ZooKeeper服务端口：clientPort（默认24002）

16.13.24 如何获取 Kafka Consumer Offset 信息

问题背景与现象

使用Kafka Consumer消费数据时，如何获取Kafka Consumer Offset相关信息?

Kafka API 简单说明

- 新Producer API

  指org.apache.kafka.clients.producer.KafkaProducer中定义的接口，在使用“kafka-console-producer.sh”时，默认使用此API。

- 旧Producer API

  指kafka.producer.Producer中定义的接口，在使用“kafka-console-producer.sh”时，加“--old-producer”参数会调用此API。

- 新Consumer API

  指org.apache.kafka.clients.consumer.KafkaConsumer中定义的接口，在使用“kafka-console-consumer.sh”时，加“--new-consumer”参数会调用此API。

- 旧Consumer API

  指kafka.consumer.ConsumerConnector中定义的接口，在使用“kafka-console-consumer.sh”时，默认使用此API。

说明

新Producer API和新Consumer API，在下文中统称为新API。

处理步骤

旧Consumer API

- 前提条件
  a. 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组）。
  b. 已安装Kafka客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。
     cd /opt/kafkaclient
  c. 执行以下命令，配置环境变量。
     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通模式跳过此步骤）
     kinit *组件业务用户*
  e. 执行以下命令，切换到Kafka客户端安装目录。
     cd Kafka/kafka/bin
  f. 执行以下命令，获取consumer offset metric信息。

```
bin/kafka-consumer-groups.sh --zookeeper <zookeeper_host:port>/kafka --list
```

```
bin/kafka-consumer-groups.sh --zookeeper <zookeeper_host:port>/kafka --describe --group test-consumer-group
```

例如：

```
kafka-consumer-groups.sh --zookeeper 192.168.100.100:2181/kafka --list
kafka-consumer-groups.sh --zookeeper 192.168.100.100:2181/kafka --describe --group test-consumer-group
```

新Consumer API

- 前提条件
  a. 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组）。
  b. 已安装Kafka客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/client”。
     cd /opt/client
  c. 执行以下命令，配置环境变量。
     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通模式跳过此步骤）
     kinit *组件业务用户*

e. 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

f. 执行以下命令，获取consumer offset metric信息。

kafka-consumer-groups.sh --bootstrap-server <broker_host:port> --describe --group my-group

例如：

kafka-consumer-groups.sh --bootstrap-server 192.168.100.100:9092 --describe --group my-group

16.13.25 如何针对 Topic 进行配置增加和删除

问题背景与现象

使用Kafka过程中常常需要对特定Topic进行配置或者修改。

Topic级别可以修改参数列表：

```
cleanup.policy
compression.type
delete.retention.ms
file.delete.delay.ms
flush.messages
flush.ms
index.interval.bytes
max.message.bytes
min.cleanable.dirty.ratio
min.insync.replicas
preallocate
retention.bytes
retention.ms
segment.bytes
segment.index.bytes
segment.jitter.ms
segment.ms
unclean.leader.election.enable
```

处理步骤

- 前提条件

  已安装Kafka客户端。

- 操作步骤

  a. 以客户端安装用户，登录安装Kafka客户端的节点。

  b. 切换到Kafka客户端安装目录，例如“/opt/client”。

  cd /opt/client

  c. 执行以下命令，配置环境变量。

  source bigdata_env

  d. 执行以下命令，进行用户认证。（普通模式跳过此步骤）

  kinit *组件业务用户*

  e. 执行以下命令，切换到Kafka客户端安装目录。

  cd Kafka/kafka/bin

  f. 执行以下命令，针对Topic修改配置和删除配置。

  kafka-topics.sh --alter --topic <topic_name> --zookeeper <zookeeper_host:port>/kafka --config <name=value>

kafka-topics.sh --alter --topic <topic_name> --zookeeper <zookeeper_host:port>/kafka --delete-config <name>

例如：

kafka-topics.sh --alter --topic test1 --zookeeper 192.168.100.100:2181/kafka --config retention.ms=86400000

kafka-topics.sh --alter --topic test1 --zookeeper 192.168.100.100:2181/kafka --delete-config retention.ms

g. 执行以下命令，查询topic信息。

kafka-topics.sh --describe -topic <topic_name> --zookeeper <zookeeper_host:port>/kafka

16.13.26 如何读取“__consumer_offsets”内部 topic 的内容

用户问题

kafka如何将consumer 消费的offset保存在内部topic“ __consumer_offsets”中?

处理步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证。（普通集群跳过此步骤）

kinit *组件业务用户*

步骤5 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤6 执行以下命令，获取consumer offset metric信息。

kafka-console-consumer.sh --topic __consumer_offsets --zookeeper <zk_host:port>/kafka --formatter "kafka.coordinator.group.GroupMetadataManager\$OffsetsMessageFormatter" --consumer.config <property file> --from-beginning

其中<property file>配置文件中需要增加如下内容。

```
exclude.internal.topics = false
```

例如：

kafka-console-consumer.sh --topic __consumer_offsets --zookeeper 10.5.144.2:2181/kafka --formatter "kafka.coordinator.group.GroupMetadataManager\$OffsetsMessageFormatter" --consumer.config ../config/consumer.properties --from-beginning

```
[example-group1,test2,0]::[OffsetMetadata[0,NO_METADATA],CommitTime 1487121209218,ExpirationTime 148720760
9218]
[example-group1,test2,1]::[OffsetMetadata[0,NO_METADATA],CommitTime 1487121209218,ExpirationTime 148720760
9218]
[example-group1,test2,0]::[OffsetMetadata[2,NO_METADATA],CommitTime 1487121269208,ExpirationTime 148720760
9208]
[example-group1,test2,1]::[OffsetMetadata[1,NO_METADATA],CommitTime 1487121269208,ExpirationTime 148720760
```

----结束

16.13.27 如何配置客户端 shell 命令的日志

用户问题

如何设置客户端shell命令的日志输出级别?

处理步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤3 切换到Kafka客户端配置目录。

cd Kafka/kafka/config

步骤4 编辑tools-log4j.properties文件，将WARN修改为INFO，并保存。

```
log4j.rootLogger=WARN, stderr

log4j.appender.stderr=org.apache.log4j.ConsoleAppender
log4j.appender.stderr.layout=org.apache.log4j.PatternLayout
log4j.appender.stderr.layout.ConversionPattern=[%d] %p %m (%c)%n
log4j.appender.stderr.Target=System.err
```

```
log4j.rootLogger=INFO, stderr

log4j.appender.stderr=org.apache.log4j.ConsoleAppender
log4j.appender.stderr.layout=org.apache.log4j.PatternLayout
log4j.appender.stderr.layout.ConversionPattern=[%d] %p %m (%c)%n
log4j.appender.stderr.Target=System.err
```

步骤5 切换到Kafka客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤6 执行以下命令，配置环境变量。

source bigdata_env

步骤7 执行以下命令，进行用户认证。（普通集群跳过此步骤）

kinit *组件业务用户*

步骤8 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤9 执行以下命令，获取topic信息，在控制台可见日志打印。

```
kafka-topics.sh --list --zookeeper 10.5.144.2:2181/kafka
[2017-02-17 14:34:27,005] INFO JAAS File name: /opt/client/Kafka/./kafka/config/jaas.conf
(org.I0Itec.zkclient.ZkClient)
[2017-02-17 14:34:27,007] INFO Starting ZkClient event thread. (org.I0Itec.zkclient.ZkEventThread)
[2017-02-17 14:34:27,013] INFO Client environment:zookeeper.version=V100R002C10, built on 05/12/2016
08:56 GMT (org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:host.name=10-10-144-2
(org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:java.version=1.8.0_72
(org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:java.vendor=Oracle Corporation
(org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:java.home=/opt/client/JDK/jdk/jre
(org.apache.zookeeper.ZooKeeper)
Test
__consumer_offsets
counter
test
test2
test3
test4
```

----结束

16.13.28 如何获取 Topic 的分布信息

用户问题

如何获取Topic在Broker实例的分布信息?

前置操作

- 前提条件
  已安装Kafka、ZooKeeper客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/client”。
     cd /opt/client
  c. 执行以下命令，配置环境变量。
     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通集群跳过此步骤）
     kinit *组件业务用户*
  e. 执行以下命令，切换到Kafka客户端安装目录。
     cd Kafka/kafka/bin
  f. 执行Kafka相关命令，获取Topic分布信息和副本同步信息，观察返回结果。
     kafka-topics.sh --describe --zookeeper <zk_host:port/chroot>
     例如：

     ```
     [root@mgtdat-sh-3-01-3 client]#kafka-topics.sh --describe --zookeeper 10.149.0.90:2181/
     kafka
     Topic:topic1   PartitionCount:2  ReplicationFactor:2     Configs:
     Topic: topic1  Partition: 0 Leader: 26 Replicas: 23,25 Isr: 26
     Topic: topic1  Partition: 1 Leader: 24 Replicas: 24,23 Isr: 24,23
     ```

     其中，Replicas对应副本分布信息，Isr对应副本同步信息。

处理方法 1

1. 在ZooKeeper中查询Broker ID的对应关系。

   sh zkCli.sh -server <zk_host:port>

2. 在ZooKeeper客户端执行如下命令。

   ls /kafka/brokers/ids

   get /kafka/brokers/ids/<查询出的Broker id>

   例如：

```
[root@node-master1gAMQ kafka]# zkCli.sh -server node-master1gAMQ:2181
Connecting to node-master1gAMQ:2181
Welcome to ZooKeeper!
JLine support is enabled

WATCHER::

WatchedEvent state:SyncConnected type:None path:null
[zk: node-master1gAMQ:2181(CONNECTED) 0] ls /kafka/brokers/
ids      seqid    topics
[zk: node-master1gAMQ:2181(CONNECTED) 0] ls /kafka/brokers/ids
[1]
[zk: node-master1gAMQ:2181(CONNECTED) 1] get /kafka/brokers/ids/1
{"listener_security_protocol_map":{"PLAINTEXT":"PLAINTEXT","SSL":"SSL"},"endpoints":["PLAINTEXT://
192.168.2.242:9092","SSL://192.168.2.242:9093"],"rack":"/default/
rack0","jmx_port":21006,"host":"192.168.2.242","timestamp":"1580886124398","port":9092,"version":4}
[zk: node-master1gAMQ:2181(CONNECTED) 2]
```

处理方法 2

获取节点和broker ID的对应关系

kafka-broker-info.sh --zookeeper <zk_host:port/chroot>

例如：

```
[root@node-master1gAMQ kafka]# bin/kafka-broker-info.sh --zookeeper 192.168.2.70:2181/kafka
Broker_ID     IP_Address
---------------------------
 1          192.168.2.242
```

16.13.29 Kafka 高可靠使用说明

Kafka 高可靠、高可用说明

Kafka消息传输保障机制，可以通过配置不同的参数来保障消息传输，进而满足不同的性能和可靠性要求的应用场景。

- **Kafka高可用、高性能**

  如果业务需要保证高可用和高性能，可以采用参数：

参数	默认值	说明
unclean.leader.election.enable	true	是否允许不在ISR中的副本被选举为Leader，若设置为true，可能会造成数据丢失。

参数	默认值	说明
auto.leader.rebalance.enable	true	是否使用Leader自动均衡功能。 如果设为true，Controller会周期性的为所有节点的每个分区均衡Leader，将Leader分配给更优先的副本。
acks	1	需要Leader确认消息是否已经接收并认为已经处理完成。该参数会影响消息的可靠性和性能。 ● acks=0 ： 如果设置为0，Producer将不会等待服务端任何响应。消息将会被认为成功。 ● acks=1 ： 如果设置为1，当副本所在Leader确认数据已写入，但是其不会等待所有的副本完全写入即返回响应。在这种情况下，如果Leader确认后但是副本未同步完成时Leader异常，那么数据就会丢失。 ● acks=-1 ： 如果设置为-1（all），意味着等待所有的同步副本确认后才认为成功，配合min.insync.replicas可以确保多副本写入成功，只要有一个副本保持活跃状态，记录将不会丢失。 **说明** 该参数在kafka客户端配置文件中配置。
min.insync.replicas	1	当Producer设置acks为-1时，指定需要写入成功的副本的最小数目。

配置高可用、高性能的影响：

须知

配置高可用、高性能模式后，数据可靠性会降低。在磁盘故障、节点故障等场景下存在数据丢失风险。

- **Kafka高可靠性配置说明**

  如果业务需要保证数据高可靠性，可以采用相关参数：

参数	建议值	说明
unclean.leader.election.enable	false	是否允许不在ISR中的副本被选举为Leader。

acks	-1	Producer需要Leader确认消息是否已经接收并认为已经处理完成。 acks=-1需要表示等待在ISR列表的副本都确认接收到消息并处理完成才表示消息成功。配合min.insync.replicas可以确保多副本写入成功，只要有一个副本保持活跃状态，记录将不会丢失。 **说明** 该参数在kafka客户端配置文件中配置。
min.insync.replicas	2	当Producer设置acks为-1时，指定需要写入成功的副本的最小数目。 需要满足min.insync.replicas <= replication.factor。

配置高可靠性的影响：

- 性能降低：

  需要所有的ISR列表副本，且满足最小成功的副本数确认写入成功。这样会导致单条消息时延增加，客户端处理能力下降，具体性能以现场实际测试数据为准。

- 可用性降低：

  不允许不在ISR中的副本被选举为Leader。如果Leader下线时，其他副本均不在ISR列表中，那么该分区将保持不可用，直到Leader节点恢复。

  需要所有的ISR列表副本，且满足最小成功的副本数确认写入成功。当分区的一个副本所在节点故障时，无法满足最小成功的副本数，那么将会导致业务写入失败。

配置影响

请根据业务场景对可靠性和性能要求进行评估，采用合理参数配置。

说明

- 对于价值数据，两种场景下建议Kafka数据目录磁盘配置raid1或者raid5，从而提高单个磁盘故障情况下数据可靠性。
- 不同Producer API对应的acks参数名称不同
  - 新Producer API

    指**org.apache.kafka.clients.producer.KafkaProducer**中定义的接口，acks配置为acks。
  - 旧Producer API

    指**kafka.producer.Producer**中定义的接口，acks配置名称为request.required.acks。
- 参数配置项均为Topic级别可修改的参数，默认采用服务级配置。可针对不同Topic可靠性要求对Topic进行单独配置。

  例如，配置Topic名称为test的可靠性参数：

  kafka-topics.sh --zookeeper 192.168.1.205:2181/kafka --alter --topic test --config unclean.leader.election.enable=false --config min.insync.replicas=2

  其中192.168.1.205为ZooKeeper业务IP地址。
- 如果修改服务级配置需要重启Kafka，建议在变更窗口做服务级配置修改。

16.13.30 Kafka 生产者写入单条记录过长问题

问题背景与现象

用户在开发一个Kafka应用，作为一个生产者调用新接口（org.apache.kafka.clients.producer.*）往Kafka写数据，单条记录大小为1100055，超过了kafka配置文件server.properties中message.max.bytes=1000012。用户修改了Kafka服务配置中message.max.bytes大小为5242880，同时也将replica.fetch.max.bytes大小修改为5242880后，仍然无法成功。报异常大致如下：

```
.....
14749 [Thread-0] INFO  com.xxxxxx.bigdata.kafka.example.NewProducer  - The ExecutionException occured : {}.
java.util.concurrent.ExecutionException: org.apache.kafka.common.errors.RecordTooLargeException: The message is 1100093 bytes when serialized which is larger than the maximum request size you have configured with the max.request.size configuration.
at org.apache.kafka.clients.producer.KafkaProducer$FutureFailure.<init>(KafkaProducer.java:739)
at org.apache.kafka.clients.producer.KafkaProducer.doSend(KafkaProducer.java:483)
at org.apache.kafka.clients.producer.KafkaProducer.send(KafkaProducer.java:430)
at org.apache.kafka.clients.producer.KafkaProducer.send(KafkaProducer.java:353)
at com.xxxxxx.bigdata.kafka.example.NewProducer.run(NewProducer.java:150)
Caused by: org.apache.kafka.common.errors.RecordTooLargeException: The message is **** bytes when serialized which is larger than the maximum request size you have configured with the max.request.size configuration.
.....
```

原因分析

经分析因为在写数据到Kafka时，Kafka客户端会先比较配置项“max.request.size ”值和本次写入数据大小，若写入数据大小超过此配置项“max.request.size ”的缺省值，则抛出上述异常。

解决办法

步骤1 在初始化Kafka生产者实例时，设置此配置项“max.request.size ”的值。

例如，参考本例，可以将此配置项设置为“5252880”：

```
// 协议类型:当前支持配置为SASL_PLAINTEXT或者PLAINTEXT
props.put(securityProtocol, kafkaProc.getValues(securityProtocol, "SASL_PLAINTEXT"));
// 服务名
props.put(saslKerberosServiceName, "kafka");
props.put("max.request.size", "5252880");
.....
```

----结束

16.13.31 Kakfa 消费者读取单条记录过长问题

问题背景与现象

和“Kafka生产者写入单条记录过长问题”相对应的，在写入数据后，用户开发一个应用，以消费者调用新接口（org.apache.kafka.clients.consumer.*）到Kafka上读取数据，但读取失败，报异常大致如下：

```
.....
1687 [KafkaConsumerExample] INFO  org.apache.kafka.clients.consumer.internals.AbstractCoordinator  - Successfully joined group DemoConsumer with generation 1
1688 [KafkaConsumerExample] INFO  org.apache.kafka.clients.consumer.internals.ConsumerCoordinator  - Setting newly assigned partitions [default-0, default-1, default-2] for group DemoConsumer
2053 [KafkaConsumerExample] ERROR com.xxxxxx.bigdata.kafka.example.NewConsumer  -
```

```
[KafkaConsumerExample], Error due to
org.apache.kafka.common.errors.RecordTooLargeException: There are some messages at [Partition=Offset]:
{default-0=177} whose size is larger than the fetch size 1048576 and hence cannot be ever returned.
Increase the fetch size on the client (using max.partition.fetch.bytes), or decrease the maximum message
size the broker will allow (using message.max.bytes).
2059 [KafkaConsumerExample] INFO  com.xxxxxx.bigdata.kafka.example.NewConsumer  -
[KafkaConsumerExample], Stopped
.......
```

原因分析

经分析因为在读取数据时Kafka客户端会比较待读取数据大小和配置项“max.partition.fetch.bytes”值，若超过此配置项值，则抛出上述异常。

解决办法

步骤1 在初始化建立Kafka消费者实例时，设置此配置项“max.partition.fetch.bytes”的值。

例如，参考本例，可以将此配置项设置为“5252880”：

```
......
// 安全协议类型
props.put(securityProtocol, kafkaProc.getValues(securityProtocol, "SASL_PLAINTEXT"));
// 服务名
props.put(saslKerberosServiceName, "kafka");

props.put("max.partition.fetch.bytes","5252880");
......
```

----结束

16.13.32 Kafka 集群节点内多磁盘数据量占用高处理办法

用户问题

Kafka流式集群节点内有多块磁盘的使用量很高。当达到100%时就会造成kafka不可用如何处理?

问题现象

客户创建的MRS Kafka流式集群节点内有多块磁盘，由于分区不合理及业务原因导致某几个磁盘的使用量很高。当达到100%时就会造成kafka不可用。

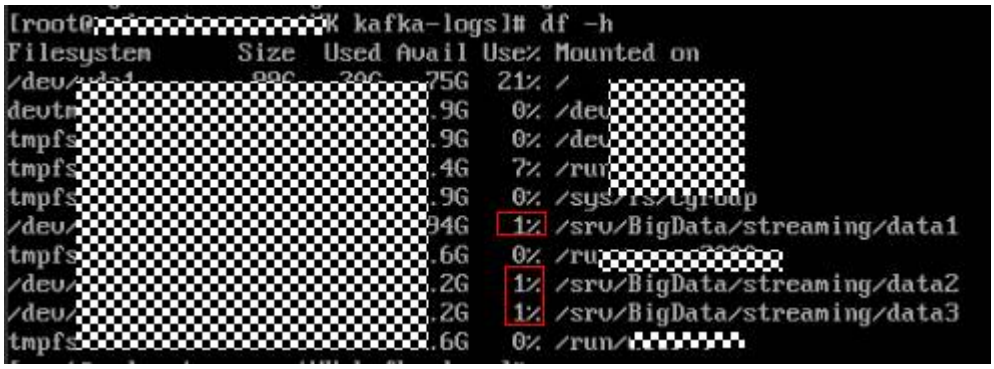
原因分析

需要提前干预处理磁盘数据，全局的log.retention.hours修改需要重启服务。为了不断服，可以将数据量大的单个topic老化时间根据需要改短。

处理步骤

步骤1 登录Kafka集群的l流式Core节点。

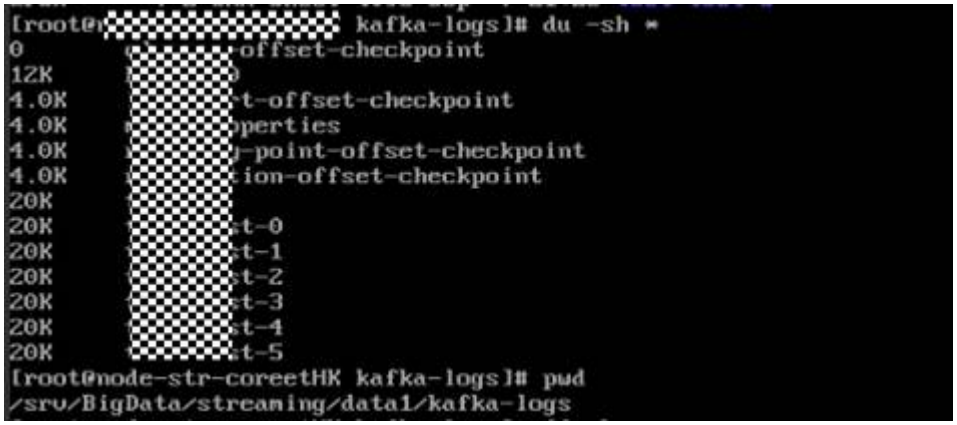
步骤2 执行**df –h**命令查看磁盘使用率。

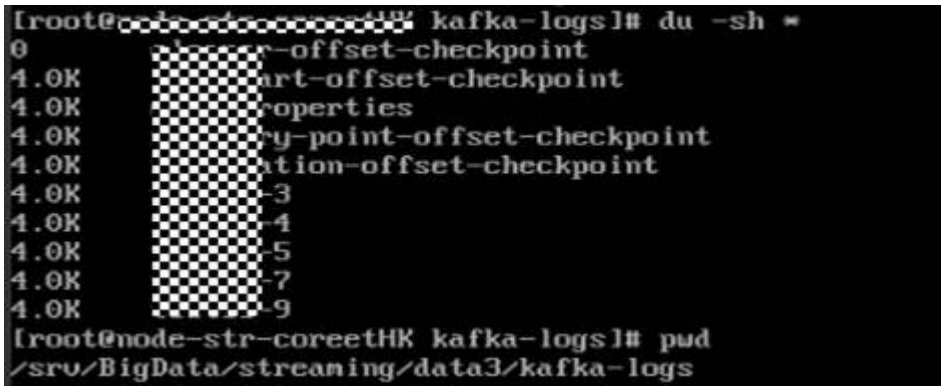


步骤3 通过kafka配置文件**opt/Bigdata/MRS_2.1.0/1_11_Broker/etc/server.properties**中的配置项**log.dirs** 获得数据存储目录。其中配置文件路径请根据时间环境的集群版本修改，当磁盘有多块时，该配置项有多个，逗号间隔。

步骤4 使用**cd**命令进入使用率较高的磁盘对应的**步骤3**中获取的数据存储目录下。

步骤5 使用**du -sh ***命令打印出当前topic的名称及大小。





步骤6 由于kafka的全局的数据保留时间默认是7天。部分topic由于业务写入量大，而这些topic的分区正好在上面使用量高的磁盘上，因此导致磁盘使用率较高。

- 可以通过修改全局数据的保留期为较短时间来释放磁盘空间，该方式需要重启Kafka服务才能生效，可能会影响业务运行。具体请参见**步骤7**。
- 可以单独将topic的数据保留期改为较短时间来释放磁盘空间，该方式无需重启Kafka服务即可生效。具体请参见**步骤8**。

步骤7 登录Manager页面，在Kafka的服务配置页面，切换为“全部配置”并搜索“log.retention.hours”配置项，该值默认为7天，请根据需要进行修改。

步骤8 可以单独将这些磁盘上的topic的数据老化时间修改为较短时间来解决该问题。

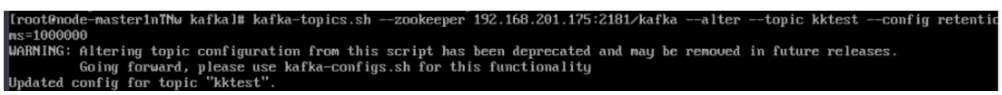
1. 查看topic数据过期时间。

   bin/kafka-topics.sh --describe --zookeeper *<ZooKeeper集群业务IP>*:2181/kafka --topic kktest

2. 设置topic数据过期时间，其中**--topic**表示具体topic名称，retention.ms=具体的数据过期时间，单位是毫秒。

   kafka-topics.sh --zookeeper *<ZooKeeper集群业务IP>*:2181/kafka --alter --topic kktest --config retention.ms=1000000

   设置数据过期时间之后会可能不会立刻执行，删除操作在参数**log.retention.check.interval.ms**所规定时间之后开始执行删，可以通过查看kafka的server.log检索是否有delete字段有判断删除操作是否生效，有delete字段

则表示已经 生效，也可以通过执行**df –h**命令查看磁盘的数据量占用情况判断设置是否生效。

```
log.retention.check.interval.ms = 300000
```

----结束

16.14 使用 Oozie

16.14.1 当并发提交大量 oozie 任务时，任务一直没有运行

用户问题

并发提交大量oozie任务的时候，任务一直没有运行。

问题现象

并发提交大量oozie任务的时候，任务一直没有运行。

原因分析

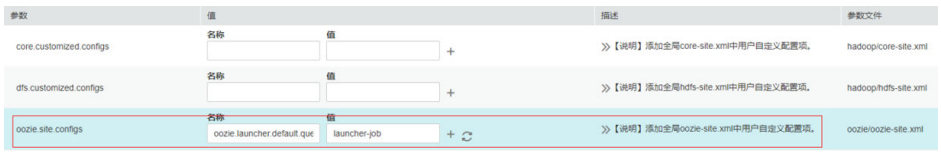
Oozie提交任务会先启动一个oozie-launcher，然后由oozie-launcher提交真正的作业运行。默认情况下launcher和真实作业会在同一个队列中。

当并发提交大量oozie任务的时候就有可能出现启动了一堆oozie-launcher，将队列的资源耗完，而没有更多资源启动真实作业，最终导致任务一直没有运行。

处理步骤

步骤1 参考“用户指南 > 管理现有集群 > 租户管理 > 添加租户”章节新建一个队列给oozie使用，也可以直接使用创建MRS集群时生成的launcher-job队列。

步骤2 在Manager页面选择“集群 > 服务 > Oozie > 配置”，搜索参数“oozie.site.configs”，在值列添加名称“oozie.launcher.default.queue”，值为“launcher-job”。



----结束

16.15 使用 Presto

16.15.1 配置 sql-standard-with-group 创建 schema 失败报 Access Denied

用户问题

配置sql-standard-with-group创建schema失败，报Access Denied的错误，如何处理?

问题现象

```
CREATE SCHEMA hive.sf2 WITH (location = 'obs://obs-zy1234/sf2');Query 20200224_031203_00002_g6gzy failed: Access Denied: Cannot create schema sf2
```

原因分析

presto创建schema需要hive的管理者权限。

处理步骤

MRS Manager界面操作：

- 方法一：
  a. 登录MRS Manager页面，选择“系统设置 > 用户管理”。
  b. 在对应用户所在行的“操作”列，单击“修改”。
  c. 单击“选择并绑定角色”，为用户添加System_administrator的权限。
  d. 单击“确定”完成修改。
- 方法二：
  a. 登录MRS Manager页面，选择“系统设置 > 角色管理”。
  b. 单击“添加角色”，并配置如下参数。
     - 角色名称：配置角色名称，例如hive_admin。
     - 权限：选择“Hive”，并勾选Hive Admin Privilege。
  c. 单击“确定”保存角色。
  d. 选择“系统设置 > 用户管理”。
  e. 在对应用户所在行的“操作”列，单击“修改”。
  f. 单击“选择并绑定角色”，为用户添加新创建的hive_admin的权限。
  g. 单击“确定”完成修改。

FusionInsight Manager界面操作：

- 方法一：
  a. 登录FusionInsight Manager页面，选择“系统 > 权限 > 用户”。
  b. 在对应用户所在行的“操作”列，单击“修改”。
  c. 单击角色后的“添加”，为用户添加System_administrator的权限。
  d. 单击“确定”完成修改。
- 方法二：

a. 登录FusionInsight Manager页面，选择“系统 > 权限 > 角色”。
b. 单击“添加角色”，并配置如下参数。
   - 角色名称：配置角色名称，例如hive_admin。
   - 配置资源权限：选择“Hive”，并勾选“Hive管理员权限”。
c. 单击“确定”保存角色。
d. 选择“系统 > 权限 > 用户”。
e. 在对应用户所在行的“操作”列，单击“修改”。
f. 单击角色后的“添加”，为用户添加新创建的hive_admin的权限。
g. 单击“确定”完成修改。

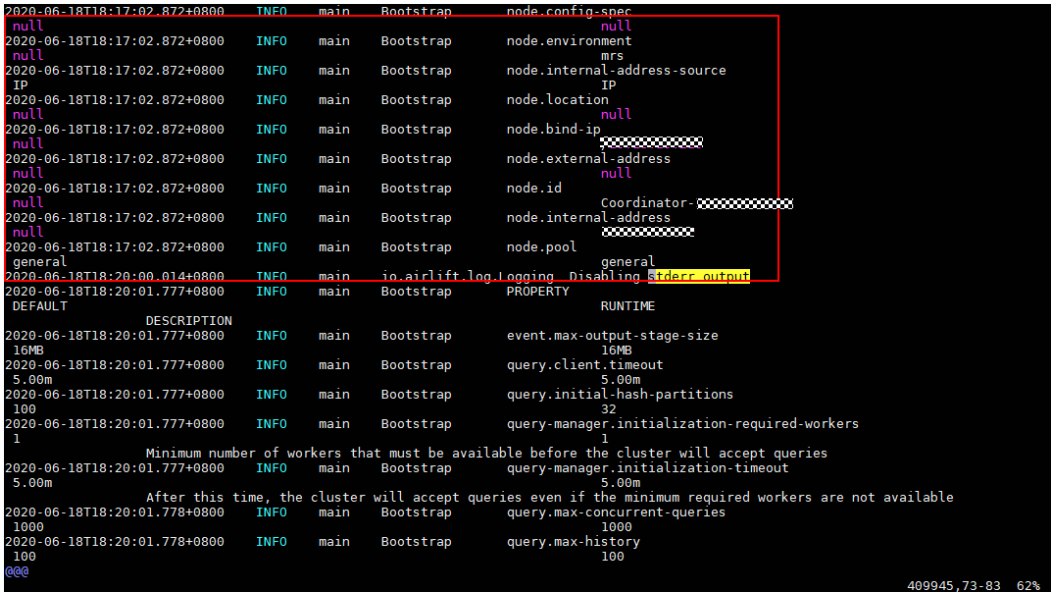
16.15.2 Presto 的 coordinator 无法正常启动

用户问题

Presto的coordinator未知原因被kill，或者presto的coordinator进程无法正常启动。

问题现象

Presto的coordinator无法正常启动，Manager页面上显示presto coordinator进程正常启动且状态正常，但查看后台日志coordinator进程未真正启动，只有如下日志：



presto的coordinator未真正启动即被Kill了，不再打印其他日志，查看presto的其他日志也未发现为何被kill。

原因分析

presto的健康检查脚本的端口检查逻辑中未做好端口的区分。

处理步骤

步骤1 使用工具分别登录集群的Master节点执行如下操作。

步骤2 执行如下命令编辑文件。

vim /opt/Bigdata/MRS_xxx/install/FusionInsight-Presto-*/ha/module/harm/plugin/script/pcd.sh

该文件中的第31行修改为“http_port_exists=$(netstat -apn | awk '{print $4, $6}' | grep :${HTTP_PORT} | grep LISTEN | wc -l)”。

```
25
26 check_status()
27 {
28         proc_exists=$(ps -ef | grep com.facebook.presto.server.PrestoServer | grep -v grep | wc -l)
29         param="-u $PRESTO_SERVER/v1/cluster"
30         if [[ ${proc_exists} == 1 ]]; then
31             http_port_exists=$(netstat -apn | awk '{print $4, $6}' | grep :${HTTP_PORT} | grep LISTEN | wc -l)
32
33             if [[ ${http_port_exists} == 1 ]] ;then
34                 log ${PCD_LOG_FILE} "INFO" "return [ normal ]"
35                 return 0
36             else
37                 log ${PCD_LOG_FILE} "ERROR" "HTTP PORT does not exist, return [ abnormal ]"
38                 return 2
39             fi
40         else
41             log ${PCD_LOG_FILE} "INFO" " coordinator process not exists, return [ abnormal ]"
42             return 2
43         fi
44 }
45
```

步骤3 保存如上修改，再在manager页面上选择“服务管理 > Presto > 实例”，重启Coordinator进程。

----结束

16.15.3 Presto 查询 Kudu 表报错

用户问题

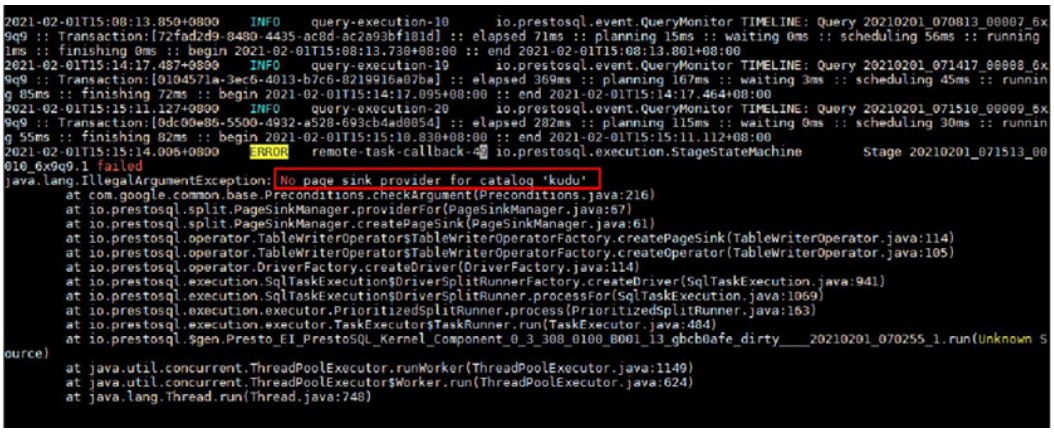
使用presto查询Kudu表报错。

问题现象

使用presto查询Kudu表，报表找不到的错误：

```
presto:default> show tables;
          Table

 impala::default.kudu_taobao
 impala::default.kudu_tt
 impala::default.kudutest
(3 rows)

Query 20210201_030636_00026_95mzd, FINISHED, 4 nodes
Splits: 53 total, 53 done (100.00%)
0:00 [3 rows, 125B] [18 rows/s, 766B/s]

presto:default> select count(*) from kudu.default.kudu_taobao;
Query 20210201_030653_00027_95mzd failed: line 1:22: Table kudu.default.kudu_taobao does not exist
select count(*) from kudu.default.kudu_taobao

presto:default> select count(*) from kudu_taobao;
Query 20210201_030939_00028_95mzd failed: line 1:22: Table kudu.default.kudu_taobao does not exist
select count(*) from kudu_taobao

presto:default>
```

后台报错：



原因分析

在实际的运行节点（worker实例所在节点）没有kudu相关配置。

处理步骤

步骤1 在集群presto所有的worker实例节点添加配置文件kudu.properties。

配置文件保存路径：/opt/Bigdata/MRS_xxx/1_x_Worker/etc/catalog/ (请根据集群实际版本修改路径)

配置文件内容：

```
connector.name=kudu
kudu.client.master-addresses=KuduMasterIP1:port,KuduMasterIP2:port,KuduMasterIP3:port
```

说明

- KuduMaster节点IP和端口请根据实际情况填写。
- 为配置文件添加和文件保存路径下其他文件一致的文件权限、属组。

步骤2 修改完成之后，请在集群详情页面选择“组件管理 > Kudu”，单击“更多 > 重启服务”。

----结束

16.15.4 Presto 查询 Hive 表无数据

用户问题

使用presto查询Hive表无数据。

问题现象

通过Tez引擎执行union相关语句写入的数据，Presto无法查询。

原因分析

由于Hive使用Tez引擎在执行union语句时，生成的输出文件会保存在HIVE_UNION_SUBDIR目录中，而Presto默认不读取子目录下的文件，所以没有读取到HIVE_UNION_SUBDIR目录下的数据。

处理步骤

步骤1 在集群详情页面选择“组件管理 > Presto > 服务配置”。

步骤2 切换“基础配置”为全部配置“。

步骤3 在左侧导航处选择“Presto > Hive”，在catalog/hive.properties文件中增加hive.recursive-directories参数，值为true。

步骤4 单击“保存配置”并勾选“重新启动受影响的服务或实例。”。

----结束

16.16 使用 Spark

16.16.1 Spark 应用下修改 split 值时报错

用户问题

在Spark应用下修改split值时报错。

问题现象

客户需要通过修改一个split最大值来实现多个mapper ，从而达到提速的目的, 但是目前执行**set $参数**命令修改Hive的配置时报错。

```
0: jdbc:hive2://192.168.1.18:21066/> set mapred.max.split.size=1000000;
Error: Error while processing statement: Cannot nodify mapred.max.split.size at runtime. It is not in list of params that are allowed to be modified at runtime( state=42000,code=1)
```

原因分析

- 在安全模式下配置白名单启停参数**hive.security.whitelist.switch**时，需要运行的参数必须在**hive.security.authorization.sqlstd.confwhitelist** 中配置。
- 默认白名单中没有包含**mapred.max.split.size**参数，所以运行的时候会提示不允许。

处理步骤

步骤1 搜索**hive.security.authorization.sqlstd.confwhitelist.append**，把**mapred.max.split.size**加进**hive.security.authorization.sqlstd.confwhitelist.append**中，详细信息可参考“组件操作指南 > > 使用Hive > 从零开始使用Hive”。

步骤2 修改完成后，保存配置，重启Hive组件。

步骤3 执行**set mapred.max.split.size=1000000;**，系统不再报错，则表示修改成功。

----结束

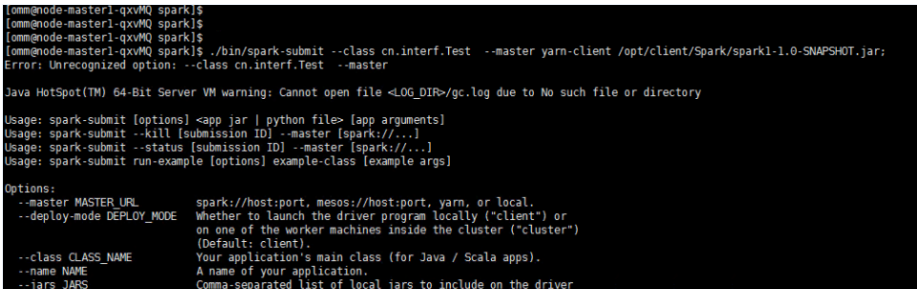
16.16.2 使用 Spark 时报错

用户问题

在使用spark时，集群运行失败。

问题现象

客户在使用spark组件时，集群运行失败。



原因分析

- 执行命令时，引入非法字符
- 上传的jar包属主属组有问题

处理步骤

步骤1 检查用户命令**./bin/spark-submit --class cn.interf.Test --master yarn-client /opt/client/Spark/spark1-1.0-SNAPSHOT.jar;**，排查是否引入非法字符。

步骤2 如果是，修改非法字符，重新执行命令。

步骤3 重新执行命令后，发生其他错误，查看该jar包的属主属组，发现全为root。

步骤4 修改jar包的属主属组为 omm:wheel，重新执行成功。

----结束

16.16.3 引入 jar 包不正确，导致 Spark 任务无法运行

用户问题

执行Spark任务，任务无法运行。

问题现象

执行Spark任务，任务无法运行。

原因分析

执行Spark任务时，引入的jar包不正确，导致Spark任务运行失败。

处理步骤

步骤1 登录任意Master节点。

步骤2 执行**cd /opt/Bigdata/MRS_*/install/FusionInsight-Spark-*/spark/examples/jars**命令， 查看样例程序的jar包。

说明

jar包名最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。

步骤3 检查OBS桶上的执行程序，执行程序可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。

说明

- OBS存储路径：以“obs://”开头。示例：obs://wordcount/program/hadoop-mapreduce-examples-2.7.x.jar
- HDFS存储路径：以“/user”开头。Spark Script需要以“.sql”结尾，MR和Spark需要以“.jar”结尾。sql、jar不区分大小写。

----结束

16.16.4 Spark 任务由于内存不够，作业卡住

用户问题

Spark提交作业内存不足导致任务长时间处于pending状态或者运行中内存溢出。

问题现象

使用Spark提交作业后，长期卡住不动。反复运行作业后报错，内容如下：

```
Exception in thread "main" org.apache.spark.SparkException: Job aborted due to stage failure:
Aborting TaskSet 3.0 because task 0 (partition 0) cannot run anywhere due to node and executor blacklist.
Blacklisting behavior can be configured via spark.blacklist.*.
```

原因分析

内存不足导致Spark提交的作业任务长时间处于pending状态。

处理步骤

步骤1 登录MRS Console页面，在现有集群中，选择集群名称，在“节点信息”页面，查看当前集群的节点规格。

步骤2 提高nodemanager进程所持有的集群资源。

MRS Manager界面操作：

1. 登录MRS Manager页面，选择“服务管理 > Yarn > 服务配置”。
2. 在“参数类别”中选择“全部配置”，然后在搜索框中搜索**yarn.nodemanager.resource.memory-mb**，查看该参数值。建议配置成节点物理内存总量的75%-90%。

FusionInsight Manager界面操作：

1. 登录FusionInsight Manager。选择“集群 > 服务 > Yarn”。
2. 单击“配置”，选择“全部配置”。然后在搜索框中搜索**yarn.nodemanager.resource.memory-mb**，查看该参数值。建议配置成节点物理内存总量的75%-90%。

步骤3 修改Spark的服务配置。

MRS Manager界面操作：

1. 登录MRS Manager页面，选择“服务管理”>“Spark” >“服务配置”。
2. 在“参数类别”中选择“全部配置”，然后在搜索框中搜索**spark.driver.memory**和**spark.executor.memory**

   根据作业的需要调大或者调小该值，具体以提交的Spark作业的复杂度和内存需要为参考（一般调大）。

FusionInsight Manager界面操作：

1. 登录FusionInsight Manager。选择“集群 > 服务 > Spark”。
2. 单击“配置”，选择“全部配置”。然后在搜索框中搜索**spark.driver.memory**和**spark.executor.memory**，根据作业的需要调大或者调小该值，具体以提交的Spark作业的复杂度和内存需要为参考（一般调大）。

说明

- 如果使用到SparkJDBC作业，搜索并修改**SPARK_EXECUTOR_MEMORY**和**SPARK_DRIVER_MEMORY**两个参数取值，具体以提交的Spark作业的复杂度和内存需要为参考（一般调大）。
- 如果对核数有要求，可以搜索并修改**spark.driver.cores**和**spark.executor.cores**的核数取值。

步骤4 Spark依赖内存做计算，如果以上还是不能满足任务的提交需要，建议扩容集群。

----结束

16.16.5 运行 Spark 报错

用户问题

运行Spark作业报找不到指定的类。

问题现象

运行Spark作业报找不到指定的类。报错内容如下：

```
Exception encountered | org.apache.spark.internal.Logging$class.logError(Logging.scala:91)
org.apache.hadoop.hbase.DoNotRetryIOException: java.lang.ClassNotFoundException:
org.apache.phoenix.filter.SingleCQKeyValueComparisonFilter
```

原因分析

用户配置的默认路径不正确。

处理步骤

步骤1 登录任意Master节点。

步骤2 修改Spark客户端目录下的配置文件。

执行**vim /opt/client/Spark/spark/conf/spark-defaults.conf**命令，打开spark-defaults.conf文件，设置“spark.executor.extraClassPath” 取值为“${PWD}/*”。

----结束

16.16.6 Driver 端提示 executor memory 超限

问题背景与现象

内存超限导致提交Spark任务失败。

原因分析

在Driver日志中直接打印申请的executor memory超过集群限制。

```
16/02/06 14:11:25 INFO Client: Verifying our application has not requested more than the maximum
memory capability of the cluster (6144 MB per container)
16/02/06 14:11:29 ERROR SparkContext: Error initializing SparkContext.
java.lang.IllegalArgumentException: Required executor memory (10240+1024 MB) is above the max
threshold (6144 MB) of this cluster!
```

Spark任务提交至Yarn上面，运行task的executor使用的资源受yarn的管理。从报错信息可看出，用户申请启动executor时，指定10G的内存，超出了Yarn设置的每个container的最大内存的限制，导致任务无法启动。

解决办法

修改Yarn的配置，提高对container的限制。如可通过调整“yarn.scheduler.maximum-allocation-mb”参数的大小，可控制启动的executor的资源，修改之后要重启Yarn服务。

配置修改方法：

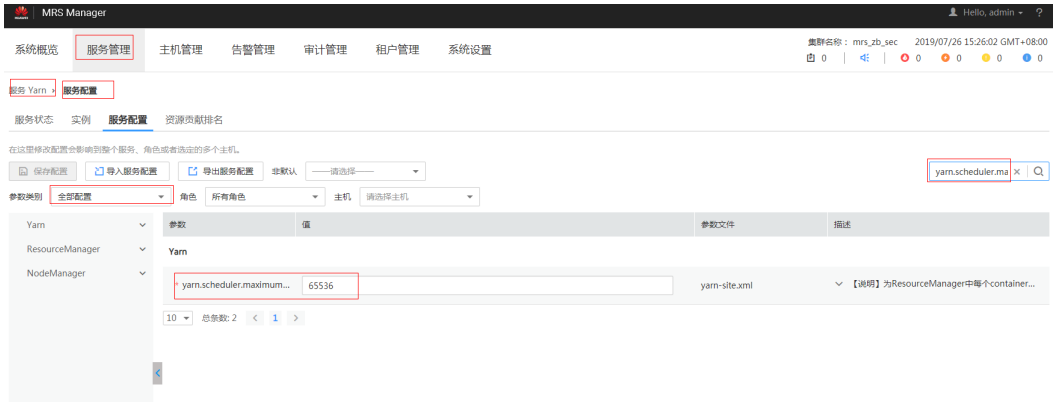
MRS Manager界面操作：

步骤1 登录MRS Manager页面。

步骤2 选择“服务管理 > Yarn > 服务配置”将“参数类别”修改为“全部配置”。

步骤3 在“搜索”栏输入“yarn.scheduler.maximum-allocation-mb”修改参数并保存重启服务。如下图所示：

图 16-56 修改 Yarn 服务参数



----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager页面。

步骤2 选择“集群 > 服务 > Yarn”，单击“配置”，选择“全部配置”。

步骤3 在“搜索”栏输入“yarn.scheduler.maximum-allocation-mb”修改参数并保存重启服务。

----结束

16.16.7 Yarn-cluster 模式下，Can't get the Kerberos realm 异常

问题背景与现象

认证异常导致提交Spark任务失败。

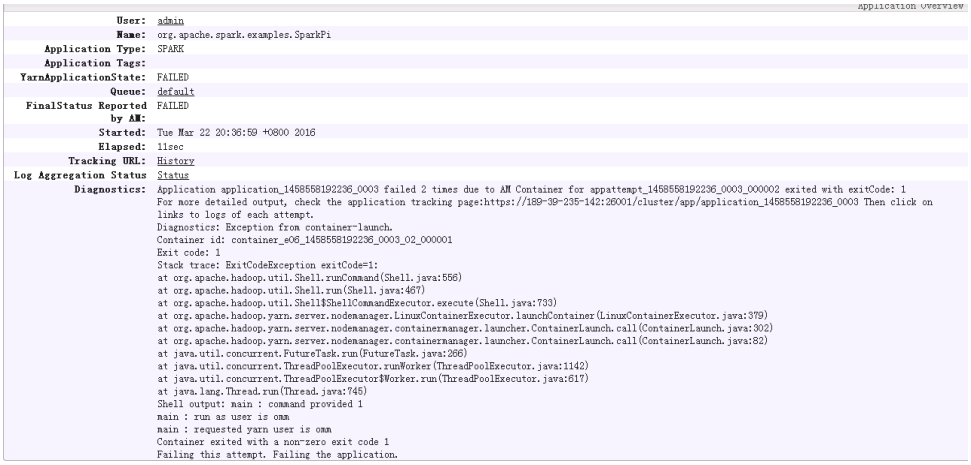
原因分析

1. 在driver端打印异常找不到连接hdfs的token，报错如下：

```
16/03/22 20:37:10 WARN Client: Exception encountered while connecting to the server :
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.token.SecretManager
$InvalidToken): token (HDFS_DELEGATION_TOKEN token 192 for admin) can't be found in cache
16/03/22 20:37:10 WARN Client: Failed to cleanup staging dir .sparkStaging/
application_1458558192236_0003
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.token.SecretManager
$InvalidToken): token (HDFS_DELEGATION_TOKEN token 192 for admin) can't be found in cache
```

2. 在Yarn原生页面显示am启动两次均失败，任务退出，如图16-57信息：

图 16-57 am 启动失败



3. 查看ApplicationMaster日志看到如下异常信息：

   ```
   Exception in thread "main" java.lang.ExceptionInInitializerError
   Caused by: org.apache.spark.SparkException: Unable to load YARN support
   Caused by: java.lang.IllegalArgumentException: Can't get Kerberos realm
   Caused by: java.lang.reflect.InvocationTargetException
   Caused by: KrbException: Cannot locate default realm
   Caused by: KrbException: Generic error (description in e-text) (60) - Unable to locate Kerberos realm
   org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1410)
   ... 86 more
   Caused by: javax.jdo.JDOFatalInternalException: Unexpected exception caught.
   NestedThrowables:java.lang.reflect.InvocationTargetException
   ... 110 more
   ```

4. 执行**./spark-submit --class yourclassname --master yarn-cluster /yourdependencyjars**任务以yarn-cluster模式提交任务，driver端会在集群中启用，由于加载的是客户端的spark.driver.extraJavaOptions，在集群节点上对应路径下找不到对应的kdc.conf文件，无法获取kerberos认证所需信息，导致am启动失败。

解决办法

在客户端提交任务时，在命令行中配置自定义的spark.driver.extraJavaOptions参数这样任务运行时就不会自动加载客户端路径下spark-defaults.conf中的spark.driver.extraJavaOptions；或者在启动spark任务时，通过--conf来指定driver的配置，如下（此处spark.driver.extraJavaOptions“=”号后面的引号部分不能缺少）。

./spark-submit -class yourclassname --master yarn-cluster --conf spark.driver.extraJavaOptions="

-Dlog4j.configuration=file:/opt/client/Spark/spark/conf/log4j.properties -Djetty.version=x.y.z -Dzookeeper.server.principal=zookeeper/hadoop.794bbab6_9505_44cc_8515_b4eddc84e6c1.com -Djava.security.krb5.conf=/opt/client/KrbClient/kerberos/var/krb5kdc/krb5.conf -Djava.security.auth.login.config=/opt/client/Spark/spark/conf/jaas.conf -Dorg.xerial.snappy.tempdir=/opt/client/Spark/tmp -Dcarbon.properties.filepath=/opt/client/Spark/spark/conf/carbon.properties" ../yourdependencyjars

16.16.8 JDK 版本不匹配启动 spark-sql，spark-shell 失败

问题背景与现象

JDK版本不匹配导致客户端启动spark-sql，spark-shell失败。

原因分析

1. 在Driver端打印异常如下：
   ```
   Exception Occurs: BadPadding 16/02/22 14:25:38 ERROR Schema: Failed initialising database. Unable to open a test connection to the given database. JDBC url = jdbc:postgresql://ip:port/sparkhivemeta, username = spark. Terminating connection pool (set lazyInit to true if you expect to start your database after your app).
   ```
2. Sparksql任务使用时，需要访问DBService以获取元数据信息，在客户端需要解密密文来访问，在使用过程中，用户没有按照流程操作，没有执行配置环境变量操作，且在其客户端环境变量中存在默认的jdk版本，导致在执行解密过程中调用的解密程序执行解密异常，会引起用户被锁。

解决办法

步骤1 使用**which java**命令查看默认的java命令是否是客户端的java。

步骤2 如果不是，请按正常的客户端执行流程。

source ${client_path}/bigdata_env

kinit *用户名*，然后输入用户名对应的密码，启动任务即可。

----结束

16.16.9 Yarn-client 模式提交 ApplicationMaster 尝试启动两次失败

问题背景与现象

Yarn-client模式提交任务AppMaster尝试启动两次失败。

原因分析

1. Driver端异常:
   ```
   16/05/11 18:10:56 INFO Client:
   client token: N/A
   diagnostics: Application application_1462441251516_0024 failed 2 times due to AM Container for appattempt_1462441251516_0024_000002 exited with  exitCode: 10
   For more detailed output, check the application tracking page:https://hdnode5:26001/cluster/app/application_1462441251516_0024 Then click on links to logs of each attempt.
   Diagnostics: Exception from container-launch.
   Container id: container_1462441251516_0024_02_000001
   ```
2. 在ApplicationMaster日志中，异常如下:
   ```
   2016-05-12 10:21:23,715 | ERROR | [main] | Failed to connect to driver at 192.168.30.57:23867, retrying ... | org.apache.spark.Logging$class.logError(Logging.scala:75)
   2016-05-12 10:21:24,817 | ERROR | [main] | Failed to connect to driver at 192.168.30.57:23867, retrying ... | org.apache.spark.Logging$class.logError(Logging.scala:75)
   2016-05-12 10:21:24,918 | ERROR | [main] | Uncaught exception:  | org.apache.spark.Logging$class.logError(Logging.scala:96)
   org.apache.spark.SparkException: Failed to connect to driver!
   at org.apache.spark.deploy.yarn.ApplicationMaster.waitForSparkDriver(ApplicationMaster.scala:426)
   at org.apache.spark.deploy.yarn.ApplicationMaster.runExecutorLauncher(ApplicationMaster.scala:292)
   ```

```
...
2016-05-12 10:21:24,925 | INFO  | [Thread-1] | Unregistering ApplicationMaster with FAILED (diag message: Uncaught exception: org.apache.spark.SparkException: Failed to connect to driver!) | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
```

Spark-client模式任务Driver运行在客户端节点上(通常是集群外的某个节点)，启动时先在集群中启动AppMaster进程，进程启动后要向Driver进程注册信息，注册成功后，任务才能继续。从AppMaster日志中可以看出，无法连接至Driver，所以任务失败。

解决办法

步骤1 请测试Driver进程所在的IP是否可以**ping**通。

步骤2 启动一个sparkpi任务，在console端会有类似如下打印信息。

```
16/05/11 18:07:20 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.1.100:23662]
16/05/11 18:07:20 INFO Utils: Successfully started service 'sparkDriver' on port 23662.
```

步骤3 在该节点，也就是**步骤2**中示例的192.168.1.100上执行**netstat - anp | grep 23662**看下此端口是否打开，如下打印标明，相关端口是打开的。

```
tcp        0      0  ip:port    :::*                    LISTEN      107274/java
tcp        0      0  ip:port    ip:port                 ESTABLISHED 107274/java
```

步骤4 在AppMaster启动的节点执行**telnet 192.168.1.100 23662**看下是否可以联通该端口，请使用**root**用户和**omm**用户都执行一遍。 如果出现**Escape character is '^]'**类似打印则说明可以联通，如果出现**connection refused**则表示失败，无法连接到相关端口。

如果相关端口打开，但是从别的节点无法联通到该端口，则需要排查下相关网络配置。

说明

23662这个端口每次都是随机的，所以要根据自己启动任务打开的端口来测试。

----结束

16.16.10 提交 Spark 任务时，连接 ResourceManager 异常

问题背景与现象

连接ResourceManager异常，导致Spark任务提交失败。

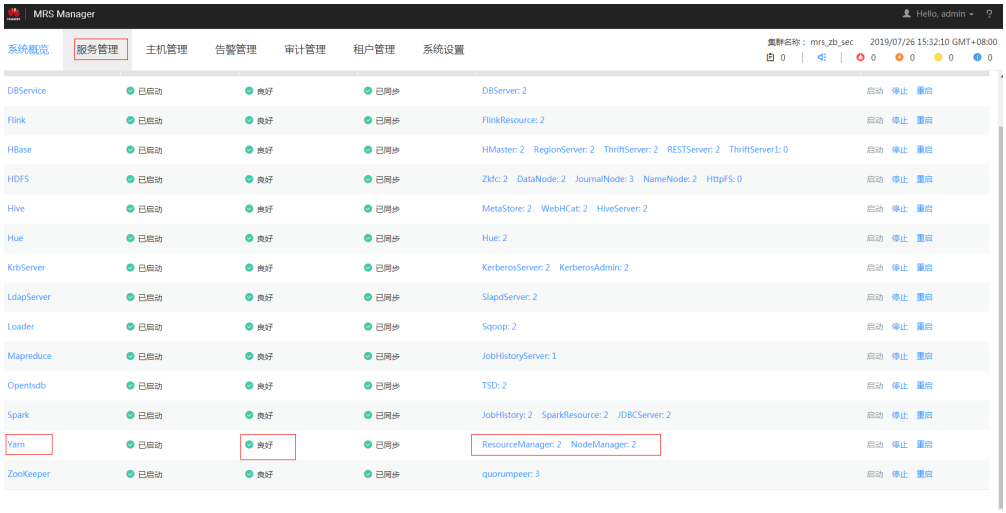
原因分析

1. 在driver端打印异常如下，打印连接两个ResourceManager主备节点的26004端口均被拒绝：

```
15/08/19 18:36:16 INFO RetryInvocationHandler: Exception while invoking getClusterMetrics of class ApplicationClientProtocolPBClientImpl over 33 after 1 fail over attempts. Trying to fail over after sleeping for 17448ms.
 java.net.ConnectException: Call From ip0 to ip1:26004 failed on connection exception: java.net.ConnectException: Connection refused.
INFO RetryInvocationHandler: Exception while invoking getClusterMetrics of class ApplicationClientProtocolPBClientImpl over 32 after 2 fail over attempts. Trying to fail over after sleeping for 16233ms.
 java.net.ConnectException: Call From ip0 to ip2:26004 failed on connection exception: java.net.ConnectException: Connection refused;
```

2. 在MRS Manager页面查看ResourceManager此时是否功能正常，如图16-58所示，如果Yarn状态故障或某个yarn服务的实例出现未知之类的异常说明此时集群的RM可能异常。

图 16-58 服务状态



3. 排查使用的客户端是否是集群最新的客户端。

   排查集群是否做过实例RM迁移相关操作（先卸载某个RM实例，然后在其他节点添加回来）。

4. 在MRS Manager页面单击“审计管理”，查看审计日志，是否有相关操作的记录。

   使用**ping**命令，查看IP是否可联通。

解决办法

- 如果RM出现异常，可参考Yarn相关章节查看解决方法。
- 如果客户端不是最新，请重新下载客户端。
- 若使用**ping**命令查看IP不通，需要协调网络管理相关人员协助排查网络。

16.16.11 DataArts Studio 调度 spark 作业失败

用户问题

DataArts Studio作业调度失败，显示读取/thriftserver/active_thriftserver路径下的数据失败。

问题现象

DataArts Studio作业调度失败，显示读取/thriftserver/active_thriftserver路径下的数据失败，

报错信息为：Can not get JDBC Connection, due to KeeperErrorCode = NoNode for /thriftserver/active_thriftserver。

原因分析

DataArts Studio提交spark作业时调用spark的JDBC方式，而Spark会启动一个名为thriftserver的进程以供客户端提供JDBC连接，JDBCServer在启动时会在zk的/

thriftserver目录下创建子目录active_thriftserver，并且注册相关连接信息。如果读不到该连接信息就会JDBC连接异常。

处理步骤

检查zookeeper下面是否有目标目录和注册的信息

步骤1 以root用户登录任意一个Master节点并初始化环境变量。

source /opt/client/bigdata_env

步骤2 执行**zkCli.sh -server 'ZookeeperIp:2181'**命令登录zk。

步骤3 执行**ls /thriftserver**查看是否有active_thriftserver目录。

- 如果有active_thriftserver目录，执行**get /thriftserver/active_thriftserver**查看该目录下是否有注册的配置信息。
  - 如果有注册的配置信息，联系支持人员处理。
  - 如果没有注册的配置信息，执行**步骤4**
- 如果没有active_thriftserver目录，执行**步骤4**。

步骤4 登录Manager界面，查看Spark的JDBCServer实例的主备状态是否未知。

- 是，执行**步骤5**。
- 否，联系运维人员处理。

步骤5 重启两个JDBCServer实例，查看主备实例状态恢复正常且zk下面有了目标目录和数据，作业即可恢复正常。若实例状态没有恢复请联系支持人员处理。

----结束

16.16.12 Spark 作业 api 提交状态为 error

用户问题

使用API提交spark作业后，作业状态显示为error。

问题类型

作业管理类。

问题现象

修改/opt/client/Spark/spark/conf/log4j.properties中的日志级别，使用API V1.1接口作业提交后，状态显示为error。

原因分析

executor会监控作业日志回显，确定作业执行结果，改为error后，检测不到输出结果，因此过期后判断作业状态为异常。

处理步骤

将/opt/client/Spark/spark/conf/log4j.properties中的日志级别修改为**info**。

建议与总结

建议客户使用V2接口提交作业接口。

16.16.13 集群反复出现 43006 告警

用户问题

集群反复出现“ALM-43006 JobHistory进程堆内存使用超出阈值”告警，且按照告警参考设置无效。

问题现象

集群出现告警“ALM-43006 JobHistory进程堆内存使用超出阈值”并且按照指导设置以后，运行一段时间又会出现同样的告警。

原因分析

可能存在JobHistory内存泄露问题，需要安装相应的补丁修复。

处理步骤

- 适当调大JobHistory进程堆内存。
- 如果已经调大堆内存，可以通过重启JobHistory实例规避。

16.16.14 在 spark-beeline 中创建/删除表失败

用户问题

客户在spark-beeline频繁创建和删除大量用户的场景下，个别用户偶现创建/删除表失败。

问题现象

创建表过程：

```
CREATE TABLE wlg_test001 (start_time STRING,value INT);
```

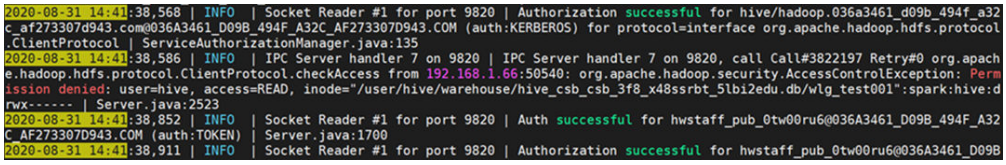
报错:

```
Error: org.apache.spark.sql.AnalysisException:
org.apache.hadoop.hive.ql.metadata.HiveException: MetaException(message:Failed to grant permission on
HDFSjava.lang.reflect.UndeclaredThrowableException); (state=,code=0)
```

原因分析

1. 查看metastore日志

2. 查看hdfs日志



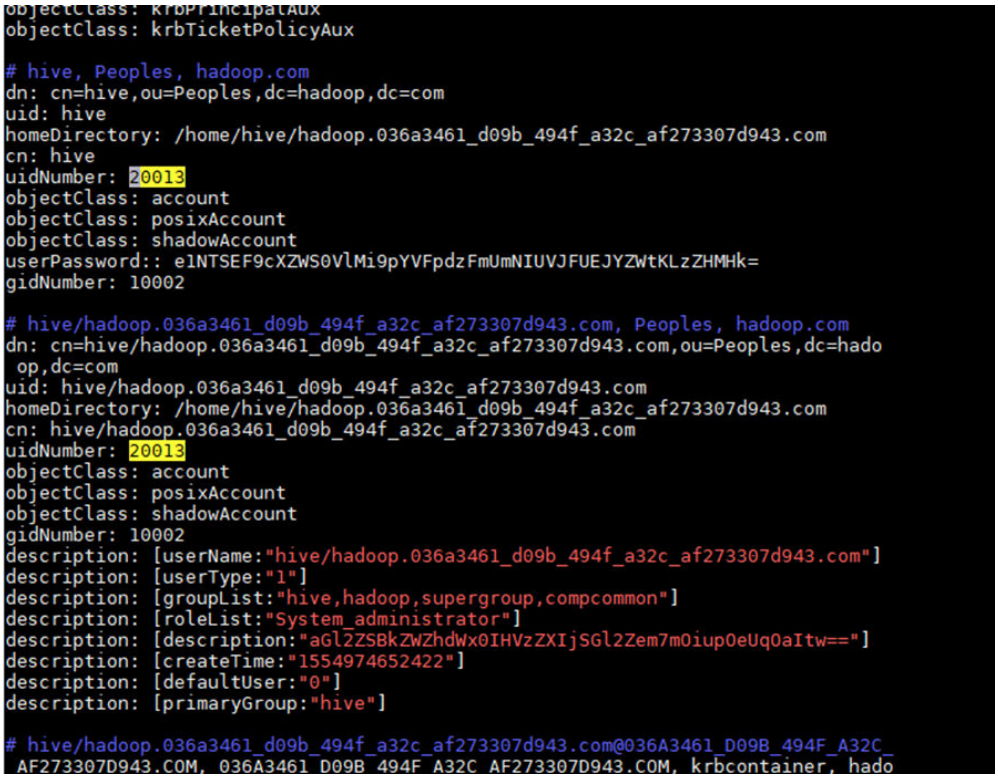
3. 权限对比（test001为异常用户创建表，test002为正常用户创建表）

4. drop表时报类似下面的错

```
0: jdbc:hive2://192.168.1.42:10000/> drop table
dataplan_modela_csbch2;
Error: Error while compiling statement: FAILED:
SemanticException Unable to fetch table dataplan_modela_csbch2.
java.security.AccessControlException: Permission denied: user=CSB_csb_3f8_x48ssrbt,
access=READ,
inode="/user/hive/warehouse/hive_csb_csb_3f8_x48ssrbt_5lbi2edu.db/
dataplan_modela_csbch2":spark:hive:drwx------
```

5. 根因分析。

   创建集群时创建的默认用户使用了相同的uid，造成用户错乱。在大量创建用户的场景下，触发了该问题，导致在创建表时偶现hive用户没有权限。



处理步骤

重启集群sssd进程。

以root用户执行**service sssd restart**命令重启sssd服务，执行**ps -ef | grep sssd**命令，查看sssd进程是否正常。

正常状态为：存在/usr/sbin/sssd进程和三个子进程/usr/libexec/sssd/sssd_be、/usr/libexec/sssd/sssd_nss、/usr/libexec/sssd/sssd_pam。

16.16.15 集群外节点提交 Spark 作业到 Yarn 报错连不上 Driver

用户问题

在集群外节点使用client模式提交Spark任务到Yarn上，任务失败，报错为连不上Driver。

问题现象

集群外节点和集群各个节点网络已经互通，在集群外节点使用client模式提交Spark任务到Yarn上，任务失败，报错为连不上Driver。

原因分析

使用client模式提交Spark任务的时候，Spark的driver进程是在客户端这边，而后面的executor都需要和Driver进行交互来运行作业。

如果NodeManager连不上客户端所在的节点，就会报错：

处理步骤

在客户端的Spark配置中指定Driver的IP地址：

“<客户端安装位置>/Spark/spark/conf/spark-defaults.conf”中添加参数“spark.driver.host=driverIP”，重新运行Spark任务即可。

建议与总结

建议客户通过cluster模式提交作业。

16.16.16 运行 Spark 任务发现大量 shuffle 结果丢失

用户问题

Spark任务运行失败，查看任务日志发现大量打印shuffle文件丢失。

问题现象

Spark任务运行失败，查看任务日志发现大量打印shuffle文件丢失。

原因分析

spark运行的时候会将临时产生的shuffle文件放在executor的临时目录中，方便后面获取。

而当某个executor异常退出时，NodeManager会把这个executor所在的container临时目录删除，随后其他executor再来申请这个executor的shuffle结果就会报文件找不到。

因此，遇到这样的问题需要确认是否executor异常退出，可以根据spark任务页面的executors便签页查看是否有dead状态的executor，查看各个dead状态的executor日志，确认异常退出的原因（其中可能有部分executor退出原因就是因为shuffle文件找不到，需要找到最早异常退出的executor）。

常见的异常退出：

- executor发生OOM
- executor运行时出现多个task任务失败
- executor所在节点被清理

处理步骤

根据executor异常退出的实际原因调整或者修改任务参数或代码，重新运行Spark任务即可。

16.16.17 JDBCServer 长时间运行导致磁盘空间不足

用户问题

连接Spark的JDBCServer服务提交spark-sql任务到yarn集群上，在运行一段时间以后会出现Core节点的数据盘被占满的情况。

问题现象

客户连接Spark的JDBCServer服务提交spark-sql任务到yarn集群上，在运行一段时间以后会出现Core节点的数据盘被占满的情况。

后台查看磁盘使用情况，主要是JDBCServer服务的APP临时文件（shuffle生成的文件）太多，并且没有进行清理占用了大量内存。

原因分析

查询Core节点有大量文件的目录，发现大部分都是类似“blockmgr-033707b6-fbbb-45b4-8e3a-128c9bcfa4bf”的目录，里面存放了计算过程中产生的shuffle临时文件。

因为JDBCServer启动了Spark的动态资源分配功能，已经将shuffle托管给NodeManager，NodeManager只会按照APP的运行周期来管理这些文件，并不会关注单个executor所在的container是否存在。因此，只有在APP结束的时候才会清理这些临时文件。任务运行时间较长时导致临时文件过多占用了大量磁盘空间。

处理步骤

启动一个定时任务来清理超过一定时间的shuffle文件，例如每个整点清理超过6个小时的文件：

步骤1 创建脚本“clean_appcache.sh”，若存在多个数据盘，请根据实际情况修改BASE_LOC中data1的值。

- 安全集群

```
#!/bin/bash
BASE_LOC=/srv/BigData/hadoop/data1/nm/localdir/usercache/spark/appcache/application_*/blockmgr*
find $BASE_LOC/ -mmin +360 -exec rmdir {} \;
find $BASE_LOC/ -mmin +360 -exec rm {} \;
```

- 普通集群

```
#!/bin/bash
BASE_LOC=/srv/BigData/hadoop/data1/nm/localdir/usercache/omm/appcache/application_*/blockmgr*
find $BASE_LOC/ -mmin +360 -exec rmdir {} \;
find $BASE_LOC/ -mmin +360 -exec rm {} \;
```

步骤2 修改脚本权限。

chmod 755 clean_appcache.sh

步骤3 增加一个定时任务来启动清理脚本，脚本路径请根据实际脚本存放位置修改。

查看定时任务：crontab -l

编辑定时任务：crontab -e

```
0 * * * * sh /root/clean_appcache.sh > /dev/null 2>&1
```

----结束

16.16.18 spark-shell 执行 sql 跨文件系统 load 数据到 hive 表失败

用户问题

使用spark-shell命令执行sql或者spark-submit提交的spark任务里面有sql的load命令，并且原数据和目标表存储位置不是同一套文件系统，上述两种方式MapReduce任务启动时会报错。

原因分析

当使用load导入数据到hive表的时候，属于需要跨文件系统的情况（例如原数据在hdfs上，而hive表数据存放在obs上），并且文件长度大于阈值（默认32M），则会触发使用distcp的MapReduce任务来执行数据迁移操作。这个MapReduce任务配置直接从spark任务配置里面提取，但是spark任务的net.topology.node.switch.mapping.impl配置项不是hadoop的默认值，需要使用spark的jar包，因此MapReduce会报类找不到。

处理步骤

方案一：

如果文件较小，则可以将默认长度设置得大于文件最大长度，例如最大的文件是95M，则设置：

```
hive.exec.copyfile.maxsize=104857600
```

方案二：

如果确实文件较大，需要使用distcp任务来提高数据迁移效率，则可以在spark任务启动的时候增加设置参数：

```
--conf spark.hadoop.net.topology.node.switch.mapping.impl=org.apache.hadoop.net.ScriptBasedMapping
```

16.16.19 Spark 任务提交失败

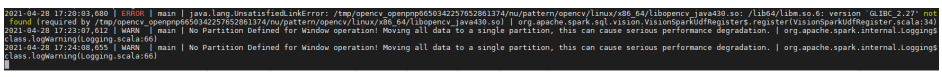
问题现象

- Spark提交任务直接提示无法提交任务。
- Spark提示无法获取到yarn的相关jar包。
- 提示多次提交一个文件。

原因分析

- 问题1：

  最常见的无法提交任务原因是认证失败，

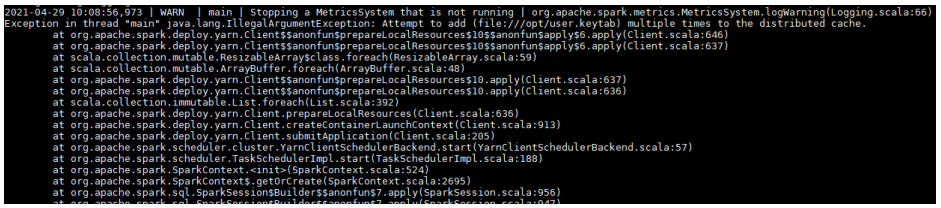
  还有可能是参数设置不正确。

- 问题2：

  集群默认会把分析节点的hadoop相关jar包添加到任务的classpath中，如果提示yarn的包找不到，一般都是因为hadoop的相关配置没有设置。

- 问题3：

  常见的场景是使用--files上传了user.keytab，然后使用--keytab又指定了同一个文件，导致一个文件多次被上传。

  
  ```
  2021-04-29 10:08:56,973 | WARN  | main | Stopping a MetricsSystem that is not running | org.apache.spark.metrics.MetricsSystem.logWarning(Logging.scala:66)
  Exception in thread "main" java.lang.IllegalArgumentException: Attempt to add (file:///opt/user.keytab) multiple times to the distributed cache.
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10$$anonfun$apply$6.apply(Client.scala:646)
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10$$anonfun$apply$6.apply(Client.scala:637)
          at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
          at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:48)
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10.apply(Client.scala:637)
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10.apply(Client.scala:636)
          at scala.collection.immutable.List.foreach(List.scala:392)
          at org.apache.spark.deploy.yarn.Client.prepareLocalResources(Client.scala:636)
          at org.apache.spark.deploy.yarn.Client.createContainerLaunchContext(Client.scala:913)
          at org.apache.spark.deploy.yarn.Client.submitApplication(Client.scala:205)
          at org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend.start(YarnClientSchedulerBackend.scala:57)
          at org.apache.spark.scheduler.TaskSchedulerImpl.start(TaskSchedulerImpl.scala:188)
          at org.apache.spark.SparkContext.<init>(SparkContext.scala:524)
          at org.apache.spark.SparkContext$.getOrCreate(SparkContext.scala:2695)
          at org.apache.spark.sql.SparkSession$Builder$$anonfun$7.apply(SparkSession.scala:956)
  ```

处理步骤

- 问题1：

  重新kinit一个用户并修改相应的配置参数。

- 问题2：

  查看hadoop相关的配置项是否正确，查看spark的conf目录下的core-site.xml，hdfs-site.xml，yarn-site.xml，mapred-site.xml等配置文件是否存在问题。

- 问题3：

  重新复制一个user.keytab，例如：

  cp user.keytab user2.keytab

  spark-submit --master yarn --files user.keytab --keytab user2.keytab .....

16.16.20 Spark 任务运行失败

问题现象

- 报错显示executor出现OOM
- 失败的task信息显示失败原因是lost task xxx

原因分析

- 问题1：一般出现executor OOM，都是因为数据量过大，也有可能是因为同一个executor上面同时运行的task太多。
- 问题2：有些task运行失败会报上述错误。当看到这个报错的时候，需要确认的是丢失的这个task在哪个节点上面运行，一般的情况是这个丢失的task异常退出导致的。

处理步骤

- 问题1：
  - 对于数据量过大，需要调整executor的内存大小的，使用--executor-memory指定内存大小；
  - 对于同时运行的task太多，主要看--executor-cores设置的vcore数量。
- 问题2：需要在相应的task的日志里面查找异常原因。如果有OOM的情况，请参照问题1。

16.16.21 JDBCServer 连接失败

问题现象

- 提示ha-cluster不识别（unknowHost或者必须加上端口）
- 提示连接JDBCServer失败

原因分析

- 问题1：使用**spark-beeline**命令连接JDBCServer，因为MRS_3.0以前的JDBCServer是ha模式，因此需要使用特定的url和MRS spark的自带的jar包来连接JDBCServer。
- 问题2：确认JDBCServer服务是否正常，查看对应的端口是否正常监听。

处理步骤

- 问题1：需要使用特定的url和MRS Spark的自带的jar包来连接JDBCServer。
- 问题2：确认JDBCServer服务是否正常，查看对应的端口是否正常监听。

16.16.22 查看 Spark 任务日志失败

问题现象

- 任务运行中查看日志失败
- 任务运行完成，但是查看不到日志

原因分析

- 问题1：可能原因是MapReduce服务异常
- 问题2：可能原因如下：
  - Spark的JobHistory服务异常。
  - 日志太大，NodeManager在做日志汇聚的时候出现超时。
  - HDFS存放日志目录权限异常（默认/tmp/logs/用户名/logs）。
  - 日志已被清理（spark的JobHistory默认存放7天的eventLog，配置项为spark.history.fs.cleaner.maxAge；MapReduce默认存放15天的任务日志，配置项为mapreduce.jobhistory.max-age-ms）。
  - 如果yarn页面上也找不到，可能是被yarn清理了（默认存放10000个历史任务，配置项为yarn.resourcemanager.max-completed-applications）。

处理步骤

- 问题1：确认MapReduce服务是否正常，如果异常，尝试重启服务。如果还是不能恢复，需要查看后台JobhistoryServer日志。
- 问题2：依次排查可能的情况：
  a. 查看Spark的JobHistory是否运行正常；
  b. 通过查看yarn的app详情页面，确认日志文件是否过大，如果日志汇聚失败，页面的“Log Aggregation Status:”应该显示为失败或者超时；
  c. 查看对应目录权限是否异常；

d. 查看目录下是否有对应的appid文件（spark的eventlog存放目录：MRS 3.x及以后版本的目录是hdfs://hacluster/spark2xJobHistory2x，MRS 3.x以前版本的目录是hdfs://hacluster/sparkJobHistory，任务运行日志存放目录是hdfs://hacluster/tmp/logs/用户名/logs）；

e. 查看appid和当前作业的id是否超过历史记录最大值。

16.16.23 Spark 连接其他服务认证问题

问题现象

- Spark连接HBase，报认证失败或者连接不到hbase表。
- Spark连接HBase报找不到jar包。

原因分析

- 问题1：HBase没有获取到当前任务的认证信息，导致连接HBase的时候认证失败，无法读取到相应数据
- 问题2：Spark默认没有加载HBase相关的jar包，需要使用--jars添加到任务中

处理步骤

- 问题1：可以尝试开启hbase认证开关：spark.yarn.security.credentials.hbase.enabled=true。但不建议直接用HBase客户端的hbase-site.xml替换Spark客户端下的hbase-site.xml，两者并不是完全相同。
- 问题2：需要将HBase相关的包使用--jars上传。

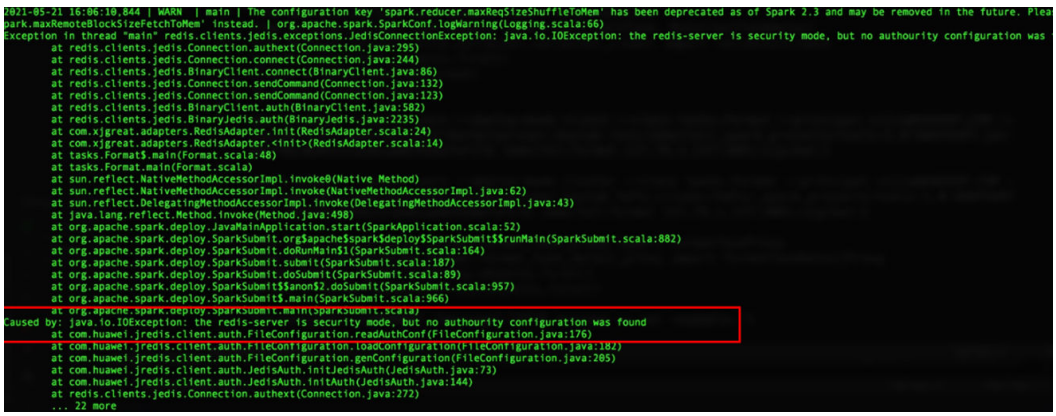
16.16.24 spark 连接 redis 报错

用户问题

使用MRS 3.x版本安全集群的spark组件访问redis报错。

问题现象

使用MRS_3.0版本安全集群的spark组件访问redis，会出现如下错误：



原因分析

Spark的jars目录下有一个MRS集群自带的jredisclient-xxx.jar包，客户使用spark任务连接redis的时候会因为加载了这个包从而出现该错误，需要手动去除redisclient包即可。

处理步骤

步骤1 清理Spark客户端下的jar包。

cd $SPARK_HOME/jars

mv jredisclient-*.jar /tmp

步骤2 清理Spark服务端下的jar包。

分别登录SparkResource2x所在的节点(一般有两个)。

mkdir /tmp/SparkResource2x

cd /opt/Bigdata/FusionInsight_Current/1_*_SparkResource2x/install/spark/jars/

mv jredisclient-*.jar /tmp/SparkResource2x

步骤3 清理HDFS上面的jredisclient文件。

1. 查看$SPARK_HOME/conf/spark-defaults.conf里面的配置项spark.yarn.archive，获取spark-archive-2x.zip包的地址。

   cat $SPARK_HOME/conf/spark-defaults.conf | grep "spark.yarn.archive"

2. 下载spark-archive-2x.zip包（本指导以MRS 3.0.5版本为例，具体命令请根据实际集群版本修改）。

   cd /opt

   mkdir sparkTmp

   cd sparkTmp

   hdfs dfs -get hdfs://hacluster/user/spark2x/jars/8.0.2.1/spark-archive-2x.zip

3. 解压spark-archive-2x.zip文件，并删除原文件。

   unzip spark-archive-2x.zip

   rm -f spark-archive-2x.zip

4. 移除jredisclient包。

   rm -f jredisclient-*.jar

5. 重新压缩spark-archive-2x.zip包。

   zip spark-archive-2x.zip ./*

6. 备份原有压缩包，上传新的压缩包。

   hdfs dfs -mv hdfs://hacluster/user/spark2x/jars/8.0.2.1/spark-archive-2x.zip /tmp

   hdfs dfs -put spark-archive-2x.zip hdfs://hacluster/user/spark2x/jars/8.0.2.1/spark-archive-2x.zip

7. 新的spark-archive-2x.zip中已删除jredisclient文件，需要重启JDBCServer服务，防止JDBCServer服务异常。

8. 清除临时文件。

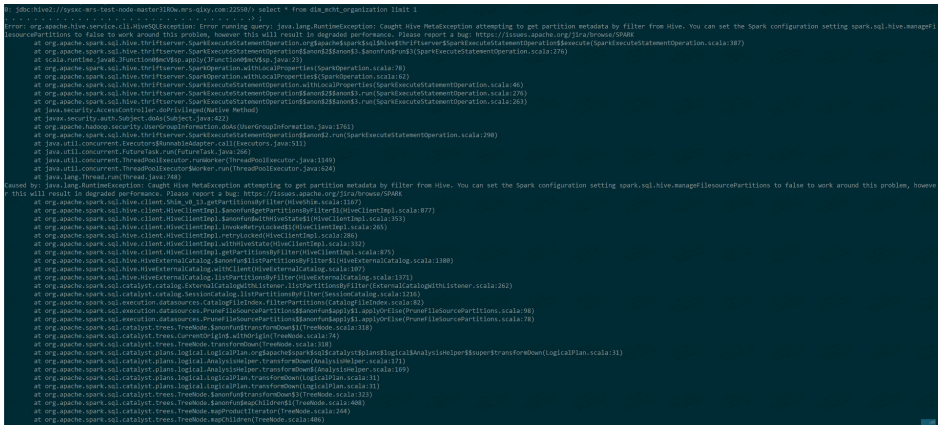
   rm -rf /opt/sparkTmp

----结束

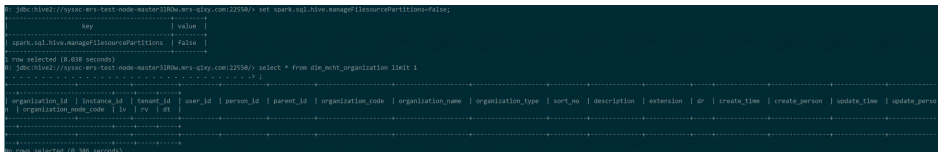
16.16.25 spark-beeline 查询 Hive 视图报错

用户问题

MRS 3.1.2版本，spark-beeline查询Hive视图报错，报错如下：



根据提示set spark.sql.hive.manageFilesourcePartitions=false后，查不到数据（但是实际在hive查询中是有数据的）。



原因分析

转换格式失败，spark-sql使用其内置的Metastore，而不是Hive中使用的Metastore，所以读取元数据时失败，spark-sql内部读取的是partque格式的数据，而Hive中默认是orc的。

处理步骤

步骤1 登录Spark客户端节点，执行如下命令，进入spark-sql：

cd *{客户端安装目录}*

source bigdata_env

source Spark2x/component_env

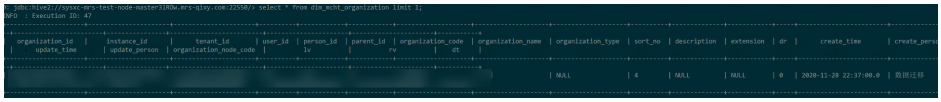
kinit *组件业务用户*（普通模式无需执行kinit）

spark-sql

步骤2 执行如下命令设置spark.sql.hive.convertMetastoreOrc=false。

set spark.sql.hive.convertMetastoreOrc=false;

步骤3 重新查询Hive视图，显示正常。



----结束

16.17 使用 Sqoop

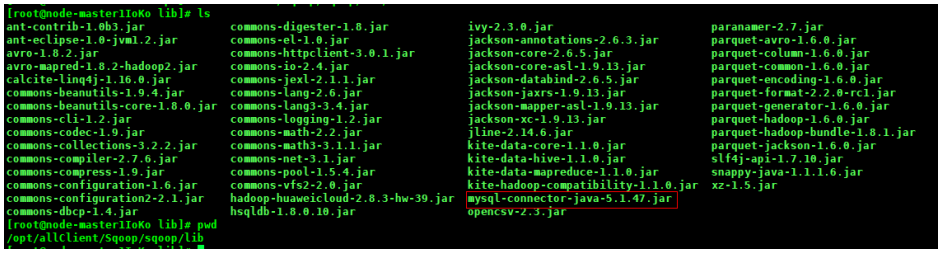
16.17.1 Sqoop 如何连接 mysql

用户问题

Sqoop如何连接mysql。

处理步骤

步骤1 在集群上安装客户端，查看客户端sqoop/lib下是否有mysql驱动包。



步骤2 在客户端目录下加载环境变量。

source bigdata_env

步骤3 Kerberos认证。

如果集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用跳过此步骤。

kinit MRS集群用户

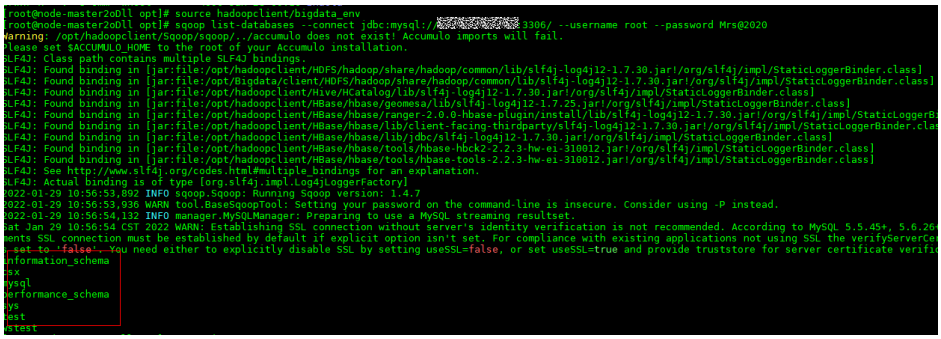
例如：

kinit admin

步骤4 连接数据库。

sqoop list-databases --connect jdbc:mysql://IP:3306/ --username 用户名 --password 密码

如下：



上图所示：则代表sqoop连接mysql成功。

----结束

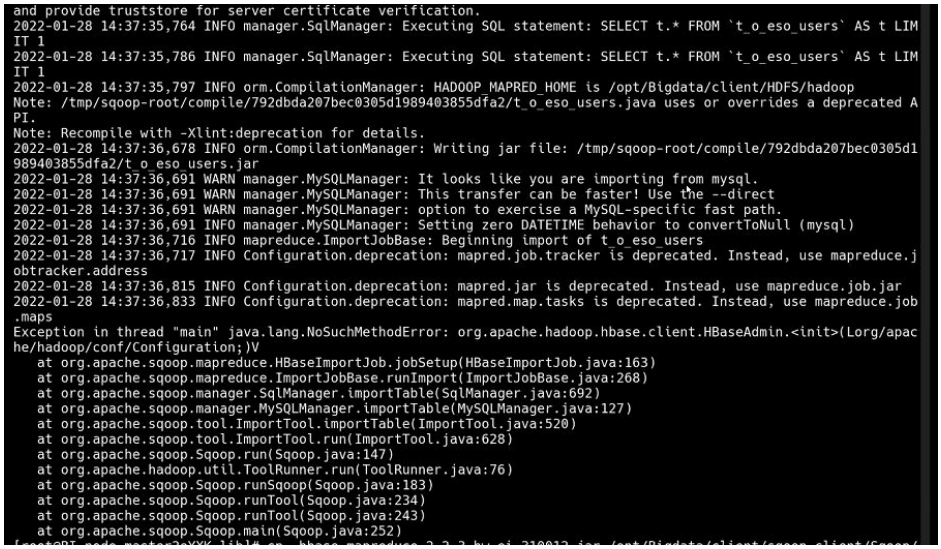
16.17.2 Sqoop 读取 MySQL 中数据到 HBase 报 HBaseAdmin.<init>方法找不到异常

问题

使用MRS的Sqoop客户端（1.4.7版本），从MySQL数据库中指定表抽取数据，存放到HBase（2.2.3版本）指定的表中，报出异常：

```
Trying to load data into HBASE through Sqoop getting below error.
Exception in thread "main" java.lang.NoSuchMethodError:
org.apache.hadoop.hbase.client.HBaseAdmin.<init>(Lorg/apache/hadoop/conf/Configuration;)V
```

完整异常信息如图所示：



执行Sqoop抽取数据命令样例：

```
sqoop import \
--connect jdbc:mysql://mysql服务器地址:端口号/database1 \
--username admin \
--password xxx \
--table table1 \
--hbase-table table2 \
--column-family info \
--hbase-row-key id \
--hbase-create-table --m 1
```

处理步骤

Sqoop客户端安装完成之后，没有直接引入HBase相关的依赖jar包，需要通过手动导入指定低版本的HBase相关依赖jar包。解决方法步骤如下：

步骤1 确认Sqoop客户端和HBase客户端是否在同一个路径下。

- 是，执行**步骤2**。
- 否，删除原有的Sqoop和HBase客户端文件，从FusionInsight Manager上下载完整的客户端安装在同一路径下。执行**步骤2**。

步骤2 以**root**用户登录Sqoop客户端安装节点。

步骤3 下载HBase 1.6.0版本的jar包上传到Sqoop客户端的“lib”目录下。

步骤4 上传包之后，修改包的权限，可以设置为755，具体执行命令为：

chmod 755 *包名称*

步骤5 在客户端目录下执行以下命令刷新Sqoop客户端：

source bigdata_env

重新执行sqoop命令

----结束

16.17.3 HUE 界面的 Sqoop 任务 HBase 到 HDFS 报错

本章节仅适用于MRS 1.9.2版本集群。

用户问题

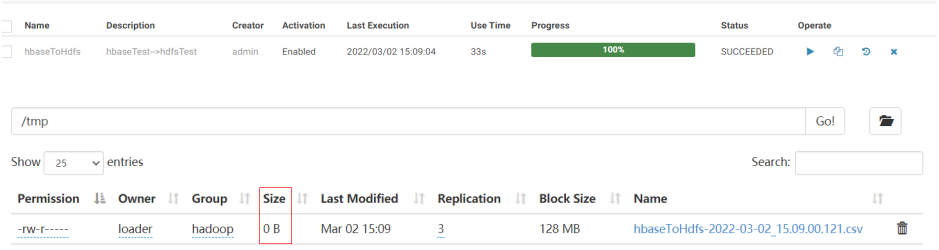
利用HUE的sqoop操作把HBase中的数据导入HDFS中报错。

Caused by: java.lang.ClassNotFoundException: org.apache.htrace.Trace

```
2022-03-02 15:09:00,264 [main] ERROR org.apache.sqoop.connector.hbase.HBaseExtractor  - An exceptional condition has occurred.
org.apache.sqoop.common.SqoopException: HBASE_CONNECTOR_0011:Failed to open table.
        at org.apache.sqoop.connector.hbase.HBaseExtractor.openDB(HBaseExtractor.java:239)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.access$100(HBaseExtractor.java:34)
        at org.apache.sqoop.connector.hbase.HBaseExtractor$1.run(HBaseExtractor.java:86)
        at org.apache.sqoop.connector.hbase.HBaseExtractor$1.run(HBaseExtractor.java:76)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.extract(HBaseExtractor.java:114)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.extract(HBaseExtractor.java:34)
        at org.apache.sqoop.job.mr.SqoopMapper.runInternal(SqoopMapper.java:156)
        at org.apache.sqoop.job.mr.SqoopMapper.run(SqoopMapper.java:79)
        at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:787)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:188)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1840)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:182)
Caused by: java.lang.reflect.InvocationTargetException
```

问题现象

Sqoop任务运行成功，但hdfs中的csv文件无内容。



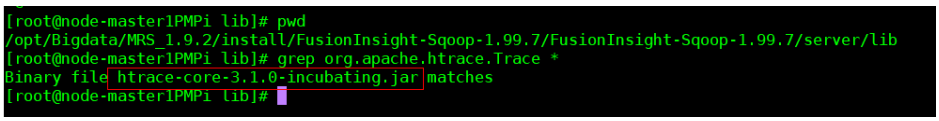
原因分析

推测jar包冲突或者缺少jar包造成的。

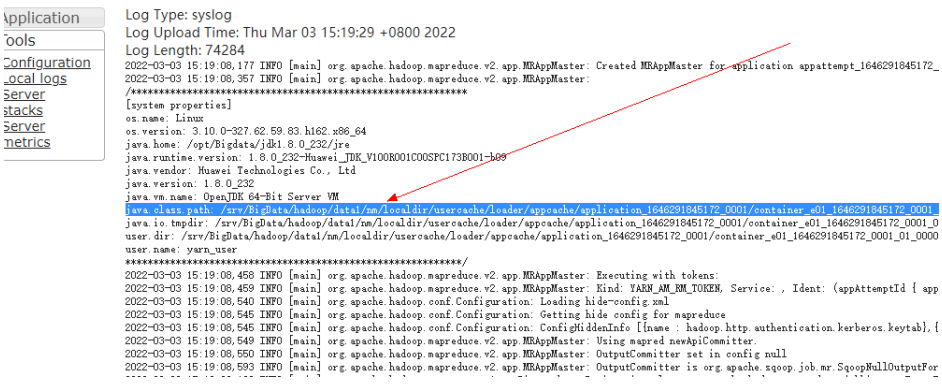
处理步骤

步骤1 去sqoop的lib下grep。

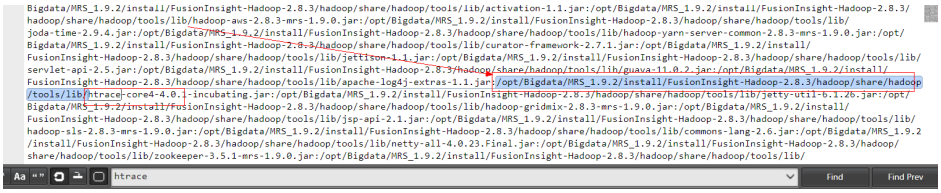
1. 进入sqoop的lib目录下，进行grep查找。



2. 进入yarn原生界面，查看运行的任务的报错具体信息。



3. 将java.class.path复制出来，搜索htrace-core。



4. 复制jar包到如下位置。

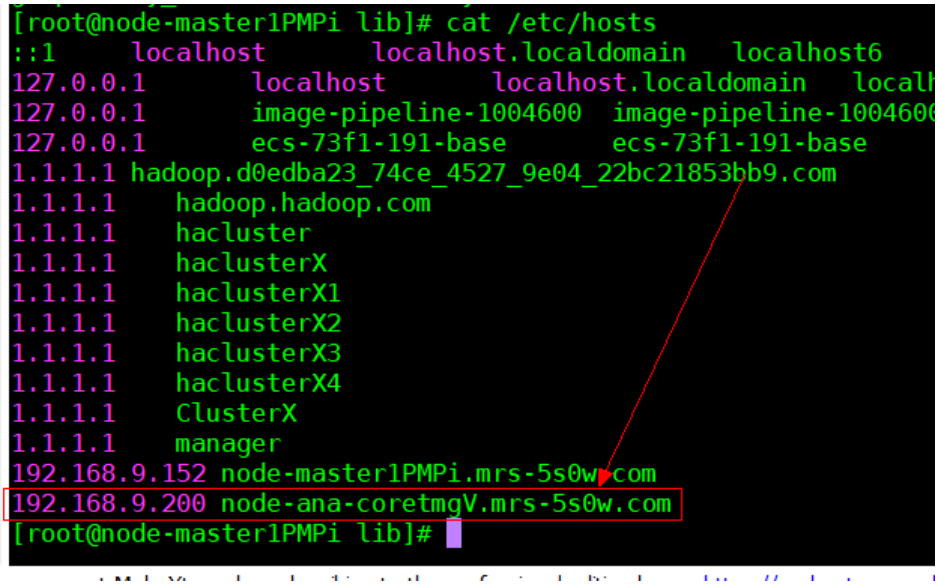
   cp /opt/Bigdata/MRS_1.9.2/install/FusionInsight-Sqoop-1.99.7/FusionInsight-Sqoop-1.99.7/server/lib/htrace-core-3.1.0-incubating.jar /opt/Bigdata/MRS_1.9.2/install/FusionInsight-Hadoop-2.8.3/hadoop/share/hadoop/common/lib/

5. 修改权限。

   chmod 777 *htrace-core-3.1.0-incubating.jar*（真实复制的jar包）

   chown omm:ficommon *htrace-core-3.1.0-incubating.jar*（真实复制的jar包）

6. 查看hosts文件，对其他所有节点进行同样的复制jar包操作。

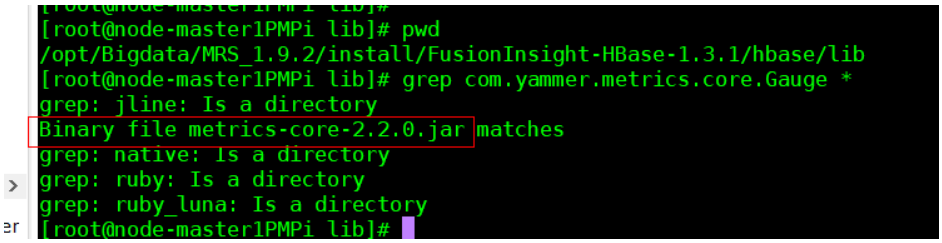


7. 重新运行sqoop任务，产生报错如下：

```
        at java.lang.Thread.run(Thread.java:745)
Caused by: com.google.protobuf.ServiceException: java.lang.NoClassDefFoundError: com/yammer/metrics/core/Gauge
        at org.apache.hadoop.hbase.ipc.AbstractRpcClient.callBlockingMethod(AbstractRpcClient.java:240)
        at org.apache.hadoop.hbase.ipc.AbstractRpcClient$BlockingRpcChannelImplementation.callBlockingMethod(AbstractRp
        at org.apache.hadoop.hbase.protobuf.generated.ClientProtos$ClientService$BlockingStub.scan(ClientProtos.java:3E
        at org.apache.hadoop.hbase.client.ClientSmallReversedScanner$SmallReversedScannerCallable.call(ClientSmallRever
        ... 9 more
Caused by: java.lang.NoClassDefFoundError: com/yammer/metrics/core/Gauge
        at org.apache.hadoop.hbase.ipc.AbstractRpcClient.callBlockingMethod(AbstractRpcClient.java:225)
        ... 12 more
Caused by: java.lang.ClassNotFoundException: com.yammer.metrics.core.Gauge
        at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:419)
        at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:352)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:352)
        ... 13 more
2022-03-03 15:45:01,714 [main] INFO  org.apache.sqoop.job.mr.SqoopMapper  - Extractor has finished
2022-03-03 15:45:01,715 [main] INFO  org.apache.sqoop.job.mr.SqoopMapper  - Stopping progress service
2022-03-03 15:45:01,727 [main] INFO  org.apache.sqoop.job.mr.SqoopOutputFormatLoadExecutor  - SqoopOutputFormatLoadExec
2022-03-03 15:45:01,776 [OutputFormatLoader-consumer] INFO  org.apache.sqoop.job.mr.SqoopOutputFormatLoadExecutor  - Lo
2022-03-03 15:45:01,777 [main] INFO  org.apache.sqoop.job.mr.SqoopOutputFormatLoadExecutor  - SqoopOutputFormatLoadExec

Log Type: stdout
Log Upload Time: Thu Mar 03 15:45:15 +0800 2022
Log Length: 0

Log Type: syslog
```

步骤2 去hbase的lib下grep。

1. 进入hbase的lib目录下，进行grep查找。



2. 继续复制jar包过去。

   cp /opt/Bigdata/MRS_1.9.2/install/FusionInsight-HBase-1.3.1/hbase/lib/metrics-core-2.2.0.jar /opt/Bigdata/MRS_1.9.2/install/FusionInsight-Hadoop-2.8.3/hadoop/share/hadoop/common/lib/

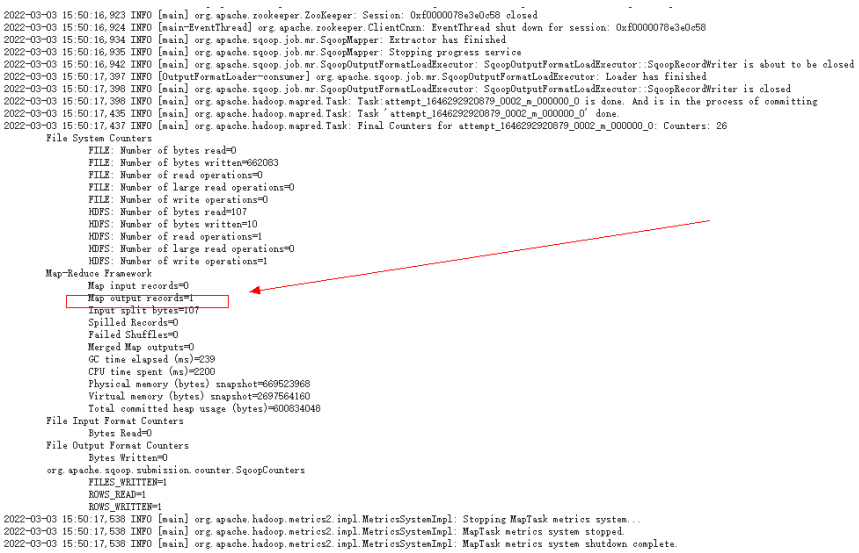
3. 修改文件权限。

   chmod 777 *metrics-core-2.2.0.jar*（真实复制的jar包）

   chown omm:ficommon *metrics-core-2.2.0.jar*（真实复制的jar包）

4. 查看hosts文件，对其他所有节点进行同样的复制jar包操作。

5. 继续运行sqoop任务，成功。



----结束

处理总结

1. 将sqoop的lib下htrace-core-3.1.0-incubating.jar和hbase的lib下的metrics-core-2.2.0.jar，复制到/opt/Bigdata/MRS_1.9.2/install/FusionInsight-Hadoop-2.8.3/hadoop/share/hadoop/common/lib/下。
2. 修改jar包的文件权限为777 和 omm:ficommon。
3. 所有节点均采取以上操作，重新运行sqoop任务即可。

16.17.4 Sqoop 从 hive 到 mysql8.0 报格式错误

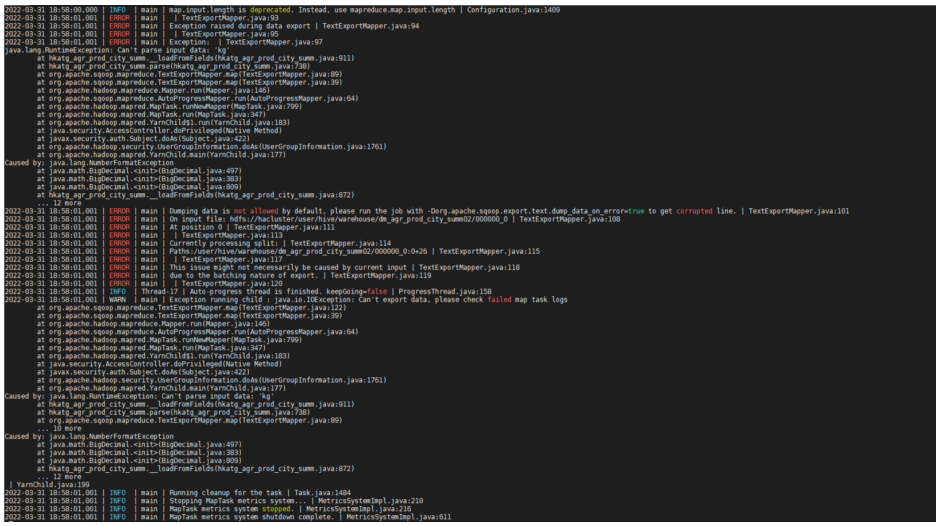
本章节仅适用于MRS 3.1.0版本集群。

用户问题

310集群Sqoop从hive到mysql8.0报格式错误。

问题现象

```
2022-03-31 19:56:44,581 ERROR mapreduce.ExportJobBase: Export job failed!
2022-03-31 19:56:44,581 ERROR tool.ExportTool: Error during export:
Export job failed!
        at org.apache.sqoop.mapreduce.ExportJobBase.runExport(ExportJobBase.java:445)
        at org.apache.sqoop.manager.SqlManager.exportTable(SqlManager.java:931)
        at org.apache.sqoop.tool.ExportTool.exportTable(ExportTool.java:80)
        at org.apache.sqoop.tool.ExportTool.run(ExportTool.java:99)
        at org.apache.sqoop.Sqoop.run(Sqoop.java:147)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:76)
        at org.apache.sqoop.Sqoop.runSqoop(Sqoop.java:183)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:234)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
        at org.apache.sqoop.Sqoop.main(Sqoop.java:252)
```



原因分析

通过日志判断得知是格式异常。

处理步骤

确认分隔符、表字段的格式无问题，在sqoop语句中添加--columns绑定对应字段。

sqoop export --connect jdbc:mysql://IP:端口号/数据库名 --username 用户名 --password 密码 --table 表名 --columns 列字段（多个列用英文逗号分开） -export-dir 导出地址 --fields-terminated-by 分隔符 --input-null-string '\\N' --input-null-non-string '\\N' -m 1

样例：

sqoop export --connect jdbc:mysql://172.16.0.6:3306/lidengpeng --username root --password Mrs@2021 --table hkatg_agr_prod_city_summ --columns year,city_name,city_code,prod_code,prod_name,prod_type,sown_area,area_unit,yield_wegt,yield_unit,total_wegt,total_wegt_unit,data_sorc_code,etl_time -export-dir hdfs://hacluster/user/hive/warehouse/dm_agr_prod_city_summ02 --fields-terminated-by ',' --input-null-string '\\N' --input-null-non-string '\\N' -m 1

16.17.5 Sqoop import 从 pg 到 hive 报错

背景

使用sqoop import命令抽取开源postgre到MRS hdfs或hive等。

用户问题

使用sqoop命令查询postgre表可以，但是执行sqoop import命令导入导出时报错：

The authentication type 5 is not supported. Check that you have configured the pg_hba.conf file to include the client's IP address or subnet。

原因分析

1. 连接postgresql MD5认证不通过，需要在pg_hba.cnf配置白名单。
2. 在执行sqoop import命令时，会启动MapReduce任务，由于MRS Hadoop安装目/opt/Bigdata/FusionInsight_HD_*/1_*_DataNode/install/hadoop/share/hadoop/common/lib下自带了postgre驱动包gsjdbc4-*.jar，与开源postgre服务不兼容导致报错。

处理步骤

1. 客户在pg_hba.cnf配置白名单。
2. 驱动重复，集群自带，将其余驱动排除出去，所有core节点上的gsjdbc4jar包去掉，在sqoop/lib下添加postgrejar包即可。

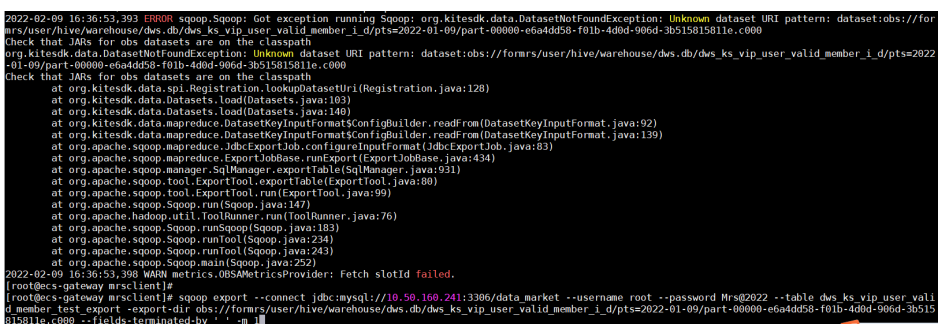
   mv /opt/Bigdata/FusionInsight_HD_*/1_*_DataNode/install/hadoop/share/hadoop/common/lib/gsjdbc4-*.jar /tmp

16.17.6 Sqoop 读 mysql，写 parquet 文件到 OBS 失败

用户问题

sqoop读mysql数据，然后直接写到obs，指定parquet格式时写入报错，不指定parquet时不报错。

问题现象



原因分析

parquet不支持hive3，用Hcatalog方式写入。

处理步骤

采用Hcatalog的方式：参数指定对应的hive库和表，需要修改SQL语句指定到具体字段（需要客户修改脚本）。

具体如下：

客户原来的脚本：

sqoop import --connect 'jdbc:mysql://10.160.5.65/xxx_pos_online_00?zeroDateTimeBehavior=convertToNull' --username root --password Mrs@2022

--split-by id

--num-mappers 2

--query 'select * from pos_remark where 1=1 and $CONDITIONS'

--target-dir obs://za-test/dev/xxx_pos_online_00/pos_remark

--delete-target-dir

--null-string '\\N‘

--null-non-string '\\N‘

--as-parquetfile

修改后的脚本（可以执行成功）：

sqoop import --connect 'jdbc:mysql://10.160.5.65/xxx_pos_online_00?zeroDateTimeBehavior=convertToNull' --username root --password Mrs@2022

--split-by id

--num-mappers 2

--query 'select id,pos_case_id,pos_transaction_id,remark,update_time,update_user,is_deleted,creator,modifier,gmt_created,gmt_modified,update_user_id,tenant_code from pos_remark where 1=1 and $CONDITIONS‘

--hcatalog-database xxx_dev

--hcatalog-table ods_pos_remark

16.18 使用 Storm

16.18.1 Storm 组件的 Storm UI 页面中 events 超链接地址无效

用户问题

Storm组件的Storm UI页面中events超链接地址无效。

问题现象

用户提交拓扑后无法查看拓扑数据处理日志，按钮events地址无效。

原因分析

MRS集群提交拓扑时默认不开启拓扑数据处理日志查看功能。

处理步骤

步骤1 登录Storm WebUI：

- MRS 2.x及之前版本：选择“Storm”，在“Storm 概述”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

  说明

  第一次访问Storm WebUI，需要在浏览器中添加站点信任以继续打开页面。

- MRS 3.x及后续版本：选择“Storm > 概览”，在“基本信息”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

步骤2 单击“Topology Summary”区域的指定拓扑名称，打开拓扑的详细信息。

步骤3 在“Topology actions”区域单击“Kill”删除已经提交的Storm拓扑。

步骤4 重新提交Storm拓扑，并开启查看拓扑数据处理日志功能，在提交Storm拓扑时增加参数“topology.eventlogger.executors”，该参数设置为一个不为0的正整数。例如：

storm jar *拓扑包路径 拓扑Main方法的类名称 拓扑名称* -c topology.eventlogger.executors=X

步骤5 在Storm UI界面，单击“Topology Summary”区域的指定拓扑名称，打开拓扑的详细信息。

步骤6 在“Topology actions”区域单击“Debug”，输入采样数据的百分比数值，并单击“OK”开始采样。

步骤7 单击拓扑的“Spouts”或“Bolts”任务名称，在“Component summary”单击“events”即可打开处理数据日志。

说明

如需开启特定“Spouts”或“Bolts”任务的拓扑数据处理日志查看功能，请单击拓扑的“Spouts”或“Bolts”任务名称后，“Topology actions”区域单击“Debug”按钮，输入采样数据的百分比数值。

----结束

16.18.2 提交拓扑失败

问题背景与现象

使用MRS流式集群，主要安装ZooKeeper、Storm、Kafka。

使用客户端命令，提交Topology失败。

可能原因

- Storm服务异常。
- 客户端用户没有进行安全认证或者认证过期。
- 提交拓扑中包含storm.yaml文件和服务端冲突。

原因分析

用户提交拓扑失败，可能原因客户端侧问题或者Storm侧问题。

1. 查看Storm状态。

   MRS Manager：

   登录MRS Manager，在MRS Manager页面，选择“服务管理 > Storm”，查看Storm服务当前状态，发现状态为“良好”，且监控指标内容显示正确。

   FusionInsight Manager界面操作：

   对于MRS 3.x及后续版本集群：登录FusionInsight Manager。选择“集群 > 服务 > Storm”，查看Storm服务当前状态，发现状态为“良好”，且监控指标内容显示正确。

2. 查看客户端提交日志，发现打印KeeperExceptionSessionExpireException异常信息，如下所示：

   ```
   org.apache.zookeeper.KeeperException$SessionExpiredException: KeeperErrorCode = Session expired
           at org.apache.zookeeper.KeeperException.create(KeeperException.java:131) ~[zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.curator.framework.imps.CuratorFrameworkImpl.checkBackgroundRetry(CuratorFrameworkImpl.java:710) [curator-framework-2.5.0.jar:na]
           at org.apache.curator.framework.imps.CuratorFrameworkImpl.processBackgroundOperation(CuratorFrameworkImpl.java:510) [curator-framework-2.5.0.jar:na]
           at org.apache.curator.framework.imps.BackgroundSyncImpl$1.processResult(BackgroundSyncImpl.java:50) [curator-framework-2.5.0.jar:na]
           at org.apache.zookeeper.ClientCnxn$EventThread.processEvent(ClientCnxn.java:484) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn$EventThread.queuePacket(ClientCnxn.java:498) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn.finishPacket(ClientCnxn.java:731) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn.conLossPacket(ClientCnxn.java:765) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn.access$2700(ClientCnxn.java:97) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn$SendThread.cleanup(ClientCnxn.java:1391) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn$SendThread.run(ClientCnxn.java:1314) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
   2016-05-31 09:25:24 | INFO  | [main] | Session: 0x10027394760ab4b closed | org.apache.zookeeper.ZooKeeper (ZooKeeper.java:968)
   Exception in thread "main" java.lang.RuntimeException: Exception while initializing NimbusLeaderElections
           at backtype.storm.nimbus.NimbusLeaderElections.init(NimbusLeaderElections.java:86)
           at backtype.storm.utils.NimbusClient.getConfiguredClient(NimbusClient.java:39)
           at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:189)
           at backtype.storm.StormSubmitter.submitTopologyWithProgressBar(StormSubmitter.java:256)
           at backtype.storm.StormSubmitter.submitTopologyWithProgressBar(StormSubmitter.java:236)
           at storm.starter.WordCountTopology.main(WordCountTopology.java:94)
   Caused by: org.apache.zookeeper.KeeperException$ConnectionLossException: KeeperErrorCode = ConnectionLoss for /storm/nimbus-leader
           at org.apache.zookeeper.KeeperException.create(KeeperException.java:99)
           at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
           at org.apache.zookeeper.ZooKeeper.exists(ZooKeeper.java:1501)
           at org.apache.curator.framework.imps.ExistsBuilderImpl$2.call(ExistsBuilderImpl.java:172)
           at org.apache.curator.framework.imps.ExistsBuilderImpl$2.call(ExistsBuilderImpl.java:161)
           at org.apache.curator.RetryLoop.callWithRetry(RetryLoop.java:107)
           at org.apache.curator.framework.imps.ExistsBuilderImpl.pathInForeground(ExistsBuilderImpl.java:157)
           at org.apache.curator.framework.imps.ExistsBuilderImpl.forPath(ExistsBuilderImpl.java:148)
           at org.apache.curator.framework.imps.ExistsBuilderImpl.forPath(ExistsBuilderImpl.java:36)
           at backtype.storm.nimbus.NimbusLeaderElections.init(NimbusLeaderElections.java:66)
           ... 5 more
   ```

   上述错误是由于在提交拓扑之前没有进行安全认证或者认证后TGT过期导致。

   解决方法参考**步骤1**。

3. 查看客户端提交日志，发现打印ExceptionInIntializerError异常信息，提示Found multiple storm.yaml resources。如下所示：

   ```
   Exception in thread "main" java.lang.ExceptionInInitializerError
           at backtype.storm.topology.TopologyBuilder.createTopology(TopologyBuilder.java:106)
           at com.huawei.streaming.storm.example.wordcount.WordCountTopology.cmdSubmit(WordCountTopology.java:117)
           at com.huawei.streaming.storm.example.wordcount.WordCountTopology.submitTopology(WordCountTopology.java:80)
           at com.huawei.streaming.storm.example.wordcount.WordCountTopology.main(WordCountTopology.java:71)
   Caused by: java.lang.RuntimeException: Found multiple storm.yaml resources. You're probably bundling the Storm jars with your topology jar.
           at backtype.storm.utils.Utils.findAndReadConfigFile(Utils.java:151)
           at backtype.storm.utils.Utils.readStormConfig(Utils.java:206)
           at backtype.storm.utils.Utils.<clinit>(Utils.java:70)
           ... 4 more
   ```

   该错误是由于业务jar包中存在storm.yaml文件，和服务端的storm.yaml文件冲突导致的。

   解决方法参考**步骤2**。

4. 如果不是上述原因，则请参考**提交拓扑失败，提示Failed to check principle for keytab**。

解决办法

步骤1 认证异常。

1. 登录客户端节点，进入客户端目录。
2. 执行以下命令重新提交任务。（业务jar包和Topology根据实际情况替换）

   source bigdata_env

   kinit *用户名*

   storm jar storm-starter-topologies-0.10.0.jar storm.starter.WordCountTopology test

步骤2 拓扑包异常。

排查业务jar，将业务jar中storm.yaml文件删除，重新提交任务。

----结束

16.18.3 提交拓扑失败，提示 Failed to check principle for keytab

问题背景与现象

使用MRS流式安全集群，主要安装ZooKeeper、Storm、Kafka等。

定义拓扑访问HDFS、HBase等组件，使用客户端命令，提交Topology失败。

可能原因

- 提交拓扑中没有包含用户的keytab文件。
- 提交拓扑中包含的keytab和提交用户不一致。
- 客户端/tmp目录下已存在user.keytab，且宿主非运行用户。

原因分析

1. 查看日志发现异常信息Can not found user.keytab in storm.jar。具体信息如下：

   ```
   [main] INFO  b.s.StormSubmitter - Get principle for stream@HADOOP.COM success
   [main] ERROR b.s.StormSubmitter - Can not found user.keytab in storm.jar.
   Exception in thread "main" java.lang.RuntimeException: Failed to check principle for keytab
   at backtype.storm.StormSubmitter.submitTopologyAs(StormSubmitter.java:219)
   at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:292)
   at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:176)
   at com.xxx.streaming.storm.example.hbase.SimpleHBaseTopology.main(SimpleHBaseTopology.java:77)
   ```

   查看提交的拓扑运行Jar，发现没有包含keytab文件。

2. 查看日志发现异常信息The submit user is invalid,the principle is 。具体信息如下：

   ```
   [main] INFO  b.s.StormSubmitter - Get principle for stream@HADOOP.COM success
   [main] WARN  b.s.s.a.k.ClientCallbackHandler - Could not login: the client is being asked for a password, but the  client code does not currently support obtaining a password from the user. Make sure that the client is configured to use a ticket cache (using the JAAS configuration setting 'useTicketCache=true)' and restart the client. If you still get this message after that, the TGT in the ticket cache has expired and must be manually refreshed. To do so, first determine if you are using a password or a keytab. If the former, run kinit in a Unix shell in the environment of the user who is running this client using the command 'kinit <princ>' (where <princ> is the name of the client's Kerberos principal). If the latter, do 'kinit -k -t <keytab> <princ>' (where <princ> is the name of the Kerberos principal, and <keytab> is the location of the keytab file). After manually refreshing your cache, restart this client. If you continue to see this message after manually refreshing your cache, ensure that your KDC host's clock is in sync with this host's clock.
   [main] ERROR b.s.StormSubmitter - The submit user is invalid,the principle is : stream@HADOOP.COM
   ```

```
Exception in thread "main" java.lang.RuntimeException: Failed to check principle for keytab
at backtype.storm.StormSubmitter.submitTopologyAs(StormSubmitter.java:219)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:292)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:176)
at com.xxx.streaming.storm.example.hbase.SimpleHBaseTopology.main(SimpleHBaseTopology.java:77)
```

业务提交拓扑时使用的认证用户为stream，但是在拓扑提交过程中提示submit user是无效用户，表明内部校验失败。

3. 查看提交的拓扑运行Jar，发现包含keytab文件。

   查看user.keytab文件，发现principal为zmk_kafka.

```
[root@8-5-148-6 client]# klist -kt user.keytab
Keytab name: FILE:user.keytab
KVNO Timestamp         Principal
---- ----------------- --------------------------------------------------------
   1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
   1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
```

   发现认证用户和user.keytab文件中principal不对应。

4. 查看日志发现异常信息Delete the tmp keytab file failed，the keytab file is ：/tmp/user.keytab，具体信息如下：

```
[main] WARN b.s.StormSubmitter - Delete the tmp keytab file failed, the keytab file is : /tmp/
user.keytab
[main] ERROR b.s.StormSubmitter - The submit user is invalid,the principle is : hbase1@HADOOP.COM
Exception in thread "main" java.lang.RuntimeException: Failed to check principle for keytab
at backtype.storm.StormSubmitter.submitTopologyAs(StormSubmitter.java:213)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:286)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:170)
at com.touchstone.storm.cmcc.CmccDataHbaseTopology.main(CmccDataHbaseTopology.java:183)
```

   查看系统/tmp目录，发现存在user.keytab文件，且文件宿主非运行用户。

解决办法

- 提交拓扑时携带用户user.keytab文件。
- 提交拓扑时的用户需要和user.keytab文件用户一致。
- 删除/tmp目录下不对应的user.keytab文件。

16.18.4 提交拓扑后 Worker 日志为空

现象描述

在Eclipse中远程提交拓扑成功之后，无法在Storm WebUI查看拓扑的详细信息，并且每个拓扑的Bolt和Spout所在Worker节点在一直变化。查看Worker日志，日志内容为空。

可能原因

Worker进程启动失败，触发Nimbus重新分配任务，在其他Supervisor上启动Worker。由于Worker启动失败后会继续重启，导致Worker节点在一直变化，且Worker日志内容为空。Worker进程启动失败的可能原因有两个：

- 提交的Jar包中包含“storm.yaml”文件。

  Storm规定，每个“classpath”中只能包含一个“storm.yaml”文件，如果多于一个那么就会产生异常。使用Storm客户端提交拓扑，由于客户端“classpath”配置和Eclipse远程提交方式“classpath”不一样，客户端会自动加载用户的Jar包到“classpath”，从而使“classpath”中存在两个“storm.yaml”文件。
- Worker进程初始化时间较长，超过Storm集群设置Worker启动超时时间，导致Worker被Kill从而一直进行重分配。

定位思路

1. 使用Storm客户端提交拓扑，检查出重复“storm.yaml”问题。
2. 重新打包Jar包，然后再提交拓扑。
3. 修改Storm集群关于Worker启动超时参数。

处理步骤

步骤1 使用Eclipse远程提交拓扑后Worker日志为空，则使用Storm客户端，提交拓扑对应的Jar包，查看提示信息。

例如，Jar包中包含两个不同路径下的“storm.yaml”文件，系统显示以下信息：

```
Exception in thread "main" java.lang.ExceptionInInitializerError
  at com.xxx.streaming.storm.example.WordCountTopology.createConf(WordCountTopology.java:132)
  at com.xxx.streaming.storm.example.WordCountTopology.remoteSubmit(WordCountTopology.java:120)
  at com.xxx.streaming.storm.example.WordCountTopology.main(WordCountTopology.java:101)
 Caused by: java.lang.RuntimeException: Found multiple storm.yaml resources. You're probably bundling the Storm jars with your topology jar. [jar:file:/opt/xxx/fi_client/Streaming/streaming-0.9.2/bin/stormDemo.jar!/storm.yaml, file:/opt/xxx/fi_client/Streaming/streaming-0.9.2/conf/storm.yaml]
  at backtype.storm.utils.Utils.findAndReadConfigFile(Utils.java:151)
  at backtype.storm.utils.Utils.readStormConfig(Utils.java:206)
  at backtype.storm.utils.Utils.<(Utils.java:70)>
```

步骤2 重新打包Jar包，且不能包含“storm.yaml”文件、“log4j”和“slf4j-log4j”相关的Jar包。

步骤3 使用IntelliJ IDEA远程提交新打包的Jar包。

步骤4 查看是否可以在WebUI查看拓扑的详细信息和Worker日志内容。

步骤5 在Manager页面修改Storm集群关于Worker启动超时参数（参数说明请参考**参考信息**），保存并重启Storm服务。

- MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Storm> 配置”。
- FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Storm > 配置"

步骤6 重新提交待运行的Jar包。

----结束

参考信息

1. nimbus.task.launch.secs和supervisor.worker.start.timeout.secs这两个参数分别代表nimbus端和supervisor端对于拓扑启动的超时容忍时间，一般nimbus.task.launch.secs的值要大于等于supervisor.worker.start.timeout.secs的值（建议相等或略大，如果超出太多会影响任务重分配的效率）。
   - nimbus.task.launch.secs：nimbus在超过该参数配置的时间内没有收到拓扑的task发的心跳时，会将该拓扑重新分配（分配给别的supervisor），同时会刷新zk中的任务信息，supervisor读到zk中的任务信息并且与自己当前所启动的拓扑进行比较，如果存在拓扑已经不属于自己，那么则会删除该拓扑的元数据，也就是/srv/Bigdata/streaming_data/stormdir/supervisor/stormdist/{worker-id}目录。
   - supervisor.worker.start.timeout.secs：supervisor启动worker后，在该参数配置的时间内没有收到worker的心跳时，supervisor会主动停掉worker，等

待worker的重新调度，一般在业务启动时间较长时适当增加该参数的值，保证worker能启动成功。

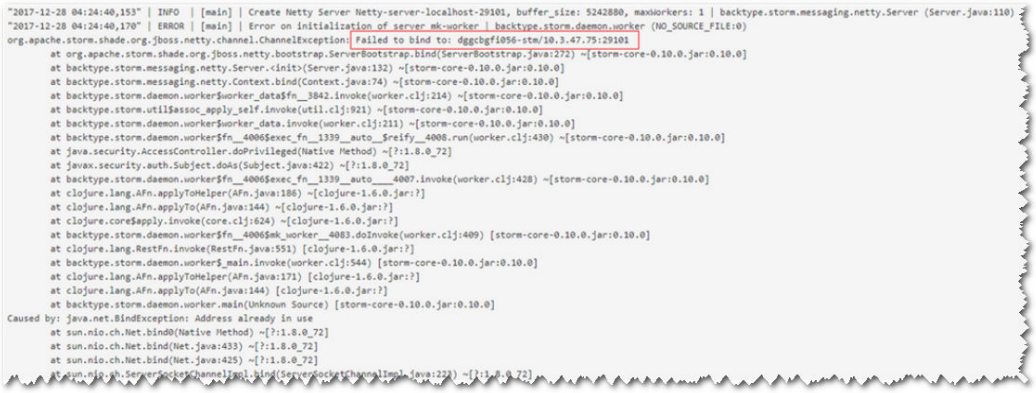
如果supervisor.worker.start.timeout.secs配置的值比nimbus.task.launch.secs的值大，那么则会出现supervisor的容忍时间没到，仍然继续让worker启动，而nimbus却认定该业务启动超时，将该业务分配给了其他主机，这时supervisor的后台线程发现任务不一致，删除了拓扑的元数据，导致接下来worker在启动过程中要读取stormconf.ser时，发现该文件已经不存在了，就会抛出FileNotFoundException。

2. nimbus.task.timeout.secs和supervisor.worker.timeout.secs这两个参数则分别代表nimbus端和supervisor端对于拓扑运行过程中心跳上报的超时容忍时间，一般nimbus.task.timeout.secs的值要大于等于supervisor.worker.timeout.secs的值（建议相等或略大），原理同上。

16.18.5 提交拓扑后 Worker 运行异常，日志提示 Failed to bind to：host:ip

现象描述

提交业务拓扑后，发现Worker无法正常启动。查看Worker日志，日志提示Failed to bind to：host:ip。



可能原因

随机端口范围配置错误。

定位思路

1、检查worker相关信息日志。

2、检查绑定端口的进程信息。

3、检查随机端口范围配置。

原因分析

1. 通过SSH登录Worker启动失败主机，通过**netstat -anp | grep <port>**命令，查看占用端口的进程ID信息。其中port修改为实际端口号。
2. 通过**ps -ef | grep <pid>**命令查看进程的详细信息，其中pid为查询出的实际进程ID。

发现占用端口的进程为worker进程，该进程为另一个拓扑业务进程。同时根据进程详细信息发现，分配给该进程的端口为29122。

3. 通过**lsof -i:<port>**命令，查看连接详细信息。其中port为实际端口号。

发现29101端口连接对端端口为21005，而21005为Kafka服务端端口。

说明业务层作为客户端连接Kafka获取消息，业务端口分配通过OS的随机端口分配范围来确定。

4. 通过**cat /proc/sys/net/ipv4/ip_local_port_range**命令查看随机端口范围。

5. 发现随机端口范围过大，和MRS的服务端口范围存在冲突。

说明

MRS的服务端口范围：20000-30000。

处理步骤

步骤1 修改随机端口范围。

```
vi /proc/sys/net/ipv4/ip_local_port_range
32768 61000
```

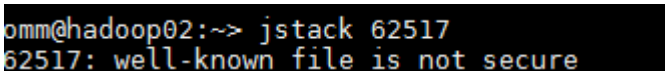
步骤2 停止占用服务端口的业务进程，释放端口。（停止业务拓扑）

----结束

16.18.6 使用 jstack 命令查看进程堆栈提示 well-known file is not secure

问题背景与现象

使用jstack命令查看进程堆栈信息报：well-known file is not secure。



原因分析

1. 由于执行命令的用户与当前查看pid信息的进程提交用户不一致导致。
2. Storm引入区分用户执行任务特性，在启动worker进程时将给进程的uid和gid改为提交用户和ficommon，目的是为了logviewer可以访问到worker进程的日志同时日志文件只开放权限到640。这样会导致切换到提交用户后对Worker进程执行jstack和jmap等命令执行失败，原因是提交用户的默认gid并不是ficommon，需要通过ldap命令修改提交用户的gid为9998（ficommon）才可执行。

解决办法

共有两种方式解决该问题。

方式一：通过storm原生页面查看进程堆栈

步骤1 登录Storm原生界面。

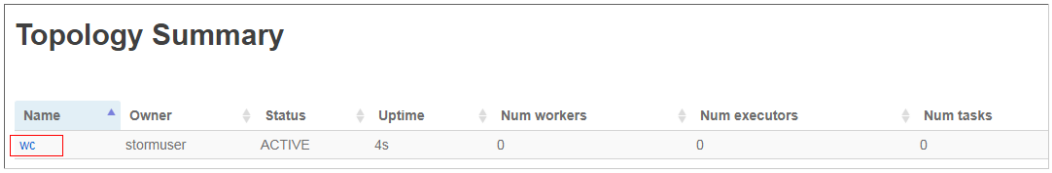
MRS Manager界面操作：

1. 访问MRS Manager。
2. 在Manager选择“服务管理 > Storm”，在“Storm 概述”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

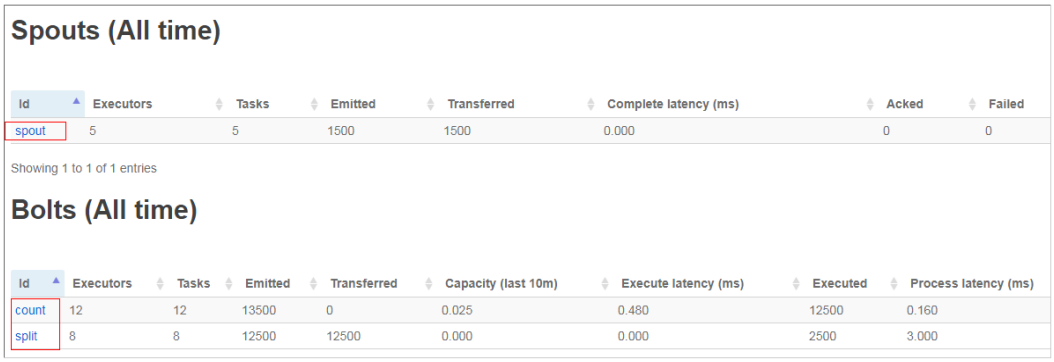
FusionInsight Manager界面操作：

1. 访问FusionInsight Manager。
2. 在Manager选择“集群 > 服务 > Storm”，在“概览”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

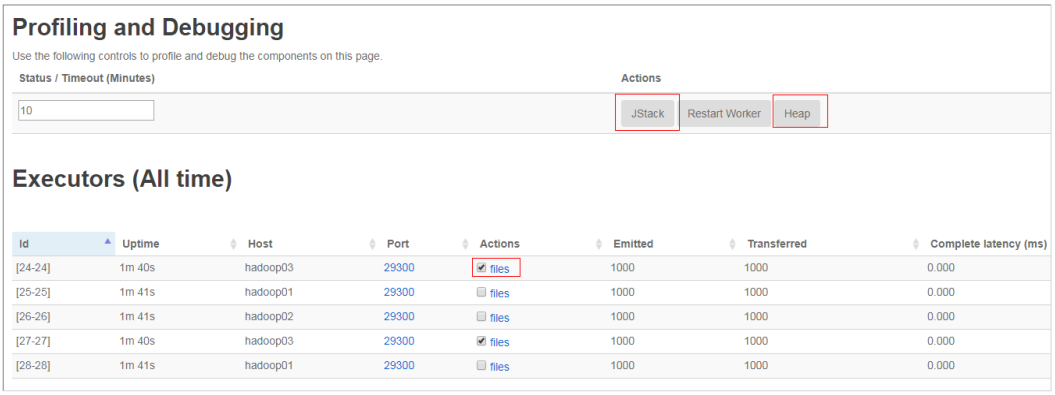
步骤2 选择要查看的拓扑。



步骤3 选择要查看的spout或者bolt。



步骤4 选择要查看的节点日志文件，再选择JStack或者Heap按钮，其中JStack对应的是堆栈信息，Heap对应的是堆信息：



----结束

方式二：通过修改自定义参数查看进程堆栈

步骤1 进入Storm服务参数配置界面。

MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Storm > 服务配置”，“参数类别”选择“全部配置”。

FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 服务 > Yarn”，单击“配置”，选择“全部配置”。

步骤2 在左侧导航栏选择“supervisor > 自定义”，添加一个变量supervisor.run.worker.as.user=false。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

步骤4 重新提交拓扑。

步骤5 后台节点切为omm用户执行jps命令即可查看worker的pid。

```
omm@hadoop02:~> jps | grep worker
22485 worker
111402 worker
```

步骤6 执行jstack pid，即可查看jstack信息。

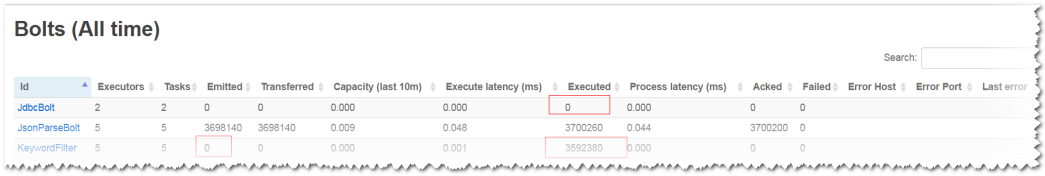
```
omm@hadoop02:~> jstack 22485
2018-05-26 08:46:24
Full thread dump Java HotSpot(TM) 64-Bit Server VM (25.144-b01 mixed mode):

"Attach Listener" #82 daemon prio=9 os_prio=0 tid=0x0000000001c95000 nid=0xb840 waiting on condition [0x0000000000000000]
   java.lang.Thread.State: RUNNABLE

"pool-14-thread-1" #81 daemon prio=5 os_prio=0 tid=0x00007f7ebc931000 nid=0x6113 waiting on condition [0x00007f7eb5ddf000]
   java.lang.Thread.State: TIMED_WAITING (parking)
        at sun.misc.Unsafe.park(Native Method)
        - parking to wait for  <0x00000000dfe820a0> (a java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject)
        at java.util.concurrent.locks.LockSupport.parkNanos(LockSupport.java:215)
        at java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject.awaitNanos(AbstractQueuedSynchronizer.java:2078)
        at java.util.concurrent.ScheduledThreadPoolExecutor$DelayedWorkQueue.take(ScheduledThreadPoolExecutor.java:1093)
        at java.util.concurrent.ScheduledThreadPoolExecutor$DelayedWorkQueue.take(ScheduledThreadPoolExecutor.java:809)
        at java.util.concurrent.ThreadPoolExecutor.getTask(ThreadPoolExecutor.java:1074)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1134)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
```

----结束

16.18.7 使用 Storm-JDBC 插件开发 Oracle 写入 Bolt，发现数据无法写入

现象描述

使用Storm-JDBC插件开发Oracle写入Bolt，发现能连上Oracle数据库，但是无法向Oracle数据库里面写数据。

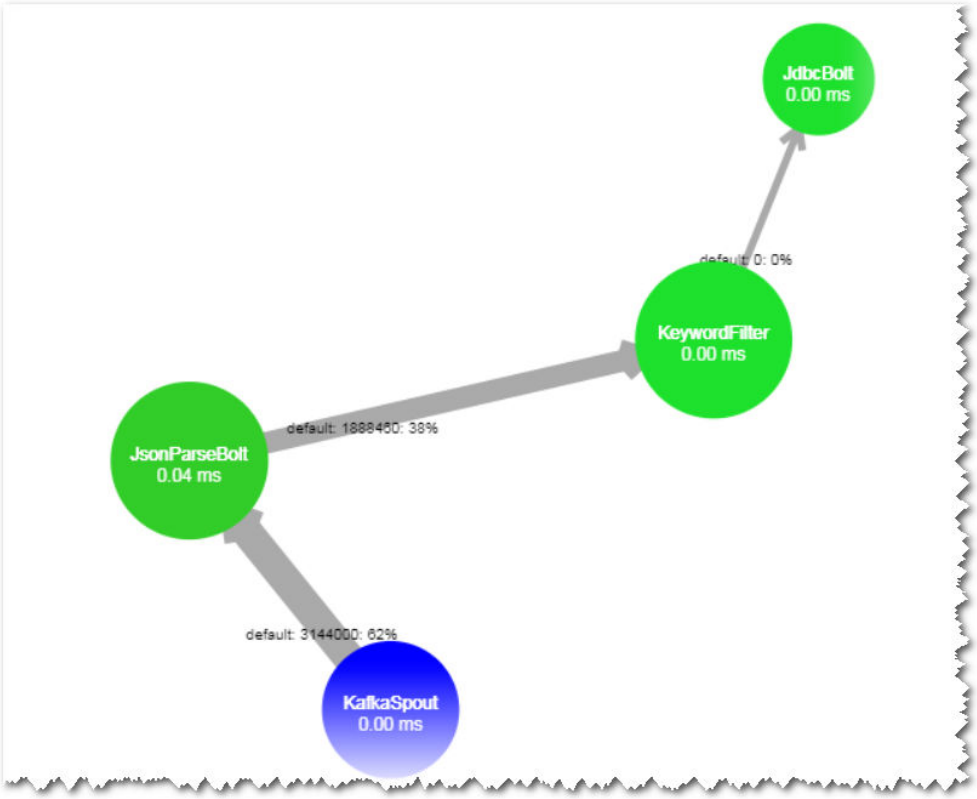


可能原因

- 拓扑定义异常。
- 数据库表结果定义异常。

原因分析

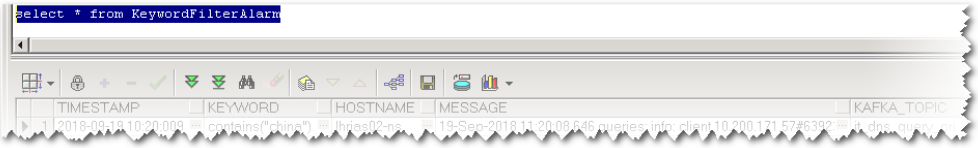
1. 通过Storm WebUI 查看拓扑DAG图，发现DAG图与拓扑定义一致。



2. 查看KeyWordFilter Bolt输出流字段定义和发送消息字段发现一致。

```
@Override
public void declareOutputFields(OutputFieldsDeclarer declarer)
{
    declarer.declare(new Fields("timestamp","keyword","hostname","message","kafka_topic") );
```

3. 查看Oracle数据库中表定义，发现字段名为大写，与流定义字段名称不一致。



4. 单独调试execute方法，发现抛出字段不存在。

处理步骤

修改流定义字段名称为大写，与数据库表定义字段一致。

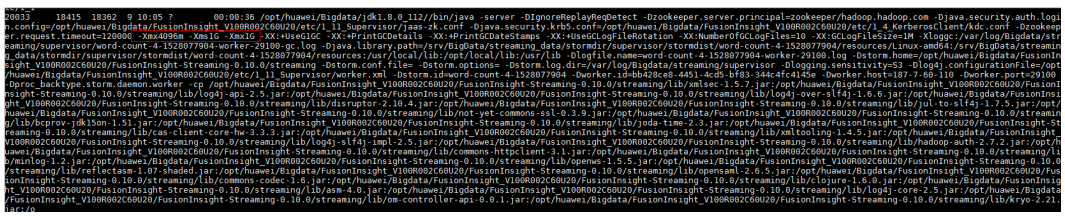
16.18.8 业务拓扑配置 GC 参数不生效

问题背景与现象

业务拓扑代码中配置参数topology.worker.childopts不生效，关键日志如下：

```
[main] INFO b.s.StormSubmitter - Uploading topology jar /opt/jar/example.jar to assigned location: /srv/
BigData/streaming/stormdir/nimbus/inbox/stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar
Start uploading file '/opt/jar/example.jar' to '/srv/BigData/streaming/stormdir/nimbus/inbox/
stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar' (65574612 bytes)
[==========================================================] 65574612 / 65574612
File '/opt/jar/example.jar' uploaded to '/srv/BigData/streaming/stormdir/nimbus/inbox/
stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar' (65574612 bytes)
[main] INFO b.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /srv/BigData/
streaming/stormdir/nimbus/inbox/stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar
[main] INFO b.s.StormSubmitter - Submitting topology word-count in distributed mode with conf
{"topology.worker.childopts":"-
Xmx4096m","storm.zookeeper.topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-59
15065013522446406:-6421330379815193999","topology.workers":1}
[main] INFO b.s.StormSubmitter - Finished submitting topology: word-count
```

通过ps -ef | grep worker命令查看worker进程信息如下：



原因分析

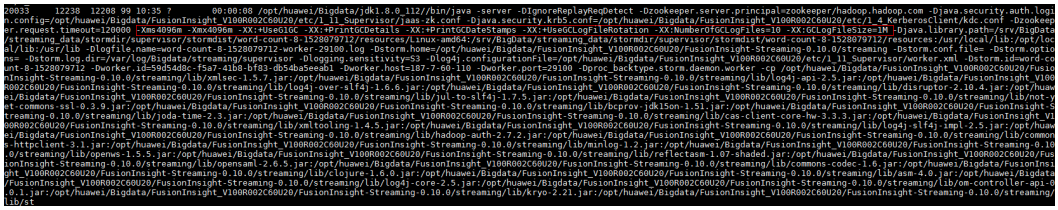
1. 由于topology.worker.gc.childopts、topology.worker.childopts和worker.gc.childopts(服务端参数)有优先级，优先级大小为：topology.worker.gc.childopts > worker.gc.childopts > topology.worker.childopts。
2. 如果设置了客户端参数topology.worker.childopts，则该参数会与服务端参数worker.gc.childopts共同配置，但是后面的相同参数会将前面的覆盖掉，如上面图有两个-Xmx，-Xmx1G会覆盖掉-Xmx4096m。
3. 如果配置了topology.worker.gc.childopts则服务端参数worker.gc.childopts会被替换。

解决办法

步骤1 如果想要修改拓扑的JVM参数，可以在命令中直接修改topology.worker.gc.childopts这个参数或者在服务端修改该参数，当topology.worker.gc.childopts为 "-Xms4096m -Xmx4096m -XX:+UseG1GC -XX:+PrintGCDetails -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M" 时，效果如下：

```
[main-SendThread(10.7.61.88:2181)] INFO o.a.s.s.o.a.z.ClientCnxn - Socket connection established, initiating session, client: /10.7.61.88:44694, server: 10.7.61.88/10.7.61.88:2181
[main-SendThread(10.7.61.88:2181)] INFO o.a.s.s.o.a.z.ClientCnxn - Session establishment complete on server 10.7.61.88/10.7.61.88:2181, sessionid = 0x16037a6e5f092575, negotiated timeout = 40000
[main-EventThread] INFO o.a.s.s.o.a.c.f.s.ConnectionStateManager - State change: CONNECTED
[main] INFO b.s.u.StormBoundedExponentialBackoffRetry - The baseSleepTimeMs [1000] the maxSleepTimeMs [1000] the maxRetries [1]
[main] INFO o.a.s.s.o.a.z.Login - successfully logged in.
[main-EventThread] INFO o.a.s.s.o.a.z.ClientCnxn - EventThread shut down for session: 0x16037a6e5f092575
[main] INFO o.a.s.s.o.a.z.ZooKeeper - Session: 0x16037a6e5f092575 closed
[main] INFO b.s.StormSubmitter - Uploading topology jar /opt/jar/example.jar to assigned location: /srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar
Start uploading file '/opt/jar/example.jar' to '/srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar' (74143745 bytes)
[==================================================] 74143745 / 74143745
File '/opt/jar/example.jar' uploaded to '/srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar' (74143745 bytes)
[main] INFO b.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar
[main] INFO b.s.StormSubmitter - Submitting topology word-count in distributed mode with conf {"storm.zookeeper.topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-7360002804241426074:-6868950379453400421","topology.worker.gc.childopts":"-Xms4096m -Xmx4096m -XX:+UseG1GC -XX:+PrintGCDetails -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M","topology.workers":1}
[main] INFO b.s.StormSubmitter - Finished submitting topology: word-count
```

步骤2 通过ps -ef | grep worker命令查看worker进程信息如下：



----结束

16.18.9 UI 查看信息显示 Internal Server Error

问题背景与现象

使用MRS版本安装集群，主要安装ZooKeeper、Strom。

通过MRS Manager中的Storm Status页面UI连接访问信息时出现Internal Server Error。

UI页面出现如下信息：

```
Internal Server Error
org.apache.thrift7.transport.TTransportException: Frame size (306030) larger than max length (1048576)!
```

可能原因

- Storm服务中Nimbus异常。
- Storm集群信息较多超过系统默认Thrift传输大小的设置。

原因分析

1. 查看Storm服务状态及监控指标：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Storm ，查看当前Storm状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Storm，
     查看当前Storm状态，发现状态为良好，且监控指标内容显示正确。
2. 选择“实例”页签，查看Nimbus实例状态，显示正常。
3. 查看当前Storm集群thrift相关配置，发现nimbus.thrift.max_buffer_size参数配置为1048576（1M）。
4. 上述配置和异常信息中信息一致，说明当前配置的Thrift的buffer size小于集群信息所需的buffer size。

解决方法

调整Storm集群中Thrift的Buffer Size大小，具体大小根据错误信息进行实际调整。

步骤1 进入Storm服务参数配置界面。

- MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Storm > 服务配置”，“参数类别”选择“全部配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 服务 > Yarn”，单击“配置”，选择“全部配置”。

步骤2 修改nimbus.thrift.max_buffer_size参数为10485760（10M）。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

16.19 使用 Ranger

16.19.1 Hive 启用 Ranger 鉴权后，在 Hue 页面能查看到没有权限的表和库

用户问题

Hive启用Ranger鉴权后，在Hue页面能查看到没有权限的表和库

问题现象

普通集群（未开启Kerberos认证）中，Hive启用Ranger鉴权后，在Hue页面能查看到没有权限的表和库。

原因分析

Hive启用Ranger鉴权后，默认的Hive策略中有2个关于database的public组策略， 所有用户都属于public组，默认给public组配有default数据库的创表和所有其他数据库的**create**权限，因此默认所有的用户都有**show databases**和**show tables**的权限，如果不想让某些用户有**show databases**和**show tables**权限，可在Ranger WEBUI中删除该默认public组策略，并赋予需要查看的用户权限，具体请参考处理步骤。

处理步骤

步骤1 登录Ranger WebUI界面。

步骤2 在“Service Manager”区域内，单击Hive组件名称，进入Hive组件安全访问策略列表页面。

步骤3 分别单击“all - database”和“default database tables columns”策略所在行的按钮。

步骤4 删除“public”组策略。

图 16-59 all - database 策略

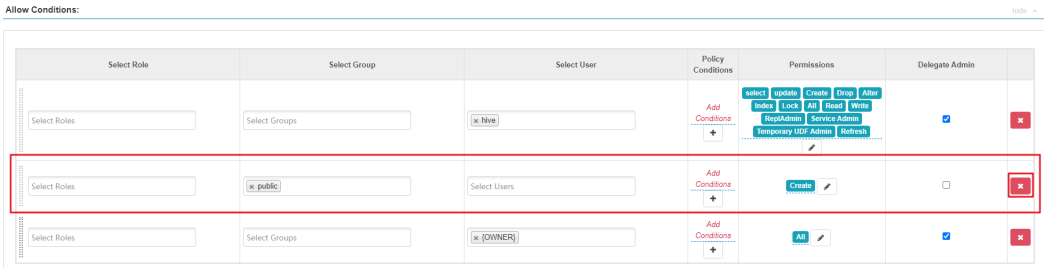
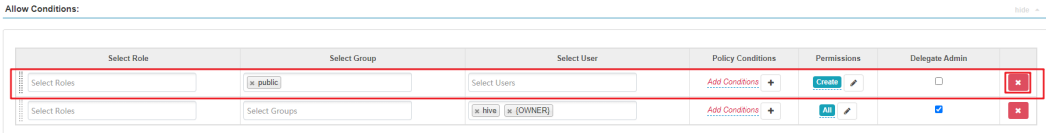


图 16-60 default database tables columns 策略



步骤5 在Hive组件安全访问策略列表页面，单击“Add New Policy”为相关用户或者用户组添加资源访问策略。

----结束

16.20 使用 Yarn

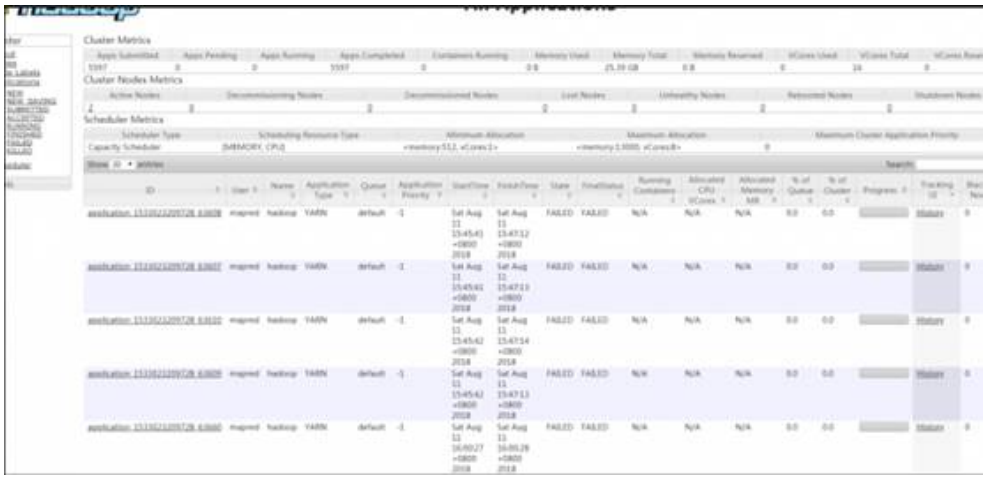
16.20.1 启动 Yarn 后发现一堆 job

用户问题

MRS 2.x及之前版本集群，构建MRS集群启动Yarn后，发现一堆job占用资源。

问题现象

客户使用MapReduce服务构建集群启动Yarn后 生成一堆job占用资源。



原因分析

- 疑似黑客攻击。
- 安全组入口方向的Any协议源地址配置为0.0.0.0/0。

处理步骤

步骤1 登录MRS集群页面，在“现有集群”中，单击对应的集群名称，进入集群详情页面。

步骤2 单击“集群管理页面”后面的“前往 Manager”，弹出“访问MRS Manager页面”。

步骤3 单击“管理安全组规则”，检查安全组规则配置。

步骤4 检查入口方向Any协议的源地址是否为0.0.0.0/0。

步骤5 如果是，修改入口方向Any协议的远端为指定IP地址。如果不是，则无需修改。

步骤6 修改成功后，重启集群虚机。

----结束

建议与总结

关闭入口方向的Any协议，或者指定入口方向的Any协议远端为指定IP。

参考信息

请参考。

16.20.2 通过客户端 hadoop jar 命令提交任务，客户端返回 GC overhead

问题背景与现象

通过客户端提交任务，客户端返回内存溢出的报错结果:

```
main path:hdfs://hacluster/user/wangyou
17/09/18 08:29:57 INFO hdfs.DFSClient: Created HDFS_DELEGATION_TOKEN token 22890097 for wangyou on ha-hdfs:hacluster
17/09/18 08:29:57 INFO security.TokenCache: Got dt for hdfs://hacluster; Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 2
2890097 for wangyou)
17/09/18 08:29:57 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
17/09/18 08:32:42 INFO retry.RetryInvocationHandler: Exception while invoking getListing of class ClientNamenodeProtocolTranslatorPB over fi1-cn-003/10.113.246.10:2500
0. Trying to fail over immediately.
java.io.IOException: com.google.protobuf.ServiceException: java.lang.OutOfMemoryError: GC overhead limit exceeded
        at org.apache.hadoop.ipc.ProtobufHelper.getRemoteException(ProtobufHelper.java:47)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.getListing(ClientNamenodeProtocolTranslatorPB.java:578)
        at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:497)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:191)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:102)
        at com.sun.proxy.$Proxy10.getListing(Unknown Source)
        at org.apache.hadoop.hdfs.DFSClient.listPaths(DFSClient.java:1757)
        at org.apache.hadoop.hdfs.DistributedFileSystem$DirListingIterator.hasNextNoFilter(DistributedFileSystem.java:1024)
        at org.apache.hadoop.hdfs.DistributedFileSystem$DirListingIterator.hasNext(DistributedFileSystem.java:999)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:304)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
        at org.apache.hadoop.mapreduce.lib.input.CombineFileInputFormat.getSplits(CombineFileInputFormat.java:217)
        at org.apache.hadoop.mapreduce.lib.input.DelegatingInputFormat.getSplits(DelegatingInputFormat.java:115)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:306)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:323)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:200)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1290)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1287)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1673)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1287)
```

原因分析

从报错堆栈可以看出是任务在提交过程中分片时在读取HDFS文件阶段内存溢出了，一般是由于该任务要读取的小文件很多导致内存不足。

解决办法

步骤1 排查启动的MR任务是否对应的HDFS文件个数很多，如果很多，减少文件数量，提前先合并小文件或者尝试使用combineInputFormat来减少任务读取的文件数量。

步骤2 增大hadoop命令执行时的内存，该内存在客户端中设置，修改对应路径“*客户端安装目录*/HDFS/component_env”文件中“CLIENT_GC_OPTS”的“-Xmx”参数，将该参数的默认值改大，比如改为512m。然后执行**source component_env**命令，使修改的参数生效。

```
export YARN_ROOT_LOGGER=INFO,console

#GC_OPTS for client operation.
CLIENT_GC_OPTS="-Xmx512m -Djava.io.tmpdir=${HADOOP_HOME}"

export HADOOP_CLIENT_OPTS="$CLIENT_GC_OPTS"
```

----结束

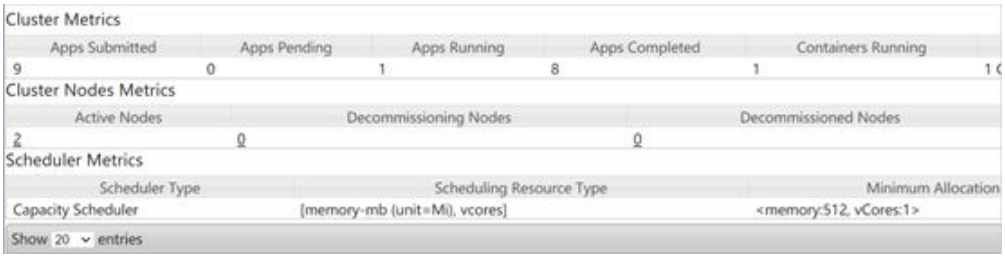
16.20.3 Yarn 汇聚日志过大导致磁盘被占满

用户问题

集群的磁盘使用率很高。

问题现象

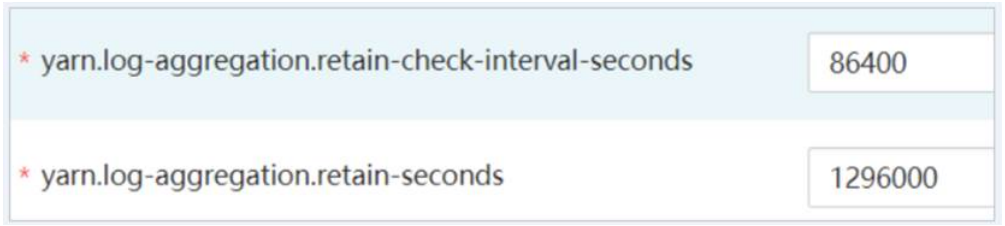
- Manager管理页面下主机管理显示磁盘使用率过高。
- Yarn WebUI界面上显示只有少量任务在运行。



- 登录到集群的Master节点执行**hdfs dfs -du -h /** 命令发现如下文件占用大量磁盘空间。

```
22.5 G   45.0 G   /tmp/logs/root/logs/application_1589278244866_0153
18.4 M   36.8 M   /tmp/logs/root/logs/application_1589278244866_0154
23.4 G   46.8 G   /tmp/logs/root/logs/application_1589278244866_0155
23.5 G   46.9 G   /tmp/logs/root/logs/application_1589278244866_0156
23.7 G   47.4 G   /tmp/logs/root/logs/application_1589278244866_0157
23.7 G   47.4 G   /tmp/logs/root/logs/application_1589278244866_0158
22.5 G   45.0 G   /tmp/logs/root/logs/application_1589278244866_0159
18.5 M   37.0 M   /tmp/logs/root/logs/application_1589278244866_0160
22.5 G   45.0 G   /tmp/logs/root/logs/application_1589278244866_0161
18.8 M   37.6 M   /tmp/logs/root/logs/application_1589278244866_0162
24.0 G   48.0 G   /tmp/logs/root/logs/application_1589278244866_0163
121.3 K  242.7 K  /tmp/logs/root/logs/application_1589278244866_0164
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0165
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0166
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0167
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0168
```

- Yarn服务的汇聚日志配置如下



原因分析

客户提交任务的操作过于频繁，且聚合后的日志文件被删除的时间配置为1296000，即聚合日志保留15天，导致汇聚的日志无法在短时时间内释放，从而引起磁盘被占满。

处理步骤

步骤1 登录Manager页面，进入MapReduce服务全部配置页面。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > MapReduce > 服务配置 > 全部配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 服务 > MapReduce > 配置 > 全部配置”。

步骤2 搜索“yarn.log-aggregation.retain-seconds”参数，并根据实际情况将yarn.log-aggregation.retain-seconds调小，比如调整为： 259200，即Yarn的聚合日志保留3天，到期后自动释放磁盘空间。

步骤3 保存配置，不勾选“重新启动受影响的服务或实例”。

步骤4 在业务空闲时执行该步骤重启MapReduce服务，重启服务会导致上层服务业务中断，影响集群的管理维护和业务，建议在空闲时执行。

1. 登录Manager页面。
2. 重启MapReduce服务。

----结束

16.20.4 MR 任务异常临时文件不删除

用户问题

MR任务异常临时文件为什么没有删除?

问题现象

HDFS临时目录文件过多，占用内存。

原因分析

MR任务提交时会将相关配置文件、jar包和-files添加的文件都放入hdfs上的临时目录，方便后面container启动以后获取相应的文件。由配置项yarn.app.mapreduce.am.staging-dir决定具体存放位置，默认值是/tmp/hadoop-yarn/staging。

正常运行的MR任务会在Job结束以后就清理这些临时文件，但是当Job对应的yarn任务是异常退出时，这些临时文件不会被清理，长时间积攒导致该临时目录下的文件数量越来越多，占用存储空间越来越多。

处理步骤

步骤1 登录集群。

1. 以root用户登录任意一个Master节点，用户密码为创建集群时用户自定义的密码。
2. 如果集群开启Kerberos认证，执行如下命令进入客户端安装目录并设置环境变量，再认证用户并按照提示输入密码，该密码请向管理员获取。

   cd *客户端安装目录*

   source bigdata_env

kinit hdfs

3. 如果集群未开启Kerberos认证，执行如下命令切换到omm用户，再进入客户端安装目录设置环境变量。

   su - omm

   cd *客户端安装目录*

   source bigdata_env

步骤2 获取文件列表。

hdfs dfs -ls /tmp/hadoop-yarn/staging/*/.staging/ | grep "^drwx" | awk '{print $8}' > job_file_list

job_file_list文件中就是所有任务的文件夹列表，文件内容参考：

```
/tmp/hadoop-yarn/staging/omm/.staging/job__<Timestamp>_<ID>
```

步骤3 统计当前运行中的任务。

mapred job -list 2>/dev/null | grep job_ | awk '{print $1}' > run_job_list

run_job_list文件里面就是当前正在运行的JobId列表，文件内容格式为：

```
job_<Timestamp>_<ID>
```

步骤4 删除job_file_list文件中正在运行中的任务。确保在删除过期数据时不会误删正在运行任务的数据。

cat run_job_list | while read line; do sed -i "/$line/d" job_file_list; done

步骤5 删除过期数据。

cat job_file_list | while read line; do hdfs dfs -rm -r $line; done

步骤6 清除临时文件。

rm -rf run_job_list job_file_list

----结束

16.20.5 提交任务的 Yarn 的 ResourceManager 报错 connection refused，且配置的 Yarn 端口为 8032

用户问题

请求提交任务的Yarn的ResourceManager报错connection refused，且配置的Yarn端口为8032。

问题现象

MRS的Yarn ResourceManager中的一个节点无法连接，且配置的Yarn端口为8032。

原因分析

该业务应用在集群外部运行，且使用的是老集群的客户端，配的Yarn端口是8032，与MRS服务的Yarn ResourceManager实际端口不同。从而使请求提交任务的Yarn的ResourceManager报错connection refused。

处理步骤

步骤1 更新MRS服务客户端。

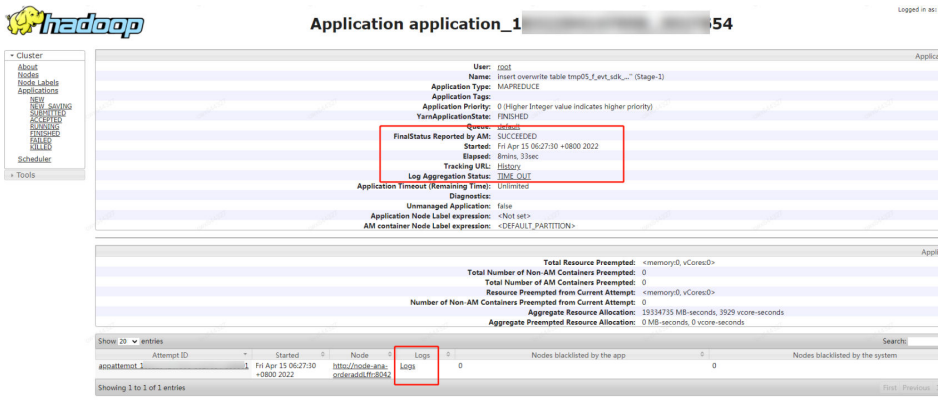
步骤2 然后重试提交作业。

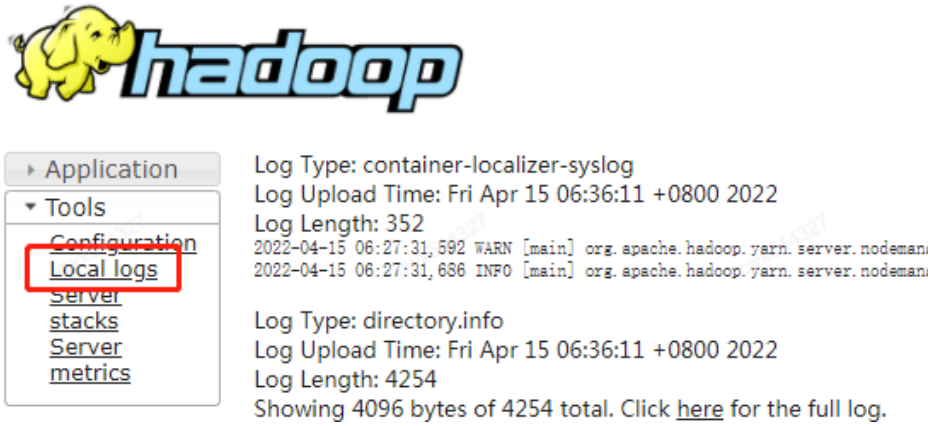
----结束

16.20.6 Yarn WebUI 作业查看日志提示“Could not access logs page!”

问题背景与现象

登录Yarn WebUI界面查看作业日志“Logs”，然后单击“Local logs”，界面提示“Could not access logs page!”





原因分析

该Local logs是用来访问服务的日志，但由于安全考虑，该功能暂不支持从Yarn WebUI界面访问。您可以登录ResourceManager主节点查看ResourceManager的日志。

处理步骤

步骤1 登录Manager界面，选择“集群 > 组件 > Yarn > 实例”，查看并获取“ResourceManager(主)”实例的业务IP地址。

步骤2 以**root**用户登录ResourceManager主节点。

步骤3 进入“/var/log/Bigdata/yarn/rm”路径查看ResourceManager的日志。

cd /var/log/Bigdata/yarn/rm

----结束

16.20.7 Yarn 页面单击队列名称报错

问题背景与现象

在Yarn使用Capacity调度器时，单击Yarn原生页面的队列名称会报500的错误。

HTTP ERROR 500 javax.servlet.ServletException: javax.servlet.ServletException: java.lang.IllegalArgumentException: Illegal character in query at index 81: https://[redacted]:20026/Yarn/ResourceManager/21/cluster/scheduler?openQueues=^default$

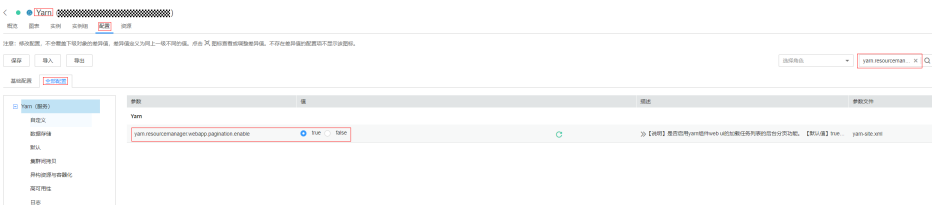
原因分析

页面链接无法识别符号“^”，导致页面访问失败。

处理步骤

步骤1 登录Manager页面，选择“集群 > 服务 > Yarn > 配置 > 全部配置”。

步骤2 在搜索框搜索“yarn.resourcemanager.webapp.pagination.enable”。



步骤3 如果该参数值为“true”（默认为“true”），请修改为“false”，并保存配置。

步骤4 在Yarn服务页面选择“实例”页签，勾选所有的ResourceManager实例，选择“更多 > 滚动重启实例”，等待实例启动完成。

----结束

16.21 使用 ZooKeeper

16.21.1 MRS 集群如何访问 ZooKeeper

用户问题

MRS集群如何访问ZooKeeper?

问题现象

客户在MRS的Master节点使用zkcli.sh访问ZooKeeper但是存在报错。

原因分析

客户使用命令有问题，造成报错的发生。

处理步骤

步骤1 获取ZooKeeper的IP地址。

步骤2 以root用户登录Master节点。

步骤3 初始化环境变量。

source /opt/client/bigdata_env

步骤4 执行**zkCli.sh -server** *ZooKeeper所在节点的IP*:**2181**即可连接上MRS的ZooKeeper。

zk所在节点的IP即为**步骤1**中查到的结果，多个IP之间以逗号间隔。

步骤5 使用**ls /**等常用的命令查看ZooKeeper上的信息。

----结束

16.22 访问 OBS

16.22.1 使用 MRS 多用户访问 OBS 功能时/tmp 目录没有权限

用户问题

在使用MRS多用户访问OBS功能，执行spark、hive、presto等作业时，出现/tmp目录没有权限的报错。

问题现象

在使用MRS多用户访问OBS功能，执行spark、hive、presto等作业时，出现/tmp目录没有权限的报错。

原因分析

作业执行过程中有临时目录，提交作业的用户对临时目录没有权限。

处理步骤

步骤1 在集群“概览”页签中，查询并记录集群所绑定的委托名称。

步骤2 登录IAM服务控制台。

步骤3 选择“权限 > 创建自定义策略”。

- 策略名称：请输入策略名称。
- 作用范围：请选择“全局级服务”。
- 策略配置方式：请选择“可视化视图”。
- 策略内容：

a. “允许”选择“允许”。
b. “云服务”选择“对象存储服务 (OBS)”。
c. “操作”勾选所有“写”、“列表”和“只读”权限。
d. “特定资源”选择：
   i. “object”选择“通过资源路径指定”，并单击“添加资源路径”分别在“路径”中输入***obs_bucket_name*/tmp/**和***obs_bucket_name*/tmp/***。此处以**/tmp**目录为例，如需其他目录权限请参考该步骤添加对应目录及该目录下所有对象的资源路径。
   ii. “bucket”选择“通过资源路径指定”，并单击“添加资源路径”在“路径”中输入***obs_bucket_name***。

   其中**obs_bucket-name**请使用实际的OBS桶名替换。若桶类型为“并行文件系统”需要在添加***obs_bucket_name*/tmp/**路径，桶类型为“对象存储”则不需要添加***obs_bucket_name*/tmp/**路径。
e. （可选）请求条件，暂不添加。

图 16-61 自定义策略

步骤4 单击“确定”完成策略添加。

步骤5 选择“委托”，并在**步骤1**中查询到的委托所在行的“操作”列单击“权限配置”。

步骤6 查询并勾选**步骤3**中创建的策略。

步骤7 单击“确定”完成委托权限配置。

----结束

16.22.2 Hadoop 客户端删除 OBS 上数据时.Trash 目录没有权限

用户问题

使用Hadoop客户端删除OBS上数据时出现.Trash目录没有权限的报错。

问题现象

执行**hadoop fs -rm obs://<obs_path>**出现如下报错：

```
exception [java.nio.file.AccessDeniedException: user/root/.Trash/Current/: getFileStatus on user/root/.Trash/Current/: status [403]
```

原因分析

hadoop删除文件时会将文件先移动到.Trash目录，若该目录没有权限则出现403报错。

处理步骤

方案一：

使用**hadoop fs -rm -skipTrash**来删除文件。

方案二：

在集群对应的委托中添加访问.Trash目录的权限。

步骤1 在集群“概览”页签中，查询并记录集群所绑定的委托名称。

步骤2 登录IAM服务控制台。

步骤3 选择“权限 > 创建自定义策略”。

- 策略名称：请输入策略名称。
- 作用范围：请选择“全局级服务”。
- 策略配置方式：请选择“可视化视图”。
- 策略内容：
  a. “允许”选择“允许”。
  b. “云服务”选择“对象存储服务 (OBS)”。
  c. “操作”勾选所有操作权限。
  d. “特定资源”选择：
     i. “object”选择“通过资源路径指定”，并单击“添加资源路径”分别在“路径”中输入.Trash目录，例如***obs_bucket_name*/user/root/.Trash/***。
     ii. “bucket”选择“通过资源路径指定”，并单击“添加资源路径”在“路径”中输入***obs_bucket_name***。

     其中**obs_bucket-name**请使用实际的OBS桶名替换。
  e. （可选）请求条件，暂不添加。

图 16-62 自定义策略

步骤4 单击“确定”完成策略添加。

步骤5 选择“委托”，并在**步骤1**中查询到的委托所在行的“操作”列单击“权限配置”。

步骤6 查询并勾选**步骤3**中创建的策略。

步骤7 单击“确定”完成委托权限配置。

步骤8 重新执行**hadoop fs -rm obs://<obs_path>**命令。

----结束

17 附录

17.1 MRS 3.x 版本操作注意事项

概述

MRS 3.x之前的版本的MRS集群使用MRS Manager对集群进行管理、监控，同时用户可通过MRS管理控制台的集群管理页面，进行集群概览查看、节点管理、组件管理、告警管理、补丁管理、文件管理、作业管理、租户管理、备份恢复、引导操作设置及标签管理。

MRS 3.x版本的MRS集群使用FusionInsight Manager对集群进行管理、监控，同时用户可通过MRS管理控制台的集群管理页面，进行集群概览查看、节点管理、组件管理、告警管理、文件管理、作业管理、引导操作设置及标签管理。

MRS 3.x版本集群的部分维护操作与历史版本有部分差异，更多详细操作可参考本**MRS Manager操作指导（适用于2.x及之前）**与**FusionInsight Manager操作指导（适用于3.x）**。

访问 MRS 集群 Manager

- 访问MRS 3.x之前的版本的MRS Manager请参考**访问MRS Manager（MRS 2.x及之前版本）**。
- 访问MRS 3.x版本的FusionInsight Manager请参考**访问FusionInsight Manager（MRS 3.x及之后版本）**。

修改 MRS 集群服务配置参数

- MRS 3.x之前的版本，用户可直接通过MRS管理控制台的集群管理页面修改各服务配置参数：
  a. 登录MRS控制台，在左侧导航栏选择“集群列表> 现有集群”，单击集群名称。
  b. 选择“组件管理 > *服务名称* > 服务配置”。
     默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。

c. 在导航树选择指定的参数分类，并在右侧修改参数值。

不确定参数的具体位置时，支持在右上角输入参数名，系统将实时进行搜索并显示结果。

d. 单击“保存配置”，并在确认对话框中单击“确定”。

e. 等待界面提示“操作成功”，单击“完成”，配置已修改。

查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。也可在保存配置时直接勾选“重新启动受影响的服务或实例。”。

- MRS 3.x版本，服务配置参数需登录FusionInsight Manager修改：

a. 登录FusionInsight Manager。

b. 选择“集群 > 服务”。

c. 单击服务视图中指定的服务名称。

d. 单击“配置”。

默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。

e. 在导航树选择指定的参数分类，并在右侧修改参数值。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

f. 单击“保存”，并在确认对话框中单击“确定”。

g. 等待界面提示“操作成功”，单击“完成”，配置已修改。

查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。