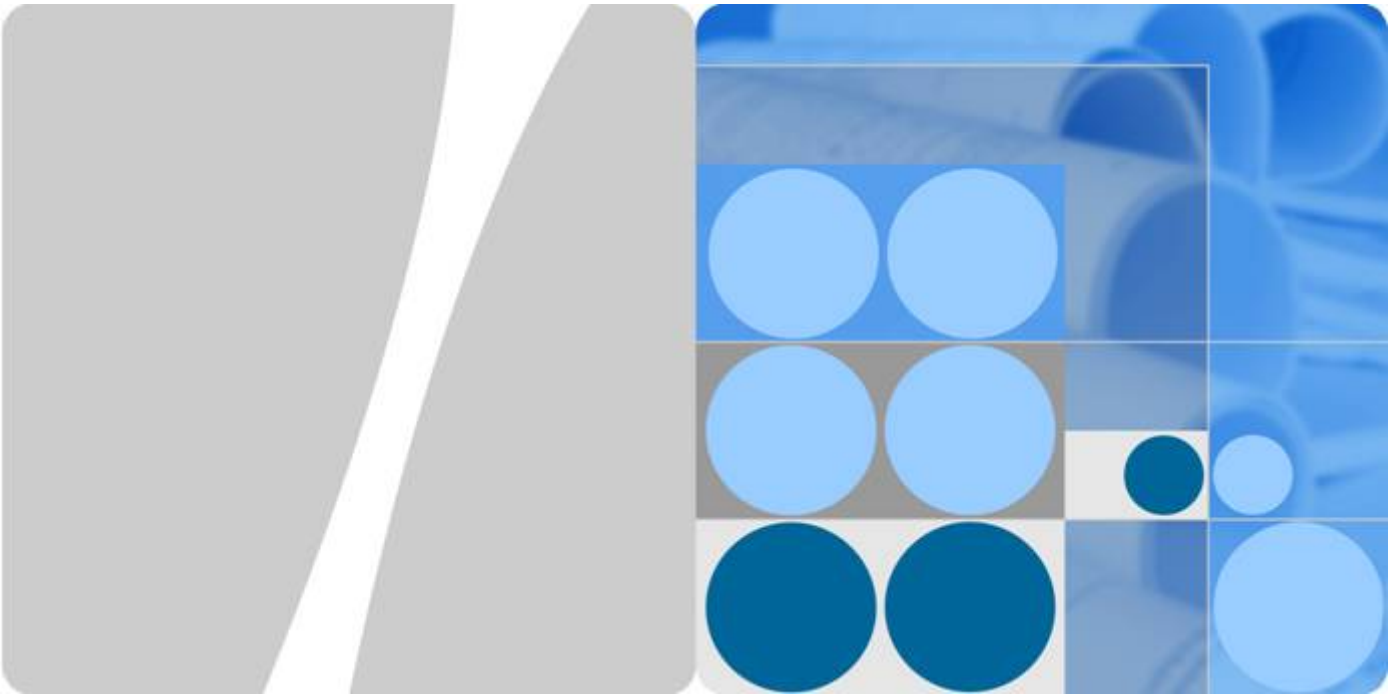


MapReduce 服务

用户指南

发布日期　2025-11-25

目 录

1 简介.....1
1.1 什么是 MRS.....1
1.2 应用场景.....4
1.3 如何选择 MRS.....6
1.4 组件介绍.....6
1.4.1 Alluxio.....6
1.4.2 CarbonData.....7
1.4.3 ClickHouse.....9
1.4.4 DBService.....13
1.4.5 Flink.....14
1.4.5.1 Flink 基本原理.....15
1.4.5.2 Flink HA 方案介绍.....20
1.4.5.3 Flink 与其他组件的关系.....22
1.4.5.4 Flink 开源增强特性.....23
1.4.5.4.1 窗口.....23
1.4.5.4.2 Job Pipeline.....25
1.4.5.4.3 Stream SQL Join.....30
1.4.5.4.4 Flink CEP in SQL.....31
1.4.6 Flume.....33
1.4.6.1 Flume 基本原理.....33
1.4.6.2 Flume 与其他组件的关系.....36
1.4.6.3 Flume 开源增强特性.....36
1.4.7 HBase.....36
1.4.7.1 HBase 基本原理.....36
1.4.7.2 HBase HA 方案介绍.....41
1.4.7.3 HBase 与其他组件的关系.....42
1.4.7.4 HBase 开源增强特性.....43
1.4.8 HDFS.....50
1.4.8.1 HDFS 基本原理.....50
1.4.8.2 HDFS HA 方案介绍.....53
1.4.8.3 HDFS 与其他组件的关系.....54
1.4.8.4 HDFS 开源增强特性.....56
1.4.9 Hive.....62

1.4.9.1 Hive 基本原理 62
1.4.9.2 Hive CBO 原理介绍 65
1.4.9.3 Hive 与其他组件的关系 69
1.4.9.4 Hive 开源增强特性 69
1.4.10 Hudi 70
1.4.11 Hue 72
1.4.11.1 Hue 基本原理 72
1.4.11.2 Hue 与其他组件的关系 74
1.4.11.3 Hue 开源增强特性 76
1.4.12 Impala 76
1.4.13 Kafka 77
1.4.13.1 Kafka 基本原理 77
1.4.13.2 Kafka 与其他组件的关系 80
1.4.13.3 Kafka 开源增强特性 81
1.4.14 KafkaManager 81
1.4.15 KrbServer 及 LdapServer 81
1.4.15.1 KrbServer 及 LdapServer 基本原理 81
1.4.15.2 KrbServer 及 LdapServer 开源增强特性 85
1.4.16 Kudu 85
1.4.17 Loader 86
1.4.17.1 Loader 基本原理 86
1.4.17.2 Loader 与其他组件的关系 88
1.4.17.3 Loader 开源增强特性 88
1.4.18 Manager 89
1.4.18.1 Manager 基本原理 89
1.4.18.2 Manager 关键特性 91
1.4.19 MapReduce 92
1.4.19.1 MapReduce 基本原理 92
1.4.19.2 MapReduce 与其他组件的关系 94
1.4.19.3 MapReduce 开源增强特性 94
1.4.20 Oozie 97
1.4.20.1 Oozie 基本原理 97
1.4.20.2 Oozie 开源增强特性 99
1.4.21 OpenTSDB 99
1.4.22 Presto 100
1.4.23 Ranger 101
1.4.23.1 Ranger 基本原理 101
1.4.23.2 Ranger 与其他组件的关系 103
1.4.24 Spark 103
1.4.24.1 Spark 基本原理 104
1.4.24.2 Spark HA 方案介绍 118
1.4.24.3 Spark 与 HDFS 和 YARN 的关系 123

1.4.24.4 Spark 开源增强特性：跨源复杂数据的 SQL 查询优化 127
1.4.25 Spark2x 129
1.4.25.1 Spark2x 基本原理 129
1.4.25.2 Spark2x HA 方案介绍 142
1.4.25.2.1 Spark2x 多主实例 142
1.4.25.2.2 Spark2x 多租户 145
1.4.25.3 Spark2x 与组件的关系 148
1.4.25.4 Spark2x 开源新特性 152
1.4.25.5 Spark2x 开源增强特性 152
1.4.25.5.1 CarbonData 简介 152
1.4.25.5.2 跨源复杂数据的 SQL 查询优化 154
1.4.26 Storm 156
1.4.26.1 Storm 基本原理 156
1.4.26.2 Storm 与其他组件的关系 160
1.4.26.3 Storm 开源增强特性 161
1.4.27 Tez 161
1.4.28 YARN 162
1.4.28.1 YARN 基本原理 162
1.4.28.2 YARN HA 方案介绍 166
1.4.28.3 Yarn 与其他组件的关系 168
1.4.28.4 YARN 开源增强特性 171
1.4.29 ZooKeeper 177
1.4.29.1 ZooKeeper 基本原理 177
1.4.29.2 ZooKeeper 与其他组件的关系 179
1.4.29.3 ZooKeeper 开源增强特性 182
1.5 产品功能 185
1.5.1 多租户 185
1.5.2 安全增强 186
1.5.3 组件 WebUI 便捷访问 187
1.5.4 可靠性增强 188
1.5.5 作业管理 189
1.5.6 自定义引导操作 189
1.5.7 元数据 190
1.5.8 集群管理 190
1.5.8.1 集群生命周期管理 190
1.5.8.2 集群扩缩容 191
1.5.8.3 自动弹性伸缩 192
1.5.8.4 创建 Task 节点 193
1.5.8.5 升级 Master 节点规格 193
1.5.8.6 隔离主机 194
1.5.8.7 标签管理 194
1.5.9 集群运维 194

1.5.10 消息通知 195
1.6 约束与限制 196
1.7 技术支持 197
1.8 权限管理 197
1.9 与其他云服务的关系 203
1.10 配额说明 205
1.11 常见概念 205
2 入门 209
2.1 如何使用 MRS 209
2.2 创建集群 209
2.3 上传示例数据和程序 211
2.4 添加作业 213
2.5 快速使用 Kerberos 认证集群 217
2.6 删除集群 223
3 准备用户 224
3.1 创建 MRS 操作用户 224
3.2 创建 MRS 自定义策略 228
3.3 IAM 用户同步 MRS 说明 233
4 配置集群 238
4.1 创建方式简介 238
4.2 快速集群 238
4.2.1 快速创建 Hadoop 分析集群 238
4.2.2 快速创建 HBase 查询集群 240
4.2.3 快速创建 Kafka 流式集群 241
4.2.4 快速创建 ClickHouse 集群 242
4.2.5 快速创建实时分析集群 244
4.3 创建自定义集群 245
4.4 创建自定义拓扑集群 256
4.5 添加集群标签 264
4.6 授权安全通信 265
4.7 配置弹性伸缩规则 268
4.7.1 简介 268
4.7.2 创建集群时配置弹性伸缩 269
4.7.3 为已有集群新增弹性伸缩策略 270
4.7.4 场景 1：单独配置弹性伸缩规则 271
4.7.5 场景 2：单独使用资源计划 272
4.7.6 场景 3：弹性伸缩规则与资源计划叠加使用 273
4.7.7 修改已有弹性伸缩策略 274
4.7.8 删除已有弹性伸缩策略 274
4.7.9 开启或关闭已有弹性伸缩策略 275
4.7.10 查看已有弹性伸缩策略 275

4.7.11 配置自动化脚本 275
4.7.12 配置弹性伸缩指标 276
4.8 管理数据连接 281
4.8.1 配置数据连接 281
4.8.2 配置 RDS 数据连接操作 284
4.8.3 配置 Ranger 数据连接 287
4.8.4 配置 Hive 数据连接 292
4.9 通过引导操作安装第三方软件 294
4.10 查看失败的集群操作任务 296
4.11 查看历史集群信息 296
5 管理集群 299
5.1 登录集群 299
5.1.1 MRS 集群节点简介 299
5.1.2 登录集群节点 300
5.1.3 如何确认 Manager 的主备管理节点 305
5.2 集群概览 305
5.2.1 集群列表简介 306
5.2.2 查看集群状态 307
5.2.3 查看集群基本信息 309
5.2.4 查看集群补丁信息 313
5.2.5 查看和定制集群监控指标 313
5.2.6 管理组件和主机监控 316
5.3 集群运维 320
5.3.1 导入导出数据 320
5.3.2 切换集群子网 323
5.3.3 配置消息通知 325
5.3.4 健康检查 327
5.3.4.1 使用前须知 327
5.3.4.2 执行健康检查 327
5.3.4.3 查看并导出检查报告 328
5.3.5 远程运维 329
5.3.5.1 运维授权 329
5.3.5.2 日志共享 329
5.3.6 查看 MRS 服务操作日志 329
5.3.7 删除集群 331
5.4 节点管理 331
5.4.1 扩容集群 331
5.4.2 缩容集群 333
5.4.3 管理主机（节点）操作 335
5.4.4 隔离主机 336
5.4.5 取消隔离主机 336
5.4.6 升级 Master 节点规格 337

5.5 作业管理 338
5.5.1 MRS 作业简介 338
5.5.2 运行 MapReduce 作业 342
5.5.3 运行 SparkSubmit 作业 346
5.5.4 运行 HiveSql 作业 351
5.5.5 运行 SparkSql 作业 355
5.5.6 运行 Flink 作业 360
5.5.7 运行 Kafka 作业 365
5.5.8 查看作业配置信息和日志 366
5.5.9 停止作业 367
5.5.10 删除作业 367
5.5.11 使用 OBS 加密数据运行作业 368
5.5.12 配置作业消息通知 374
5.6 组件管理 375
5.6.1 对象管理简介 375
5.6.2 查看配置 375
5.6.3 管理服务操作 376
5.6.4 配置服务参数 378
5.6.5 配置服务自定义参数 378
5.6.6 同步服务配置 380
5.6.7 管理角色实例操作 380
5.6.8 配置角色实例参数 381
5.6.9 同步角色实例配置 382
5.6.10 退服和入服角色实例 382
5.6.11 启动及停止集群 383
5.6.12 同步集群配置 383
5.6.13 导出集群的配置数据 384
5.6.14 支持滚动重启 384
5.7 告警管理 388
5.7.1 查看告警列表 388
5.7.2 查看事件列表 390
5.7.3 查看与手动清除告警 393
5.8 补丁管理 394
5.8.1 MRS 1.7.0 前版本补丁操作指导 394
5.8.2 MRS 1.7.0 至 2.0.1 版本补丁操作指导 395
5.8.3 滚动补丁 396
5.8.4 修复隔离主机补丁 398
5.9 租户管理 399
5.9.1 使用前须知 399
5.9.2 租户简介 399
5.9.3 添加租户 400
5.9.4 添加子租户 402

5.9.5 删除租户 404
5.9.6 管理租户目录 405
5.9.7 恢复租户数据 406
5.9.8 添加资源池 407
5.9.9 修改资源池 408
5.9.10 删除资源池 408
5.9.11 配置队列 409
5.9.12 配置资源池的队列容量策略 411
5.9.13 清除队列配置 411
5.10 引导操作 412
5.10.1 引导操作简介 412
5.10.2 准备引导操作脚本 413
5.10.3 查看执行记录 413
5.10.4 添加引导操作 414
5.10.5 修改引导操作 415
5.10.6 删除引导操作 415
6 使用 MRS 客户端 417
6.1 安装客户端 417
6.1.1 安装客户端（3.x 及之后版本） 417
6.1.2 安装客户端（3.x 之前版本） 422
6.2 更新客户端 426
6.2.1 更新客户端（3.x 及之后版本） 426
6.2.2 更新客户端（3.x 之前版本） 428
6.3 各组件客户端使用实践 431
6.3.1 使用 ClickHouse 客户端 431
6.3.2 使用 Flink 客户端 434
6.3.3 使用 Flume 客户端 441
6.3.4 使用 HBase 客户端 446
6.3.5 使用 HDFS 客户端 448
6.3.6 使用 Hive 客户端 450
6.3.7 使用 Impala 客户端 453
6.3.8 使用 Kafka 客户端 456
6.3.9 使用 Kudu 客户端 458
6.3.10 使用 Oozie 客户端 459
6.3.11 使用 Storm 客户端 460
6.3.12 使用 Yarn 客户端 461
7 配置存算分离 463
7.1 存算分离简介 463
7.2 配置存算分离集群（委托方式） 463
7.3 配置存算分离集群（AKSK 方式） 470
7.4 使用存算分离集群 473
7.4.1 Flink 对接 OBS 文件系统 473

7.4.2 Flume 对接 OBS 文件系统 474
7.4.3 HDFS 客户端对接 OBS 文件系统 475
7.4.4 Hive 对接 OBS 文件系统 476
7.4.5 MapReduce 对接 OBS 文件系统 479
7.4.6 Spark2x 对接 OBS 文件系统 480
7.4.7 Sqoop 对接外部存储系统 482
7.4.8 Hudi 对接 OBS 文件系统 485
8 访问 MRS 集群上托管的开源组件 Web 页面 488
8.1 开源组件 Web 站点 488
8.2 开源组件端口列表 491
8.3 通过专线访问 505
8.4 通过弹性公网 IP 访问 506
8.5 通过 Windows 弹性云服务器访问 507
8.6 创建连接 MRS 集群的 SSH 隧道并配置浏览器 508
9 Jupyter Notebook 采用自定义 Python 对接 MRS 512
9.1 简介 512
9.2 安装集群外节点客户端 512
9.3 安装 Python3 513
9.4 配置 MRS 客户端 516
9.5 安装 Jupyter Notebook 516
9.6 验证 Jupyter Notebook 访问 MRS 517
9.7 常见问题 518
10 访问集群 Manager 521
10.1 访问 FusionInsight Manager（MRS 3.x 及之后版本） 521
10.2 访问 MRS Manager（MRS 2.x 及之前版本） 523
11 FusionInsight Manager 操作指导（适用于 3.x） 527
11.1 从这里开始 527
11.1.1 FusionInsight Manager 入门指导 527
11.1.2 查询 FusionInsight Manager 版本号 528
11.1.3 登录管理系统 529
11.1.4 登录管理节点 529
11.2 主页 530
11.2.1 主页概述 530
11.2.2 管理监控指标数据报表 531
11.3 集群 533
11.3.1 管理集群 533
11.3.1.1 集群管理概述 533
11.3.1.2 滚动重启集群 534
11.3.1.3 管理配置过期 536
11.3.1.4 下载客户端 536
11.3.1.5 修改集群属性 537

11.3.1.6 管理集群配置 538
11.3.1.7 静态服务池 539
11.3.1.7.1 静态服务资源 539
11.3.1.7.2 配置集群静态资源 540
11.3.1.7.3 查看集群静态资源 541
11.3.1.8 客户端管理 542
11.3.1.8.1 管理客户端 542
11.3.1.8.2 批量升级客户端 543
11.3.1.8.3 批量刷新 hosts 文件 545
11.3.2 管理服务 545
11.3.2.1 服务管理概述 545
11.3.2.2 其他服务管理操作 549
11.3.2.2.1 服务详情概述 549
11.3.2.2.2 执行角色实例主备倒换 551
11.3.2.2.3 资源监控 551
11.3.2.2.4 采集堆栈信息 554
11.3.2.2.5 切换 Ranger 鉴权 555
11.3.2.3 服务配置 556
11.3.2.3.1 修改服务配置参数 556
11.3.2.3.2 修改服务自定义配置参数 557
11.3.3 管理实例 559
11.3.3.1 实例管理概述 559
11.3.3.2 入服与退服实例 561
11.3.3.3 管理实例配置 562
11.3.3.4 查看实例配置文件 563
11.3.3.5 实例组 563
11.3.3.5.1 管理实例组 563
11.3.3.5.2 查看实例组信息 565
11.3.3.5.3 配置实例组参数 565
11.4 主机 566
11.4.1 主机管理页面 566
11.4.1.1 查看主机列表 566
11.4.1.2 查看主机概览 567
11.4.1.3 查看主机进程及资源 568
11.4.2 主机维护操作 568
11.4.2.1 启动、停止主机上的所有实例 568
11.4.2.2 执行主机健康检查 568
11.4.2.3 分配机架 569
11.4.2.4 隔离主机 571
11.4.2.5 导出主机信息 572
11.4.3 资源概况 572
11.4.3.1 分布 572

11.4.3.2 趋势 574
11.4.3.3 集群 575
11.4.3.4 主机 576
11.5 运维 576
11.5.1 告警 576
11.5.1.1 告警与事件概述 576
11.5.1.2 配置阈值 579
11.5.1.3 配置告警屏蔽状态 591
11.5.2 日志 592
11.5.2.1 在线检索日志 592
11.5.2.2 下载日志 594
11.5.3 健康检查 594
11.5.3.1 查看健康检查任务 595
11.5.3.2 管理健康检查报告 595
11.5.3.3 修改健康检查配置 596
11.5.4 备份恢复设置 596
11.5.4.1 创建备份任务 596
11.5.4.2 创建恢复任务 597
11.5.4.3 其他任务管理说明 598
11.6 审计 598
11.6.1 审计管理页面概述 598
11.6.2 配置审计日志转储 599
11.7 租户资源 600
11.7.1 多租户介绍 601
11.7.1.1 简介 601
11.7.1.2 技术原理 601
11.7.1.2.1 多租户管理页面概述 601
11.7.1.2.2 相关模型 604
11.7.1.2.3 资源概述 607
11.7.1.2.4 动态资源 608
11.7.1.2.5 存储资源 610
11.7.1.3 多租户使用 610
11.7.1.3.1 使用说明 610
11.7.1.3.2 流程概述 611
11.7.2 使用 Superior 调度器的租户业务 612
11.7.2.1 创建租户 613
11.7.2.1.1 添加租户 613
11.7.2.1.2 添加子租户 616
11.7.2.1.3 添加用户并绑定租户的角色 619
11.7.2.2 管理租户 620
11.7.2.2.1 管理租户目录 620
11.7.2.2.2 恢复租户数据 622

11.7.2.2.3 删除租户 623
11.7.2.3 管理资源 623
11.7.2.3.1 添加资源池 624
11.7.2.3.2 修改资源池 624
11.7.2.3.3 删除资源池 625
11.7.2.3.4 配置队列 625
11.7.2.3.5 配置资源池的队列容量策略 627
11.7.2.3.6 清除队列容量配置 628
11.7.2.4 管理全局用户策略 628
11.7.3 使用 Capacity 调度器的租户业务 629
11.7.3.1 创建租户 629
11.7.3.1.1 添加租户 629
11.7.3.1.2 添加子租户 632
11.7.3.1.3 添加用户并绑定租户的角色 635
11.7.3.2 管理租户 636
11.7.3.2.1 管理租户目录 636
11.7.3.2.2 恢复租户数据 638
11.7.3.2.3 删除租户 639
11.7.3.2.4 Capacity Scheduler 模式下清除租户非关联队列 640
11.7.3.3 管理资源 641
11.7.3.3.1 添加资源池 641
11.7.3.3.2 修改资源池 641
11.7.3.3.3 删除资源池 642
11.7.3.3.4 配置队列 642
11.7.3.3.5 配置资源池的队列容量策略 644
11.7.3.3.6 清除队列容量配置 644
11.7.4 切换调度器 645
11.8 系统设置 647
11.8.1 权限设置 647
11.8.1.1 用户管理 647
11.8.1.1.1 创建用户 647
11.8.1.1.2 修改用户信息 648
11.8.1.1.3 导出用户信息 649
11.8.1.1.4 锁定用户 649
11.8.1.1.5 解锁用户 650
11.8.1.1.6 删除用户 650
11.8.1.1.7 修改用户密码 651
11.8.1.1.8 初始化用户密码 652
11.8.1.1.9 导出认证凭据文件 653
11.8.1.2 用户组管理 653
11.8.1.3 角色管理 654
11.8.1.4 安全策略 656

11.8.1.4.1 配置密码策略 656
11.8.1.4.2 配置私有属性 659
11.8.2 对接设置 660
11.8.2.1 配置 SNMP 北向参数 660
11.8.2.2 配置 Syslog 北向参数 662
11.8.2.3 配置监控指标数据转储 666
11.8.3 导入证书 668
11.8.4 OMS 管理 669
11.8.4.1 OMS 维护页面概述 669
11.8.4.2 修改 OMS 服务配置参数 670
11.8.5 部件管理 672
11.8.5.1 查看部件包 672
11.9 集群管理 673
11.9.1 配置客户端 673
11.9.1.1 安装客户端 673
11.9.1.2 使用客户端 677
11.9.1.3 更新已安装客户端的配置 678
11.9.2 集群互信管理 679
11.9.2.1 集群互信概述 679
11.9.2.2 修改 Manager 系统域名 680
11.9.2.3 配置跨 Manager 集群互信 683
11.9.2.4 配置跨集群互信后的用户权限 685
11.9.3 配置定时备份告警与审计信息 685
11.9.4 修改 FusionInsight Manager 添加的路由表 686
11.9.5 切换维护模式 689
11.9.6 例行维护 690
11.10 日志管理 692
11.10.1 关于日志 692
11.10.2 Manager 日志清单 708
11.10.3 配置日志级别与文件大小 716
11.10.4 配置审计日志本地备份数 718
11.10.5 查看角色实例日志 719
11.11 备份恢复管理 720
11.11.1 备份恢复简介 720
11.11.2 备份数据 725
11.11.2.1 备份 Manager 数据 726
11.11.2.2 备份 CDL 数据 729
11.11.2.3 备份 ClickHouse 元数据 730
11.11.2.4 备份 ClickHouse 业务数据 732
11.11.2.5 备份 DBService 数据 735
11.11.2.6 备份 HBase 元数据 738
11.11.2.7 备份 HBase 业务数据 741

11.11.2.8 备份 NameNode 数据 745
11.11.2.9 备份 HDFS 业务数据 748
11.11.2.10 备份 Hive 业务数据 753
11.11.2.11 备份 IoTDB 元数据 757
11.11.2.12 备份 IoTDB 业务数据 760
11.11.2.13 备份 Kafka 元数据 762
11.11.3 恢复数据 765
11.11.3.1 恢复 Manager 数据 765
11.11.3.2 恢复 CDL 数据 768
11.11.3.3 恢复 ClickHouse 元数据 770
11.11.3.4 恢复 ClickHouse 业务数据 772
11.11.3.5 恢复 DBService 数据 774
11.11.3.6 恢复 HBase 元数据 777
11.11.3.7 恢复 HBase 业务数据 780
11.11.3.8 恢复 NameNode 数据 783
11.11.3.9 恢复 HDFS 业务数据 786
11.11.3.10 恢复 Hive 业务数据 790
11.11.3.11 恢复 IoTDB 元数据 793
11.11.3.12 恢复 IoTDB 业务数据 796
11.11.3.13 恢复 Kafka 元数据 798
11.11.4 启用集群间拷贝功能 801
11.11.5 管理本地快速恢复任务 801
11.11.6 修改备份任务 802
11.11.7 查看备份恢复任务 803
11.11.8 创建 ClickHouse 备份任务路径选择 RemoteHDFS 时的环境配置 804
11.12 安全管理 805
11.12.1 安全概述 805
11.12.1.1 权限模型 805
11.12.1.2 权限机制 807
11.12.1.3 认证策略 808
11.12.1.4 鉴权策略 809
11.12.1.5 用户账号一览表 811
11.12.1.6 默认权限信息一览 835
11.12.1.7 FusionInsight Manager 安全功能 838
11.12.2 账户管理 838
11.12.2.1 账户安全设置 838
11.12.2.1.1 解锁 LDAP 用户和管理账户 838
11.12.2.1.2 解锁系统内部用户 839
11.12.2.1.3 修改集群组件鉴权配置开关 840
11.12.2.1.4 使用普通模式集群用户在非集群节点登录 842
11.12.2.2 修改系统用户密码 844
11.12.2.2.1 修改 admin 密码 844

11.12.2.2.2 修改操作系统用户密码.....845
11.12.2.3 修改系统内部用户密码.....845
11.12.2.3.1 修改 Kerberos 管理员密码.....845
11.12.2.3.2 修改 OMS Kerberos 管理员密码.....846
11.12.2.3.3 修改 LDAP 管理员和 LDAP 用户密码（含 OMS LDAP）.....847
11.12.2.3.4 修改 LDAP 管理账户密码.....848
11.12.2.3.5 修改组件运行用户密码.....849
11.12.2.4 修改默认数据库用户密码.....850
11.12.2.4.1 修改 OMS 数据库管理员密码.....851
11.12.2.4.2 修改 OMS 数据库访问用户密码.....851
11.12.2.4.3 修改组件数据库用户密码.....852
11.12.2.4.4 重置组件数据库用户密码.....853
11.12.2.4.5 修改 DBService 数据库 compdbuser 用户密码.....853
11.12.3 安全加固.....854
11.12.3.1 加固策略.....854
11.12.3.2 配置受信任 IP 访问 LDAP.....855
11.12.3.3 加密 HFile 和 WAL 内容.....858
11.12.3.4 配置 Hadoop 安全参数.....862
11.12.3.5 配置 HBase 允许修改操作的 IP 地址白名单.....865
11.12.3.6 更新集群密钥.....865
11.12.3.7 加固 LDAP.....866
11.12.3.8 配置 Kafka 数据传输加密.....867
11.12.3.9 配置 HDFS 数据传输加密.....868
11.12.3.10 配置 Controller 与 Agent 间通信加密.....870
11.12.3.11 更新 omm 用户 ssh 密钥.....871
11.12.4 安全维护.....872
11.12.4.1 账户维护建议.....872
11.12.4.2 密码维护建议.....872
11.12.4.3 日志维护建议.....873
11.12.5 安全声明.....873
12 MRS Manager 操作指导（适用于 2.x 及之前）.....875
12.1 MRS Manager 简介.....875
12.2 查看集群运行任务.....877
12.3 监控管理.....878
12.3.1 系统概览.....878
12.3.2 管理服务和主机监控.....880
12.3.3 管理资源分布.....884
12.3.4 配置监控指标转储.....884
12.4 告警管理.....885
12.4.1 查看与手动清除告警.....885
12.4.2 配置监控与告警阈值.....886
12.4.3 配置 Syslog 北向参数.....888

12.4.4 配置 SNMP 北向参数 890
12.5 对象管理 892
12.5.1 对象管理简介 892
12.5.2 查看配置 893
12.5.3 管理服务操作 893
12.5.4 配置服务参数 894
12.5.5 配置服务自定义参数 895
12.5.6 同步服务配置 896
12.5.7 管理角色实例操作 897
12.5.8 配置角色实例参数 897
12.5.9 同步角色实例配置 898
12.5.10 退服和入服务角色实例 899
12.5.11 管理主机操作 899
12.5.12 隔离主机 900
12.5.13 取消隔离主机 900
12.5.14 启动及停止集群 901
12.5.15 同步集群配置 901
12.5.16 导出集群的配置数据 902
12.6 日志管理 902
12.6.1 关于日志 902
12.6.2 Manager 日志清单 913
12.6.3 查看及导出审计日志 920
12.6.4 导出服务日志 922
12.6.5 配置审计日志转储参数 922
12.7 健康检查管理 924
12.7.1 执行健康检查 924
12.7.2 查看并导出检查报告 925
12.7.3 配置健康检查报告保存数 926
12.7.4 管理健康检查报告 926
12.7.5 DBService 健康检查指标项说明 927
12.7.6 Flume 健康检查指标项说明 927
12.7.7 HBase 健康检查指标项说明 927
12.7.8 Host 健康检查指标项说明 928
12.7.9 HDFS 健康检查指标项说明 934
12.7.10 Hive 健康检查指标项说明 934
12.7.11 Kafka 健康检查指标项说明 935
12.7.12 KrbServer 健康检查指标项说明 936
12.7.13 LdapServer 健康检查指标项说明 936
12.7.14 Loader 健康检查指标项说明 937
12.7.15 MapReduce 健康检查指标项说明 938
12.7.16 OMS 健康检查指标项说明 939
12.7.17 Spark 健康检查指标项说明 942

12.7.18 Storm 健康检查指标项说明 942
12.7.19 Yarn 健康检查指标项说明 943
12.7.20 ZooKeeper 健康检查指标项说明 943
12.8 静态服务池管理 944
12.8.1 查看静态服务池状态 944
12.8.2 配置静态服务池 946
12.9 租户管理 948
12.9.1 租户简介 948
12.9.2 添加租户 949
12.9.3 添加子租户 951
12.9.4 删除租户 953
12.9.5 管理租户目录 954
12.9.6 恢复租户数据 955
12.9.7 添加资源池 956
12.9.8 修改资源池 957
12.9.9 删除资源池 957
12.9.10 配置队列 958
12.9.11 配置资源池的队列容量策略 959
12.9.12 清除队列配置 959
12.10 备份与恢复 960
12.10.1 备份与恢复简介 960
12.10.2 备份元数据 962
12.10.3 恢复元数据 963
12.10.4 修改备份任务 965
12.10.5 查看备份恢复任务 966
12.11 安全管理 967
12.11.1 未开启 Kerberos 认证集群中的默认用户清单 967
12.11.2 开启 Kerberos 认证集群中的默认用户清单 970
12.11.3 修改操作系统用户密码 975
12.11.4 修改 admin 密码 975
12.11.5 修改 Kerberos 管理员密码 977
12.11.6 修改 LDAP 管理员和 LDAP 用户密码 978
12.11.7 修改组件运行用户密码 979
12.11.8 修改 OMS 数据库管理员密码 980
12.11.9 修改 OMS 数据库数据访问用户密码 980
12.11.10 修改组件数据库用户密码 981
12.11.11 更换 HA 证书 982
12.11.12 更新集群密钥 983
12.12 权限管理 984
12.12.1 创建角色 984
12.12.2 创建用户组 989
12.12.3 创建用户 990

12.12.4 修改用户信息 992
12.12.5 锁定用户 992
12.12.6 解锁用户 993
12.12.7 删除用户 993
12.12.8 修改操作用户密码 993
12.12.9 初始化系统用户密码 994
12.12.10 下载用户认证文件 995
12.12.11 修改密码策略 996
12.13 MRS 多用户权限管理 997
12.13.1 MRS 集群中的用户与权限 997
12.13.2 开启 Kerberos 认证集群中的默认用户清单 1001
12.13.3 创建角色 1006
12.13.4 创建用户组 1011
12.13.5 创建用户 1013
12.13.6 修改用户信息 1014
12.13.7 锁定用户 1015
12.13.8 解锁用户 1015
12.13.9 删除用户 1016
12.13.10 修改操作用户密码 1016
12.13.11 初始化系统用户密码 1017
12.13.12 下载用户认证文件 1018
12.13.13 修改密码策略 1019
12.13.14 配置跨集群互信 1020
12.13.15 配置并使用互信集群的用户 1024
12.13.16 配置 MRS 多用户访问 OBS 细粒度权限 1025
12.14 补丁操作指导 1029
12.14.1 MRS 1.7.0 前版本补丁操作指导 1029
12.14.2 支持滚动补丁 1031
12.15 修复隔离主机补丁 1033
12.16 支持滚动重启 1033
13 安全性说明 1037
13.1 集群（未启用 Kerberos 认证）安全配置建议 1037
13.2 安全认证原理和认证机制 1037
14 高危操作一览表 1041
15 故障排除 1063
15.1 Web 页面访问类 1063
15.1.1 无法访问 MRS 集群管理页面（MRS Manager 界面） 1063
15.1.2 升级 Python 后，无法登录 MRS Manager 页面 1064
15.1.3 用户修改域名后无法登录 MRS Manager 页面 1064
15.1.4 登录 Manager，页面空白不显示 1066
15.1.5 用户名过长时下载认证凭据失败 1066

15.2 集群管理类 1067
15.2.1 缩容 Task 节点失败 1067
15.2.2 MRS 集群添加新磁盘 1068
15.2.3 MRS 集群更换磁盘（适用于 2.x 及之前） 1071
15.2.4 MRS 集群更换磁盘（适用于 3.x） 1074
15.2.5 MRS 备份失败 1077
15.2.6 Core 节点出现 df 显示的容量和 du 显示的容量不一致 1078
15.2.7 如何解除关联子网 1078
15.2.8 修改 hostname，导致 MRS 状态异常 1079
15.2.9 如何定位进程被 kill 1079
15.2.10 MRS 集群使用 pip3 安装 python 包提示网络不可达 1081
15.2.11 MRS 集群客户端无法下载 1082
15.2.12 扩容失败 1082
15.2.13 MRS 通过 beeline 执行插入命令的时候出错 1083
15.2.14 MRS 集群如何进行 Euleros 系统漏洞升级？ 1084
15.2.15 使用 CDM 迁移数据至 HDFS 1087
15.2.16 MRS 集群频繁产生告警 1088
15.2.17 PMS 进程占用内存高问题处理 1090
15.2.18 Knox 进程占用内存高 1091
15.2.19 安全集群外节点安装客户端访问 HBase 很慢 1092
15.2.20 作业无法提交如何定位？ 1093
15.2.21 HBase 日志文件过大导致 OS 盘空间不足 1096
15.2.22 Manager 页面新建的租户删除失败 1097
15.3 使用 Alluixo 1097
15.3.1 Alluixo 在 HA 模式下出现 Does not contain a valid host:port authority 报错 1097
15.4 使用 ClickHouse 1098
15.4.1 ZooKeeper 上数据错乱导致 ClickHouse 启动失败问题 1098
15.5 使用 DBservice 1100
15.5.1 DBServer 实例状态异常 1100
15.5.2 DBServer 实例一直处于 Restoring 状态 1101
15.5.3 默认端口 20050 或 20051 被占用 1102
15.5.4 /tmp 目录权限不对导致 DBserver 实例状态一直处于 Restoring 1102
15.5.5 DBService 备份失败 1104
15.5.6 DBService 状态正常，组件无法连接 DBService 1104
15.5.7 DBServer 启动失败 1105
15.5.8 浮动 IP 不通导致 DBService 备份失败 1106
15.5.9 DBService 配置文件丢失导致启动失败 1107
15.6 使用 Flink 1109
15.6.1 安装客户端执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file :"security.kerberos.login.keytab" 1109
15.6.2 安装客户端修改配置后执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file 1110
15.6.3 创建 Flink 集群时执行 yarn-session.sh 命令失败 1111

15.6.4 使用不同用户，执行 yarn-session 创建集群失败 1112
15.6.5 Flink 业务程序无法读取 NFS 盘上的文件 1113
15.6.6 自定义 Flink log4j 日志输出级别 1114
15.7 使用 Flume 1114
15.7.1 Flume 向 Spark Streaming 提交作业，提交到集群后报类找不到 1114
15.7.2 Flume 客户端安装失败 1115
15.7.3 Flume 客户端无法连接服务端 1115
15.7.4 Flume 数据写入组件失败 1116
15.7.5 Flume 服务端进程故障 1117
15.7.6 Flume 数据采集慢 1117
15.7.7 Flume 启动失败 1118
15.8 使用 HBase 1119
15.8.1 连接到 HBase 响应慢 1119
15.8.2 HBase 用户认证失败 1119
15.8.3 端口被占用导致 RegionServer 启动失败 1120
15.8.4 节点剩余内存不足导致 HBase 启动失败 1121
15.8.5 HDFS 性能差导致 HBase 服务不可用告警 1121
15.8.6 参数不合理导致 HBase 启动失败 1122
15.8.7 残留进程导致 Regionsever 启动失败 1123
15.8.8 HDFS 上设置配额导致 HBase 启动失败 1123
15.8.9 HBase version 文件损坏导致启动失败 1124
15.8.10 无业务情况下，RegionServer 占用 CPU 高 1124
15.8.11 HBase 启动失败，RegionServer 日志中提示 FileNotFoundException 异常 1126
15.8.12 HBase 启动后原生页面显示 RegionServer 个数多于实际个数 1127
15.8.13 RegionServer 实例异常，处于 Restoring 状态 1128
15.8.14 新安装的集群 HBase 启动失败 1129
15.8.15 acl 表目录丢失导致 HBase 启动失败 1129
15.8.16 集群上下电之后 HBase 启动失败 1130
15.8.17 文件块过大导致 HBase 数据导入失败 1131
15.8.18 使用 Phoenix 创建 HBase 表后，向索引表中加载数据报错 1132
15.8.19 在 MRS 集群客户端无法执行 **hbase shell** 命令 1133
15.8.20 HBase shell 客户端在使用中有 INFO 信息打印在控制台导致显示混乱 1134
15.8.21 RegionServer 剩余内存不足导致 HBase 服务启动失败 1135
15.9 使用 HDFS 1136
15.9.1 修改集群 HDFS 服务的 NameNode RPC 端口后，NameNode 都变为备状态 1136
15.9.2 通过公网 IP 连接主机，使用 HDFS 客户端报错 1137
15.9.3 使用 Python 远程连接 HDFS 的端口失败 1138
15.9.4 HDFS 容量使用达到 100%，导致上层服务 HBase、Spark 等上报服务不可用 1138
15.9.5 启动 HDFS 和 Yarn 报错 1139
15.9.6 HDFS 权限设置问题 1140
15.9.7 HDFS 的 DataNode 一直显示退服中 1141
15.9.8 内存不足导致 HDFS 启动失败 1143

15.9.9 ntpdate 修改时间导致 HDFS 出现大量丢块 1145
15.9.10 DataNode 概率性出现 CPU 占用接近 100%，导致节点丢失（ssh 连得很慢或者连不上） 1147
15.9.11 单 NameNode 长期故障，如何使用客户端手动 checkpoint 1148
15.9.12 文件读写常见故障 1149
15.9.13 文件最大打开句柄数设置太小导致读写文件异常 1150
15.9.14 客户端写文件 close 失败 1151
15.9.15 文件错误导致上传文件到 HDFS 失败 1153
15.9.16 界面配置 dfs.blocksize 后 put 数据，block 大小还是原来的大小 1153
15.9.17 读取文件失败，FileNotFoundException 1154
15.9.18 HDFS 写文件失败，item limit of / is exceeded 1155
15.9.19 调整 shell 客户端日志级别 1155
15.9.20 读文件失败 No common protection layer 1155
15.9.21 HDFS 目录配额（quota）不足导致写文件失败 1156
15.9.22 执行 balance 失败，Source and target differ in block-size 1157
15.9.23 查询或者删除文件失败，父目录可以看见此文件（不可见字符） 1158
15.9.24 非 HDFS 数据残留导致数据分布不均衡 1159
15.9.25 客户端安装在数据节点导致数据分布不均衡 1160
15.9.26 节点内 DataNode 磁盘使用率不均衡处理指导 1160
15.9.27 执行 balance 常见问题定位方法 1161
15.9.28 HDFS 显示磁盘空间不足，其实还有 10%磁盘空间 1162
15.9.29 普通集群在 Core 节点安装 hdfs 客户端，使用时报错 1162
15.9.30 集群外节点安装客户端使用 hdfs 上传文件失败 1163
15.9.31 HDFS 写并发较大时，报副本不足的问题 1164
15.9.32 HDFS 客户端无法删除超长目录 1164
15.9.33 集群外节点访问 MRS HDFS 报错 1166
15.10 使用 Hive 1167
15.10.1 Hive 各个日志里都存放了什么信息？ 1167
15.10.2 Hive 启动失败问题的原因有哪些？ 1168
15.10.3 安全集群执行 set 命令的时候报 Cannot modify xxx at runtime 1168
15.10.4 怎样在 Hive 提交任务的时候指定队列？ 1170
15.10.5 客户端怎么设置 Map/Reduce 内存？ 1170
15.10.6 如何在导入表时指定输出的文件压缩格式 1171
15.10.7 desc 描述表过长时，无法显示完整 1171
15.10.8 增加分区列后再 insert 数据显示为 NULL 1172
15.10.9 创建新用户，执行查询时报无权限 1173
15.10.10 执行 SQL 提交任务到指定队列报错 1174
15.10.11 执行 load data inpath 命令报错 1174
15.10.12 执行 load data local inpath 命令报错 1175
15.10.13 执行 create external table 报错 1176
15.10.14 在 beeline 客户端执行 dfs -put 命令报错 1176
15.10.15 执行 set role admin 报无权限 1177
15.10.16 通过 beeline 创建 UDF 时候报错 1178

15.10.17 Hive 服务健康状态和 Hive 实例健康状态的区别 1178
15.10.18 Hive 中的告警有哪些以及触发的场景 1179
15.10.19 Shell 客户端连接提示"authentication failed" 1180
15.10.20 客户端提示访问 ZooKeeper 失败 1180
15.10.21 使用 udf 函数提示"Invalid function" 1182
15.10.22 Hive 服务状态为 Unknown 总结 1182
15.10.23 Hiveserver 或者 Metastore 实例的健康状态为 unknown 1182
15.10.24 Hiveserver 或者 Metastore 实例的健康状态为 Concerning 1183
15.10.25 TEXTFILE 类型文件使用 ARC4 压缩时 select 结果乱码 1183
15.10.26 hive 任务运行过程中失败，重试成功 1184
15.10.27 执行 select 语句报错 1184
15.10.28 drop partition 操作，有大量分区时操作失败 1186
15.10.29 localtask 启动失败 1186
15.10.30 WebHCat 启动失败 1188
15.10.31 切域后 Hive 二次开发样例代码报错 1188
15.10.32 DBService 超过最大连接数，导致 metastore 异常 1189
15.10.33 beeline 报 Failed to execute session hooks: over max connections 错误 1190
15.10.34 beeline 报 OutOfMemoryError 错误 1192
15.10.35 输入文件数超出设置限制导致任务执行失败 1193
15.10.36 任务执行中报栈内存溢出导致任务执行失败 1194
15.10.37 对同一张表或分区并发写数据导致任务失败 1195
15.10.38 Hive 任务失败，报没有 HDFS 目录的权限 1196
15.10.39 Load 数据到 Hive 表失败 1197
15.10.40 HiveServer 和 HiveHCat 进程故障 1197
15.10.41 Hive 执行 insert into 语句报错，命令界面报错信息不明 1198
15.10.42 增加 Hive 表字段超时 1200
15.10.43 Hive 服务重启失败 1202
15.10.44 hive 执行删除表失败 1202
15.10.45 Hive 执行 msck repair table table_name 报错 1203
15.10.46 在 Hive 中 drop 表后，如何完全释放磁盘空间 1204
15.10.47 客户端执行 SQL 报错连接超时 1204
15.10.48 WebHCat 健康状态异常导致启动失败 1205
15.10.49 mapred-default.xml 文件解析异常导致 WebHCat 启动失败 1206
15.11 使用 Hue 1207
15.11.1 Hue 上有 job 在运行 1207
15.11.2 使用 IE 浏览器在 Hue 中执行 HQL 失败 1207
15.11.3 Hue（主）无法打开 web 网页 1208
15.11.4 Hue WebUI 访问失败 1208
15.11.5 Hue 界面无法加载 HBase 表 1209
15.12 使用 Impala 1210
15.12.1 用户连接 impala-shell 失败 1210
15.12.2 创建 Kudu 表报错 1211

15.12.3 Impala 客户端登录失败 1211
15.13 使用 Kafka 1213
15.13.1 运行 Kafka 获取 topic 报错 1213
15.13.2 Flume 可以正常连接 Kafka，但是发送消息失败。 1214
15.13.3 Producer 发送数据失败，抛出 NullPointerException 1215
15.13.4 Producer 发送数据失败，抛出 TOPIC_AUTHORIZATION_FAILED 1217
15.13.5 Producer 偶现发送数据失败，日志提示 Too many open files in system 1220
15.13.6 Consumer 初始化成功，但是无法从 Kafka 中获取指定 Topic 消息 1222
15.13.7 Consumer 消费数据失败，Consumer 一直处于等待状态 1226
15.13.8 SparkStreaming 消费 Kafka 消息失败，提示 Error getting partition metadata 1228
15.13.9 新建集群 Consumer 消费数据失败，提示 GROUP_COORDINATOR_NOT_AVAILABLE 1230
15.13.10 SparkStreaming 消费 Kafka 消息失败，提示 Couldn't find leader offsets 1231
15.13.11 Consumer 消费数据失败，提示 SchemaException: Error reading field 'brokers' 1233
15.13.12 Consumer 消费数据是否丢失排查 1234
15.13.13 账号锁定导致启动组件失败 1234
15.13.14 Kafka Broker 上报进程异常，日志提示 IllegalArgumentException 1235
15.13.15 执行 Kafka Topic 删除操作，发现无法删除 1236
15.13.16 执行 Kafka Topic 删除操作，提示 AdminOperationException 1238
15.13.17 执行 Kafka Topic 创建操作，发现无法创建提示 NoAuthException 1240
15.13.18 执行 Kafka Topic 设置 ACL 操作失败，提示 NoAuthException 1241
15.13.19 执行 Kafka Topic 创建操作，发现无法创建提示 NoNode for /brokers/ids 1243
15.13.20 执行 Kakfa Topic 创建操作，发现无法创建提示 replication factor larger than available brokers 1244
15.13.21 Consumer 消费数据存在重复消费现象 1245
15.13.22 执行 Kafka Topic 创建操作，发现 Partition 的 Leader 显示为 none 1247
15.13.23 Kafka 安全使用说明 1248
15.13.24 如何获取 Kafka Consumer Offset 信息 1252
15.13.25 如何针对 Topic 进行配置增加和删除 1254
15.13.26 如何读取“__consumer_offsets”内部 topic 的内容 1255
15.13.27 如何配置客户端 shell 命令的日志 1256
15.13.28 如何获取 Topic 的分布信息 1257
15.13.29 Kafka 高可靠使用说明 1258
15.13.30 Kafka 生产者写入单条记录过长问题 1261
15.13.31 Kakfa 消费者读取单条记录过长问题 1261
15.13.32 Kafka 集群节点内多磁盘数据量占用高处理办法 1262
15.14 使用 Oozie 1265
15.14.1 当并发提交大量 oozie 任务时，任务一直没有运行 1265
15.15 使用 Presto 1265
15.15.1 配置 sql-standard-with-group 创建 schema 失败报 Access Denied 1266
15.15.2 Presto 的 coordinator 无法正常启动 1267
15.15.3 Presto 查询 Kudu 表报错 1268
15.15.4 Presto 查询 Hive 表无数据 1269
15.16 使用 Spark 1270

15.16.1 Spark 应用下修改 split 值时报错.....1270
15.16.2 使用 Spark 时报错.....1271
15.16.3 引入 jar 包不正确，导致 Spark 任务无法运行.....1271
15.16.4 Spark 任务由于内存不够，作业卡住.....1272
15.16.5 运行 Spark 报错.....1273
15.16.6 Driver 端提示 executor memory 超限.....1274
15.16.7 Yarn-cluster 模式下，Can't get the Kerberos realm 异常.....1275
15.16.8 JDK 版本不匹配启动 spark-sql，spark-shell 失败.....1277
15.16.9 Yarn-client 模式提交 ApplicationMaster 尝试启动两次失败.....1277
15.16.10 提交 Spark 任务时，连接 ResourceManager 异常.....1278
15.16.11 DataArts Studio 调度 spark 作业失败.....1279
15.16.12 Spark 作业 api 提交状态为 error.....1280
15.16.13 集群反复出现 43006 告警.....1281
15.16.14 在 spark-beeline 中创建/删除表失败.....1281
15.16.15 集群外节点提交 Spark 作业到 Yarn 报错连不上 Driver.....1283
15.16.16 运行 Spark 任务发现大量 shuffle 结果丢失.....1284
15.16.17 JDBCServer 长时间运行导致磁盘空间不足.....1285
15.16.18 spark-shell 执行 sql 跨文件系统 load 数据到 hive 表失败.....1286
15.16.19 Spark 任务提交失败.....1286
15.16.20 Spark 任务运行失败.....1287
15.16.21 JDBCServer 连接失败.....1288
15.16.22 查看 Spark 任务日志失败.....1288
15.16.23 Spark 连接其他服务认证问题.....1289
15.16.24 spark 连接 redis 报错.....1289
15.16.25 spark-beeline 查询 Hive 视图报错.....1291
15.17 使用 Sqoop.....1292
15.17.1 Sqoop 如何连接 mysql.....1292
15.17.2 Sqoop 读取 MySQL 中数据到 HBase 报 HBaseAdmin.<init>方法找不到异常.....1293
15.17.3 HUE 界面的 Sqoop 任务 HBase 到 HDFS 报错.....1294
15.17.4 Sqoop 从 hive 到 mysql8.0 报格式错误.....1297
15.17.5 Sqoop import 从 pg 到 hive 报错.....1299
15.17.6 Sqoop 读 mysql，写 parquet 文件到 OBS 失败.....1299
15.18 使用 Storm.....1300
15.18.1 Storm 组件的 Storm UI 页面中 events 超链接地址无效.....1300
15.18.2 提交拓扑失败.....1302
15.18.3 提交拓扑失败，提示 Failed to check principle for keytab.....1303
15.18.4 提交拓扑后 Worker 日志为空.....1304
15.18.5 提交拓扑后 Worker 运行异常，日志提示 Failed to bind to：host:ip.....1306
15.18.6 使用 jstack 命令查看进程堆栈提示 well-known file is not secure.....1307
15.18.7 使用 Storm-JDBC 插件开发 Oracle 写入 Bolt，发现数据无法写入.....1310
15.18.8 业务拓扑配置 GC 参数不生效.....1311
15.18.9 UI 查看信息显示 Internal Server Error.....1313

15.19 使用 Ranger 1313
15.19.1 Hive 启用 Ranger 鉴权后，在 Hue 页面能查看到没有权限的表和库 1314
15.20 使用 Yarn 1315
15.20.1 启动 Yarn 后发现一堆 job 1315
15.20.2 通过客户端 hadoop jar 命令提交任务，客户端返回 GC overhead 1316
15.20.3 Yarn 汇聚日志过大导致磁盘被占满 1317
15.20.4 MR 任务异常临时文件不删除 1318
15.20.5 提交任务的 Yarn 的 ResourceManager 报错 connection refused，且配置的 Yarn 端口为 8032 1319
15.20.6 Yarn WebUI 作业查看日志提示“Could not access logs page!” 1320
15.20.7 Yarn 页面单击队列名称报错 1321
15.21 使用 ZooKeeper 1321
15.21.1 MRS 集群如何访问 ZooKeeper 1321
15.22 访问 OBS 1322
15.22.1 使用 MRS 多用户访问 OBS 功能时/tmp 目录没有权限 1322
15.22.2 Hadoop 客户端删除 OBS 上数据时.Trash 目录没有权限 1323
16 附录 1325
16.1 MRS 所使用的裸金属服务器规格 1325
16.2 数据迁移方案介绍 1325
16.2.1 准备工作 1326
16.2.2 元数据导出 1326
16.2.3 数据拷贝 1327
16.2.4 数据恢复 1328
16.3 MRS 3.x 版本操作注意事项 1329
16.4 安装 Flume 客户端 1330
16.4.1 安装 MRS 3.x 之前版本 Flume 客户端 1330
16.4.2 安装 MRS 3.x 及之后版本 Flume 客户端 1332
17 修订记录 1335

1 简介

1.1 什么是 MRS

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

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

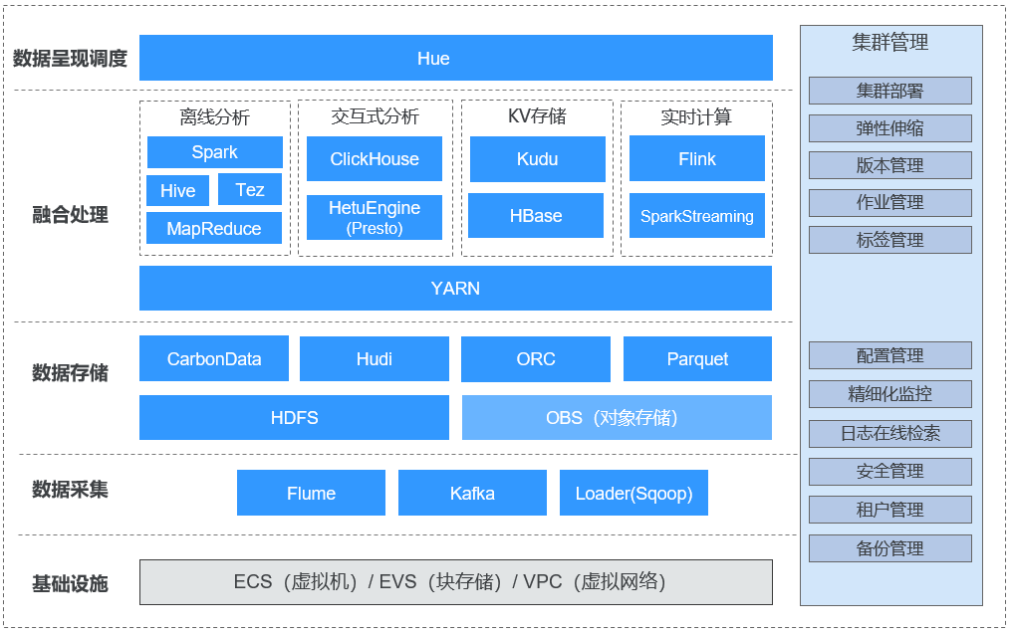
产品架构

MRS逻辑架构如图1-1所示。

说明

MRS 3.x及之后版本暂不支持在管理控制台执行补丁管理操作。

图 1-1 MRS 架构



MRS架构包括了基础设施和大数据处理流程各个阶段的能力。

- 基础设施

  MRS基于弹性云服务器ECS构建的大数据集群，充分利用了其虚拟化层的高可靠、高安全的能力。
  - 虚拟私有云（VPC）为每个租户提供的虚拟内部网络，默认与其他网络隔离。
  - 云硬盘（EVS）提供高可靠、高性能的存储。
  - 弹性云服务器（ECS）提供的弹性可扩展虚拟机，结合VPC、安全组、EVS数据多副本等能力打造一个高效、可靠、安全的计算环境。
- 数据采集

  数据采集层提供了数据接入到MRS集群的能力，包括Flume（数据采集）、Loader（关系型数据导入）、Kafka（高可靠消息队列），支持各种数据源导入数据到大数据集群中。使用云数据迁移云服务也可以将外部数据导入至MRS集群中。
- 数据存储

  MRS支持结构化和非结构化数据在集群中的存储，并且支持多种高效的格式来满足不同计算引擎的要求。
  - HDFS是大数据上通用的分布式文件系统。
  - OBS是对象存储服务，具有高可用低成本的特点。
  - HBase支持带索引的数据存储，适合高性能基于索引查询的场景。
- 数据融合处理
  - MRS提供多种主流计算引擎：MapReduce（批处理）、Tez（DAG模型）、Spark（内存计算）、SparkStreaming（微批流计算）、Storm（流计算）、Flink（流计算），满足多种大数据应用场景，将数据进行结构和逻辑的转换，转化成满足业务目标的数据模型。
  - 基于预设的数据模型，使用易用SQL的数据分析，用户可以选择Hive（数据仓库），SparkSQL以及Presto交互式查询引擎。

- 数据呈现调度

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

- 集群管理

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

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

产品优势

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

MRS具有如下优势：

- 高性能

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

- 低成本

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

- 高安全

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

- 易运维

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

- 高可靠

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

首次使用 MRS

如果您是首次使用MRS的用户，建议您学习并了解如下信息：

- **基础知识了解**

  通过MRS***组件介绍***章节的内容，了解MRS相关的基础知识，包含MRS各组件的基本原理和增强特性介绍，以及MRS服务的特有概念和功能的详细介绍。

- **入门使用**

  您可以参考***快速入门***学习并上手使用MRS。快速入门提供了样例的详细操作指导，您可以基于此操作指导，创建和使用MRS集群。

- **使用更多的功能，并查看其相关操作指导**

  如果您是一个MRS集群使用和运维人员，可以参考*用户指南*完成集群的生命周期管理、扩缩容以及作业管理等操作。集群中组件的使用指导可以详细参考***组件操作指南***。

  如果您是一个开发者，可以通过API调用完成MRS集群管理、作业执行等相关操作，您可以参考***API参考***获取详情。

1.2 应用场景

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

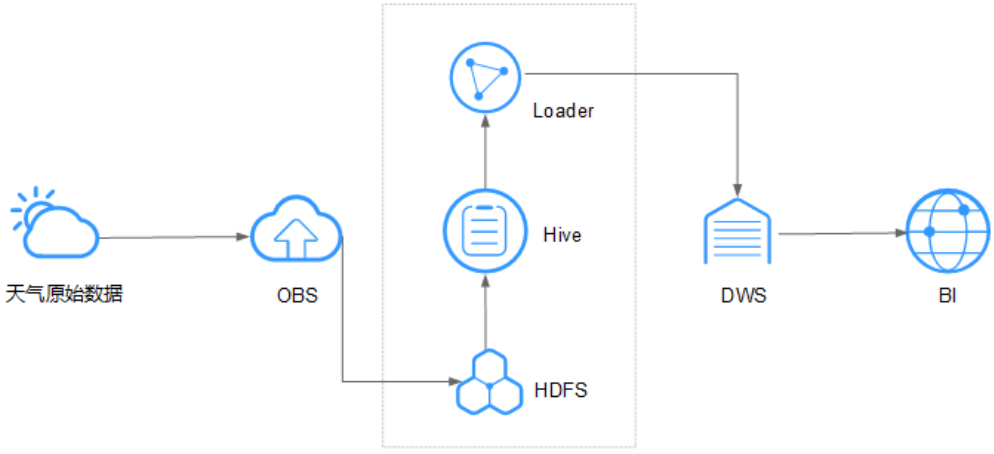
海量数据分析场景

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

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

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

图 1-2 环保行业海量数据分析场景



该场景下MRS的优势如下所示。

- 低成本：利用OBS实现低成本存储。
- 海量数据分析：利用Hive实现TB/PB级的数据分析。

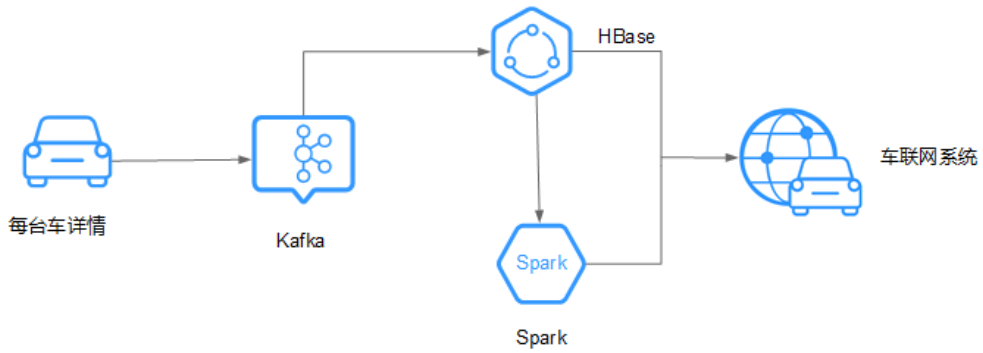
- 可视化的导入导出工具：通过可视化导入导出工具Loader，将数据导出到DWS，完成BI分析。

海量数据存储场景

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

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

图 1-3 车联网行业海量数据存储场景



该场景下MRS的优势如下所示。

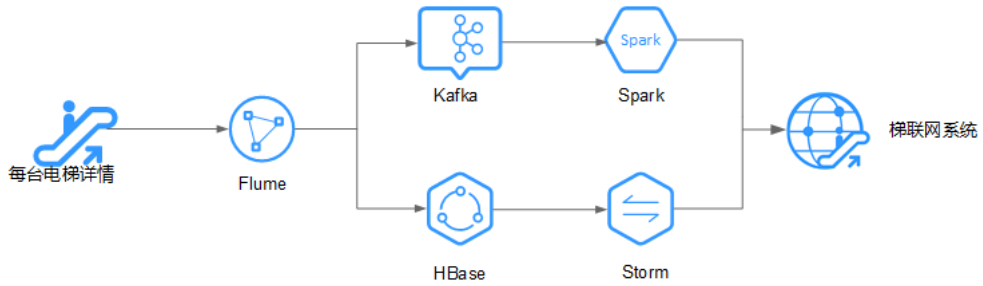
- 实时：利用Kafka实现海量汽车的消息实时接入。
- 海量数据存储：利用HBase实现海量数据存储，并实现毫秒级数据查询。
- 分布式数据查询：利用Spark实现海量数据的分析查询。

实时数据处理

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

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

图 1-4 梯联网行业低时延流式处理场景



该场景下MRS的优势如下所示。

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

1.3 如何选择 MRS

MRS 版本说明

- **普通版**
  - **功能说明**

    普通版支持集群基础操作如配置、管理和运维等。
  - **组件介绍**

    普通版支持Presto、Impala、Kudu、Sqoop等组件，可以根据普通版的不同版本选择不同组件。
- **LTS版**
  - **功能说明**

    LTS版除支持集群基础操作外，还提供版本升级能力。

    LTS版支持多可用区部署。
  - **组件介绍**

    LTS版除支持普通版本支持的组件外，还支持HetuEngine组件。

选择建议

- LTS版集群支持版本升级能力，如果您需要使用版本升级能力，您可以选择购买LTS版集群。
- LTS版集群具备多可用区部署能力，可以实现集群可用区级别的容灾。如果您需要MRS集群具备更高的安全性能和容灾能力，您可以选择购买LTS版集群。
- LTS版集群支持HetuEngine组件，如果您需要使用HetuEngine组件，您可以选择购买LTS版集群。

  说明

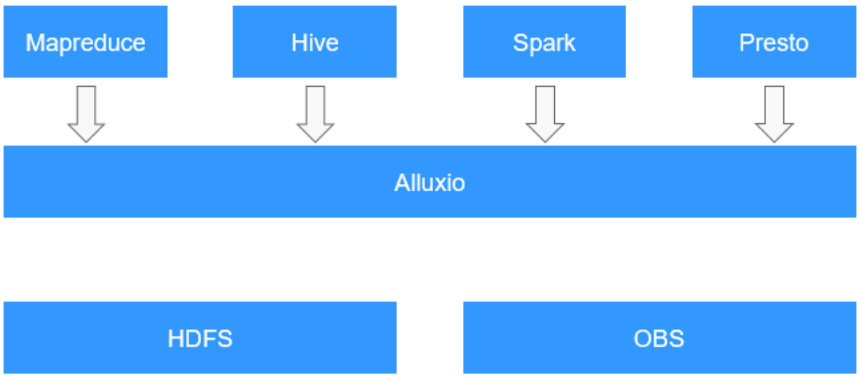
  由于已购买的LTS版集群无法切换为普通版，请根据需要选择购买。

1.4 组件介绍

1.4.1 Alluxio

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

图 1-5 Alluxio 架构



优势：

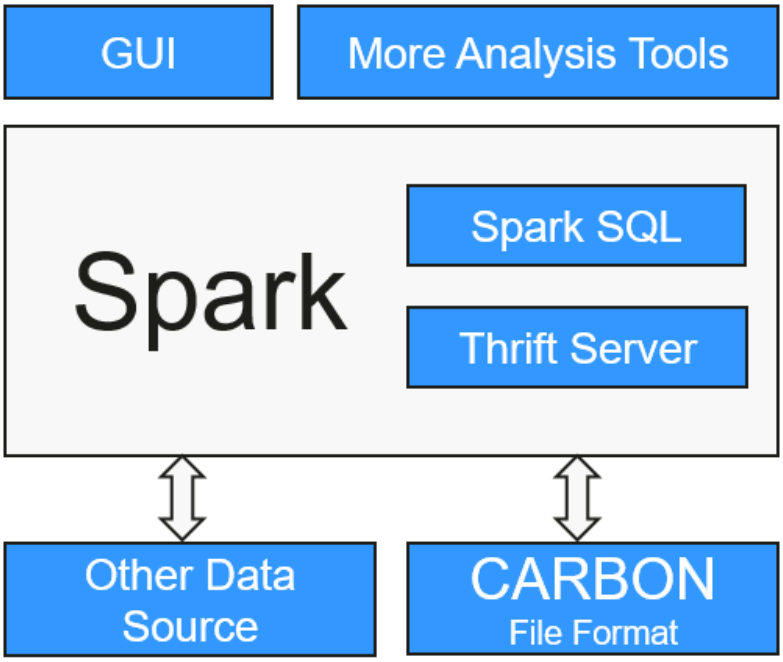
- 提供内存级 I/O 吞吐率，同时降低具有弹性扩张特性的数据驱动型应用的成本开销
- 简化云存储和对象存储接入
- 简化数据管理，提供对多数据源的单点访问
- 应用程序部署简易

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

1.4.2 CarbonData

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

图 1-6 CarbonData 基本架构



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

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

CarbonData特性

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

CarbonData独有的特点

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

- 支持一种数据格式应用于多种用例场景：例如交互式OLAP-style查询，顺序访问（ big scan ），随机访问（ narrow scan ）。

CarbonData关键技术和优势

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

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

1.4.3 ClickHouse

ClickHouse 简介

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

ClickHouse核心的功能特性介绍如下：

完备的DBMS功能

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

- DDL （数据定义语言）：可以动态地创建、修改或删除数据库、表和视图，而无须重启服务。
- DML（数据操作语言）：可以动态查询、插入、修改或删除数据。
- 权限控制：可以按照用户粒度设置数据库或者表的操作权限，保障数据的安全性。
- 数据备份与恢复：提供了数据备份导出与导入恢复机制，满足生产环境的要求。
- 分布式管理：提供集群模式，能够自动管理多个数据库节点。

列式存储与数据压缩

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

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

例如在传统的行式数据库系统中，数据按如下**表1-1**顺序存储：

表 1-1 行式数据库

row	ID	Flag	Name	Event	Time
0	12345678901	0	name1	1	2020/1/11 15:19

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

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

表 1-2 列式数据库

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

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

向量化执行引擎

ClickHouse利用CPU的SIMD指令实现了向量化执行。SIMD的全称是Single Instruction Multiple Data，即用单条指令操作多条数据，通过数据并行以提高性能的一种实现方式（其他的还有指令级并行和线程级并行），它的原理是在CPU寄存器层面实现数据的并行操作。

关系模型与SQL查询

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

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

数据分片与分布式查询

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

ClickHouse提供了本地表（Local Table）与分布式表（Distributed Table）的概念。一张本地表等同于一份数据的分片。而分布式表本身不存储任何数据，它是本地表的

访问代理，其作用类似分库中间件。借助分布式表，能够代理访问多个数据分片，从而实现分布式查询。

ClickHouse 应用场景

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

ClickHouse 开源增强特性

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

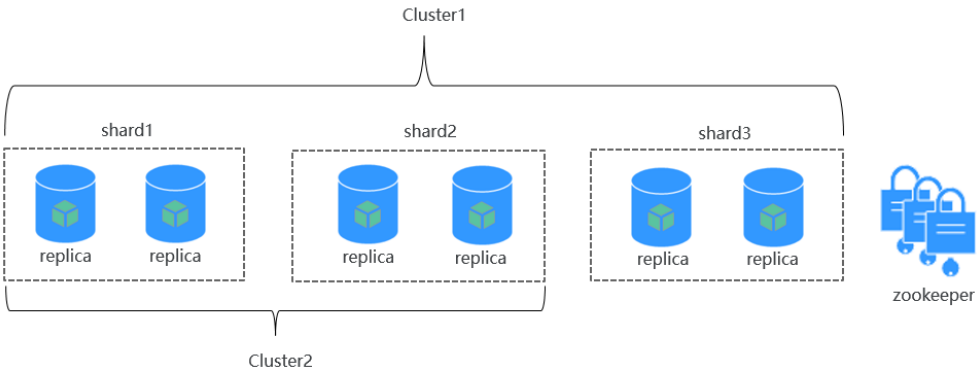
- 手动挡集群模式升级

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

  图 1-7 ClickHouse 集群

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

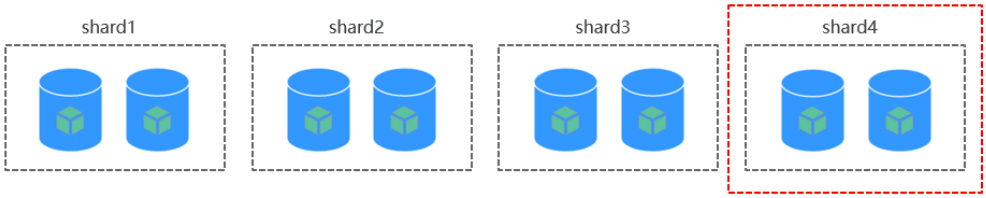
  图 1-8 ClickHouse 的 cluster 结构

- 平滑的弹性扩容能力

  随着业务的快速增长，面对集群存储容量或者CPU计算资源接近极限等场景，MRS服务提供了ClickHouse数据迁移工具，该工具可以将某几个ClickHouseServer实例节点上的一个或多个MergeTree引擎分区表的部分分区迁移至其他ClickHouseServer节点上相同的表中，以便保障业务可用性，实现了更加平滑的扩容能力。

  在用户对集群进行扩容ClickHouse节点时，可以使用该工具将原节点上的部分数据迁移至新增节点上，从而达到扩容后的数据均衡。

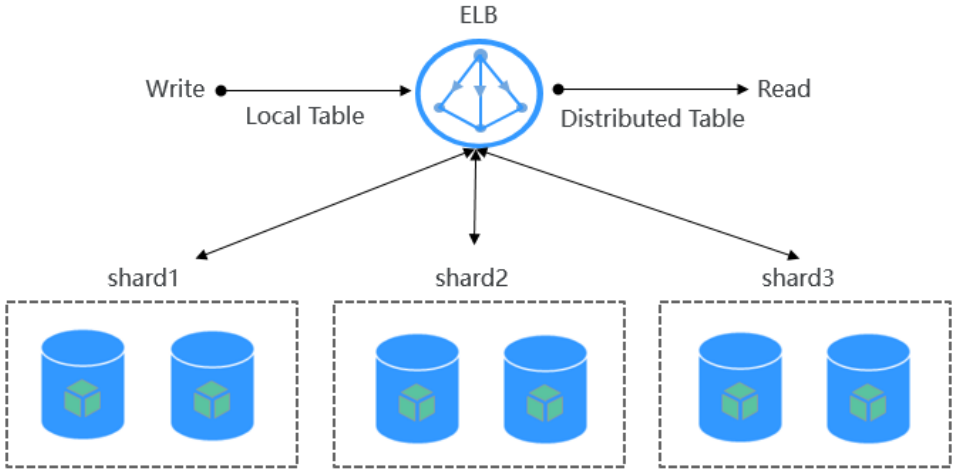


- 高可用HA部署架构

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

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

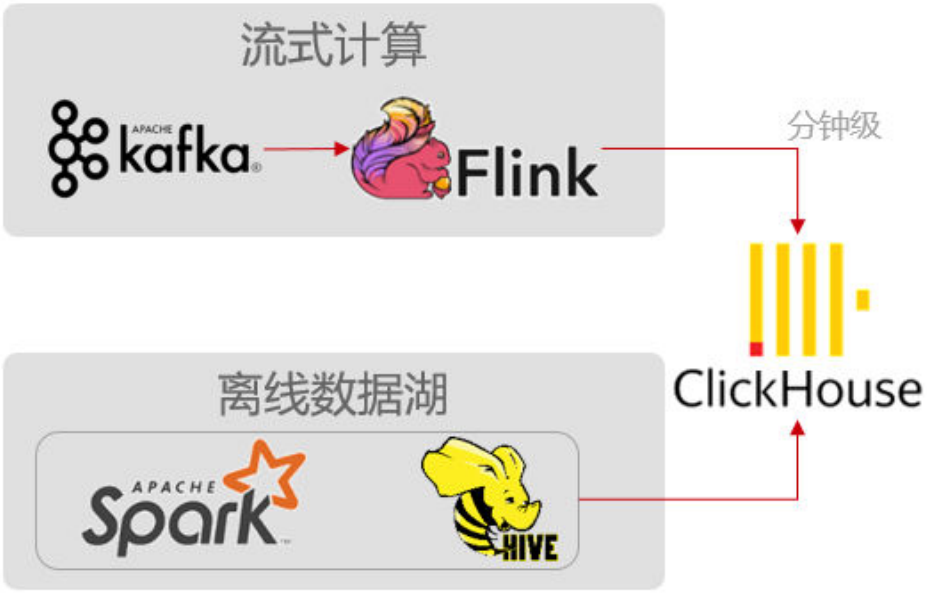
图 1-9 高可用 HA 部署架构图



ClickHouse 与其他组件的关系

ClickHouse安装部署依赖Zookeeper服务。

ClickHouse通过Flink流计算应用加工生成通用的报表数据（明细宽表），准实时写入到ClickHouse，通过Hive/Spark作业加工生成通用的报表数据（明细宽表），批量导入到ClickHouse。



说明

ClickHouse暂不支持对接Kafka普通模式和HDFS安全模式。

1.4.4 DBService

DBService 简介

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

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

DBService 结构

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

其逻辑结构如图1-10所示。

图 1-10 DBService 结构

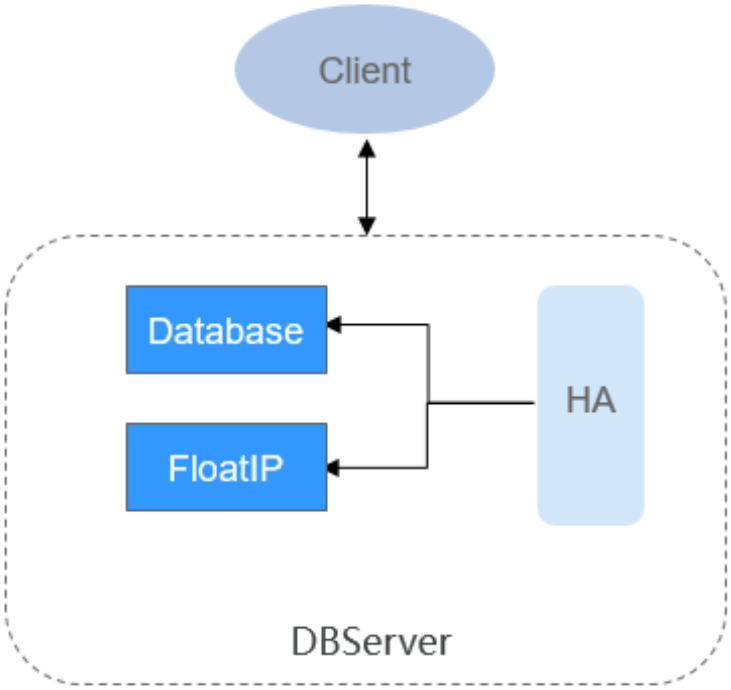


图1-10中各模块的说明如表1-3所示。

表 1-3 模块说明

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

DBService 与其他组件的关系

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

1.4.5 Flink

1.4.5.1 Flink 基本原理

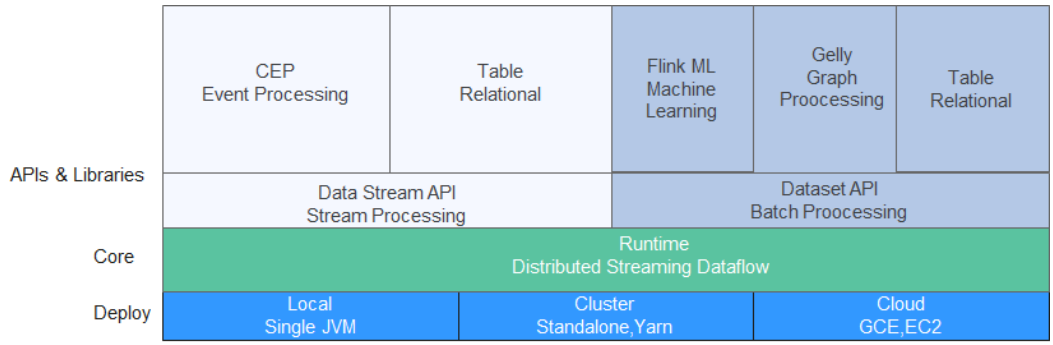
Flink 简介

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

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

Flink技术栈如图1-11所示。

图 1-11 Flink 技术栈



Flink在当前版本中重点构建如下特性：

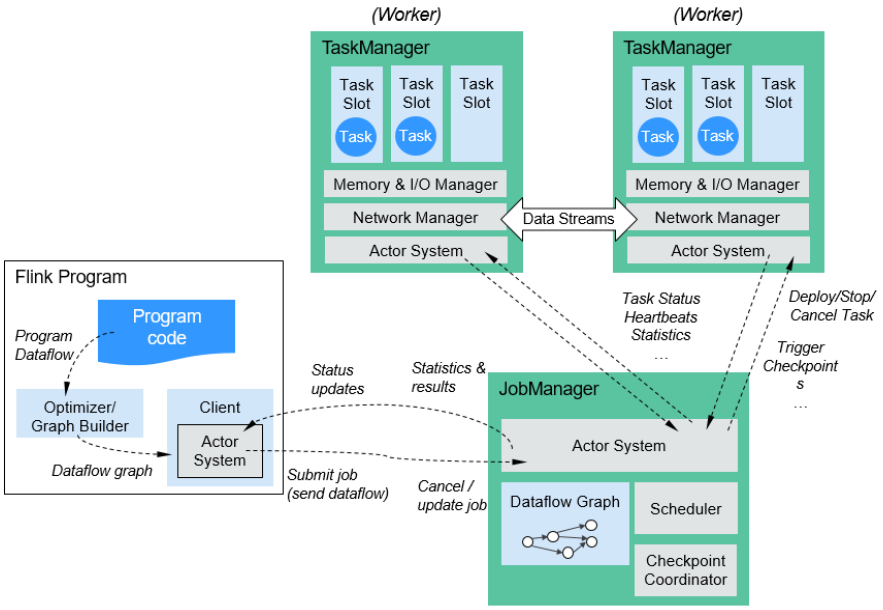
- DataStream
- Checkpoint
- 窗口
- Job Pipeline
- 配置表

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

Flink 结构

Flink结构如图1-12所示。

图 1-12 Flink 结构



Flink整个系统包含三个部分：

- Client

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

- TaskManager

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

- JobManager

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

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

Flink 原理

- **Stream & Transformation & Operator**

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

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

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

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

图 1-13 Flink DataStream 示例

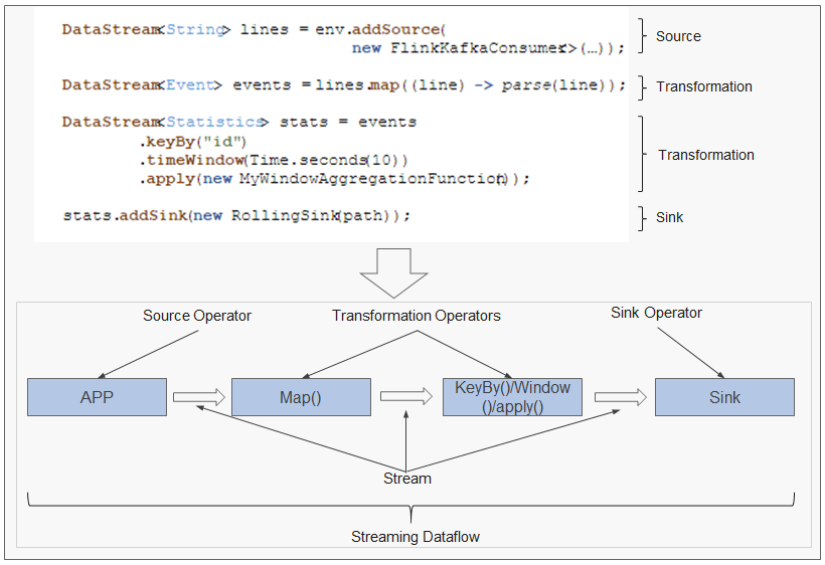


图1-13中“FlinkKafkaConsumer”是一个Source Operator，Map、KeyBy、TimeWindow、Apply是Transformation Operator，RollingSink是一个Sink Operator。

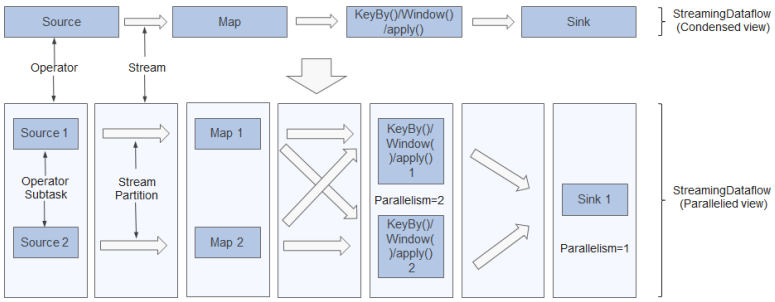
- **Pipeline Dataflow**

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

  Flink内部有一个优化的功能，根据上下游算子的紧密程度来进行优化。

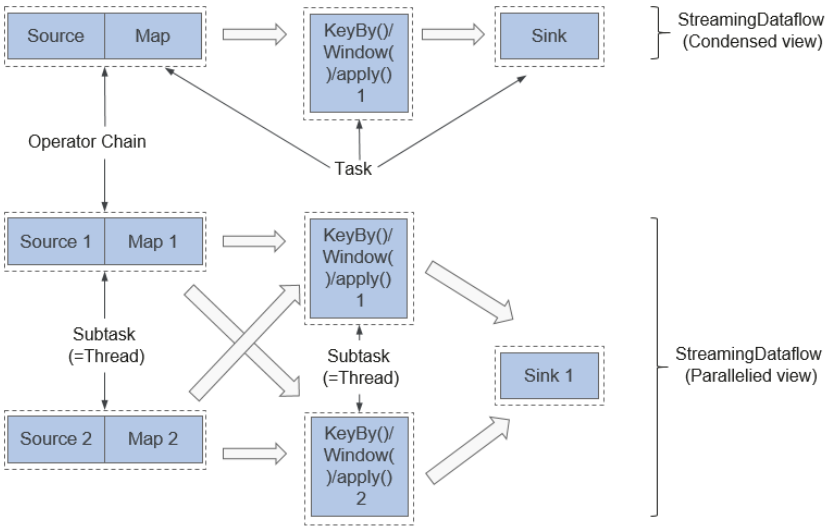
  - 紧密度低的算子则不能进行优化，而是将每一个Operator Subtask放在不同的线程中独立执行。一个Operator的并行度，等于Operator Subtask的个数，一个Stream的并行度（分区总数）等于生成它的Operator的并行度，如图1-14所示。

    图 1-14 Operator

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

图 1-15 Operator chain



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

Flink 关键特性

- 流式处理

  高吞吐、高性能、低时延的实时流处理引擎，能够提供ms级时延处理能力。

- 丰富的状态管理

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

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

- 丰富的时间语义支持

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

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

- 容错机制

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

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

- Flink SQL

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

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

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

- CEP in SQL

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

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

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

1.4.5.2 Flink HA 方案介绍

Flink HA 方案介绍

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

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

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

YARN模式

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

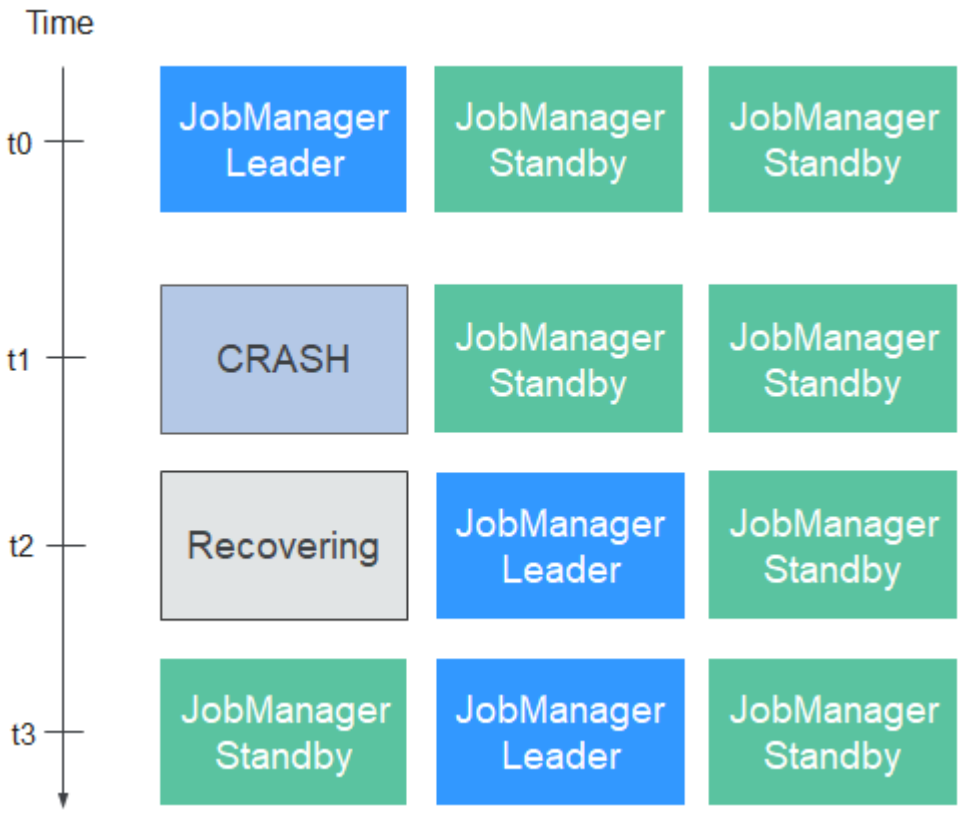
YARN模式的HA方案的更多信息，可参考链接：

http://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/ResourceManagerHA.html

Standalone模式

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

图 1-16 恢复过程



TaskManager恢复

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

JobManager恢复

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

Job恢复

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

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

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

以下场景的异常，都会导致job重新恢复：

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

说明

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

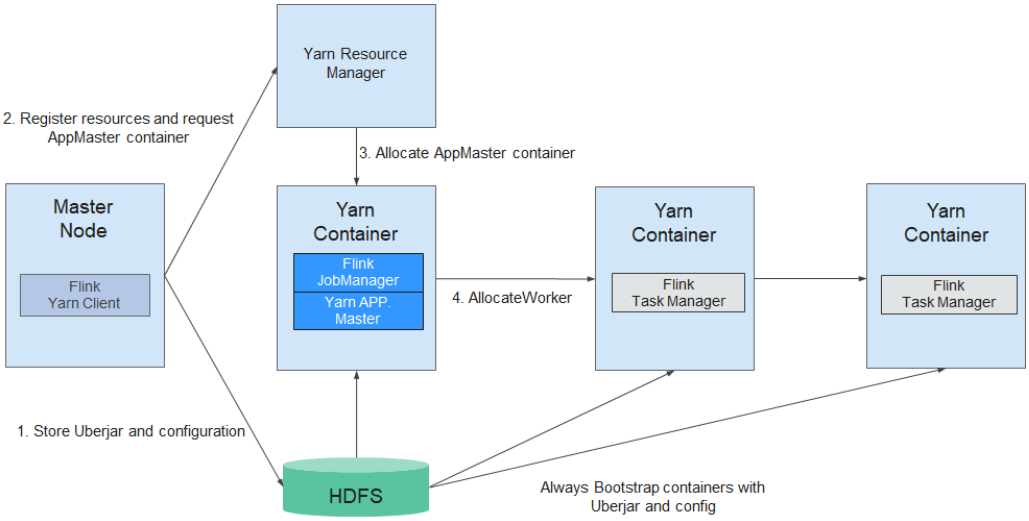
1.4.5.3 Flink 与其他组件的关系

Flink 与 YARN 的关系

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

Flink基于YARN的集群部署如图1-17所示。

图 1-17 Flink 基于 YARN 的集群部署



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

1.4.5.4 Flink 开源增强特性

1.4.5.4.1 窗口

Flink 开源特性增强：窗口

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

窗口介绍

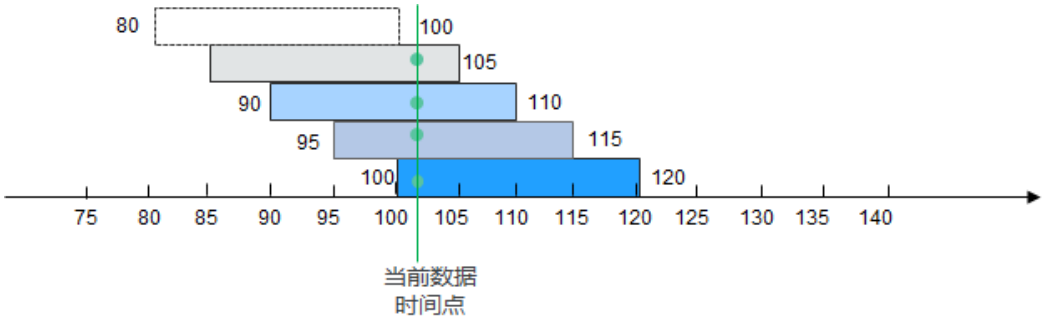
窗口中数据的保存形式主要有中间结果和原始数据两种，对窗口中的数据使用公共算子，如sum等操作时（window(SlidingEventTimeWindows.of(Time.seconds(20), Time.seconds(5))).sum）仅会保留中间结果；当用户使用自定义窗口时（window(SlidingEventTimeWindows.of(Time.seconds(20), Time.seconds(5))).apply(new UDF)）时保存所有的原始数据。

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

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

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

图 1-18 窗口原始结构示例



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

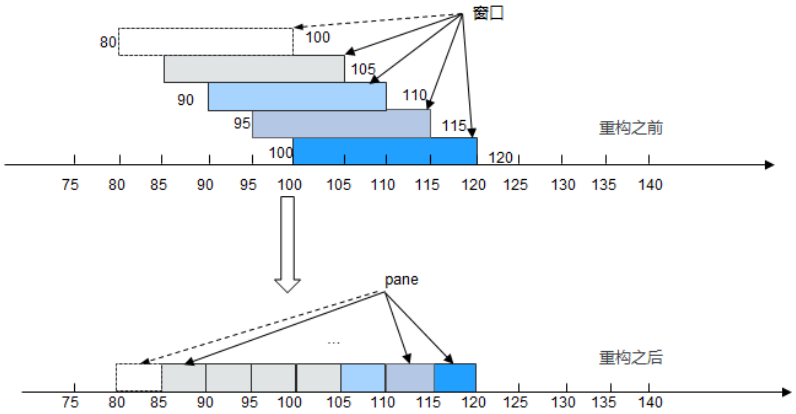
窗口优化

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

1. 以滑动周期为单位，将窗口划分为若干相互不重合的pane。

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

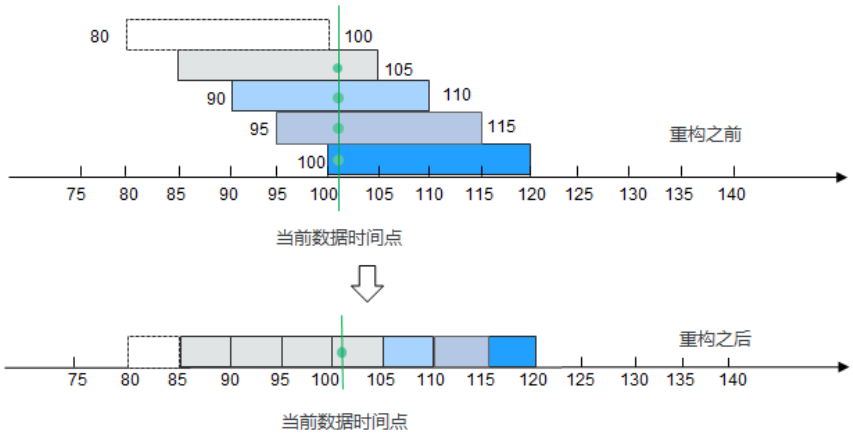
图 1-19 窗口重构示例



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

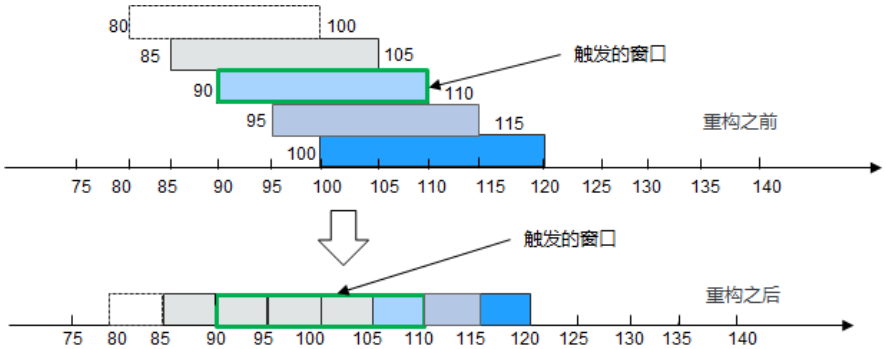
   一个数据仅保存在一个pane中，内存中只有一份。

图 1-20 窗口保存数据示例



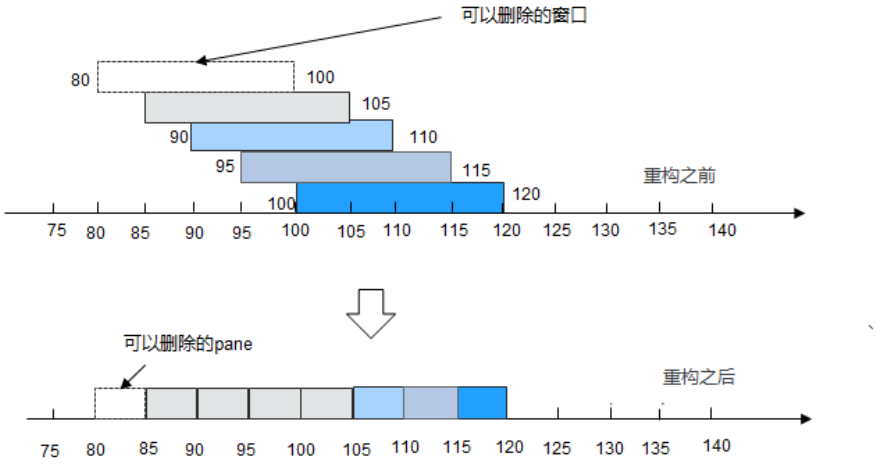
3. 当需要触发某个窗口时，计算该窗口包含的所有pane，并取出合并成一个完整的窗口计算。

图 1-21 窗口触发计算示例



4. 当某个pane不再需要时，将其从内存中删除。

图 1-22 窗口删除示例



通过优化，可以大幅度降低数据在内存以及快照中的数量。

1.4.5.4.2 Job Pipeline

Flink 开源增强特性：Job Pipeline

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

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

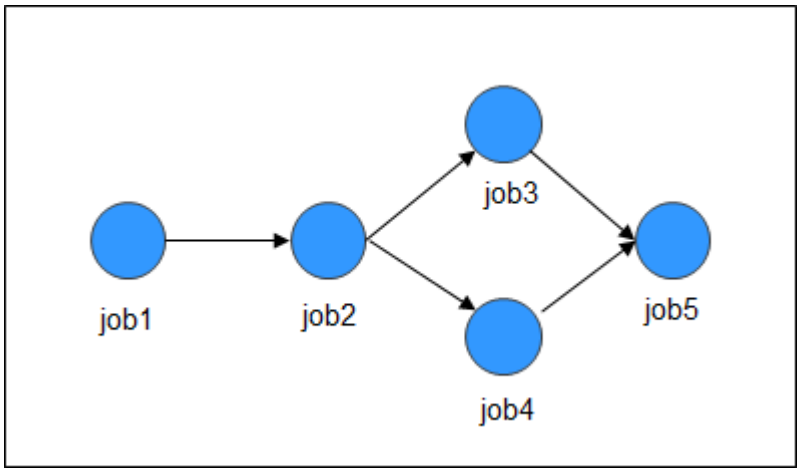
拆分成多个作业目前还存在问题。

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

Job Pipeline流图结构

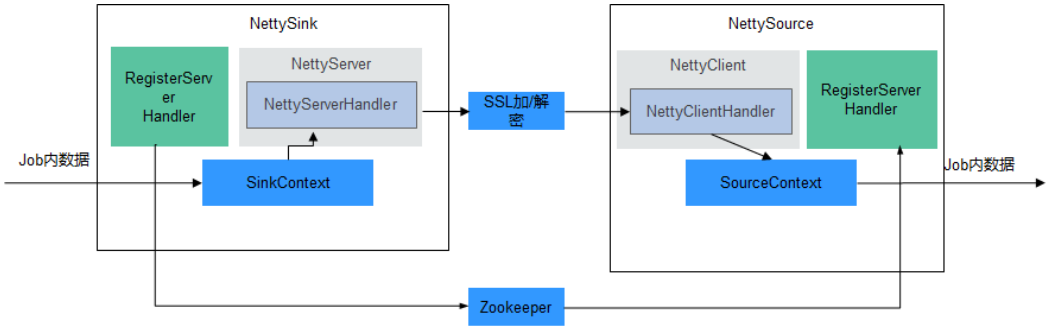
Pipeline是由Flink的多个Job通过TCP连接起来，上游Job可以直接向下游Job发送数据。这种发送数据的流图称为Job Pipeline，如图1-23所示。

图 1-23 Job Pipeline 流图



Job Pipeline原理介绍

图 1-24 Job Pipeline 原理图



- NettySink和NettySource

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

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

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

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

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

Job Pipeline功能介绍

- **NettySink**

  NettySink由以下几个重要模块组成：

  - RichParallelSinkFunction

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

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

    也可以通过其属性获取以下信息：

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

  - RegisterServerHandler

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

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

说明

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

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

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

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

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

– NettyServer

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

- 端口范围

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

- 子网

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

说明

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

– NettyServerHandler

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

- **NettySource**

NettySource由以下几个重要模块组成：

– RichParallelSourceFunction

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

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

也可以通过其属性获取以下信息：

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

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

- RegisterServerHandler

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

- NettyClient

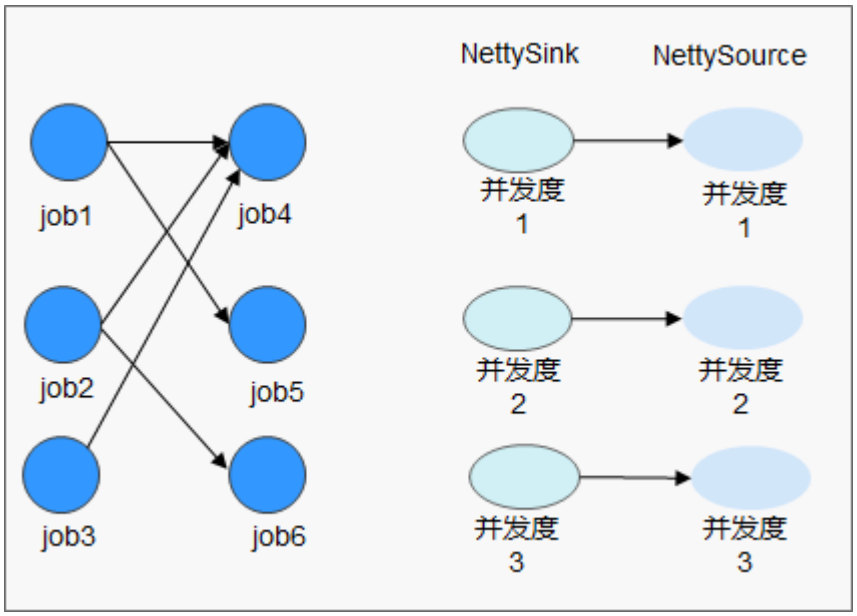
  NettyClient与NettyServer建立连接，并通过NettyClientHandler接收数据。每个NettySource算子必须具有唯一的name（由用户来保障）。NettyServer通过唯一的name确定每个Client来自不同的NettySource。当NettyClient与NettyServer建立连接时，首先向NettyServer注册NettyClient，将NettyClient的NettySource name传递给NettyServer。

- NettyClientHandler

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

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

图 1-25 关系图



1.4.5.4.3 Stream SQL Join

Flink 开源增强特性：Stream SQL Join

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

Stream SQL Join介绍

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

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

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

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

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

说明

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

1.4.5.4.4 Flink CEP in SQL

SQL 中的 Flink CEP

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

SQL 查询语法

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

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

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

MATCH_RECOGNIZE子句的语法元素定义如下：

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

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

-MEASURES [可选]：指定匹配成功的事件的属性值。

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

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

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

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

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

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

此外，还支持以下函数：

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

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

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

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

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

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

查询示例

以下查询发现股票价格数据流中的V型模式。

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

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

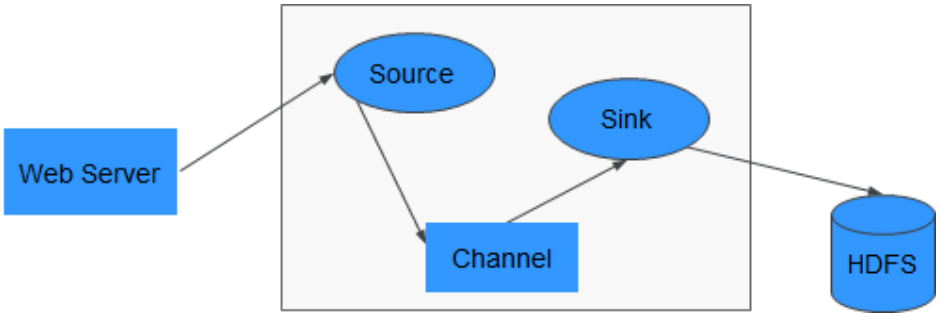
```
SELECT *
   FROM Ticker
   MATCH_RECOGNIZE (
     MEASURES
       AVG(E.price) AS avgPrice
     ONE ROW PER MATCH
     AFTER MATCH SKIP PAST LAST ROW
     PATTERN (A B+ C)
     SUBSET E = (A,C)
     DEFINE
       A AS A.price < 30,
       B AS B.price < 20,
       C AS C.price < 30
   )
```

1.4.6 Flume

1.4.6.1 Flume 基本原理

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

图 1-26 Flume-NG 架构



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

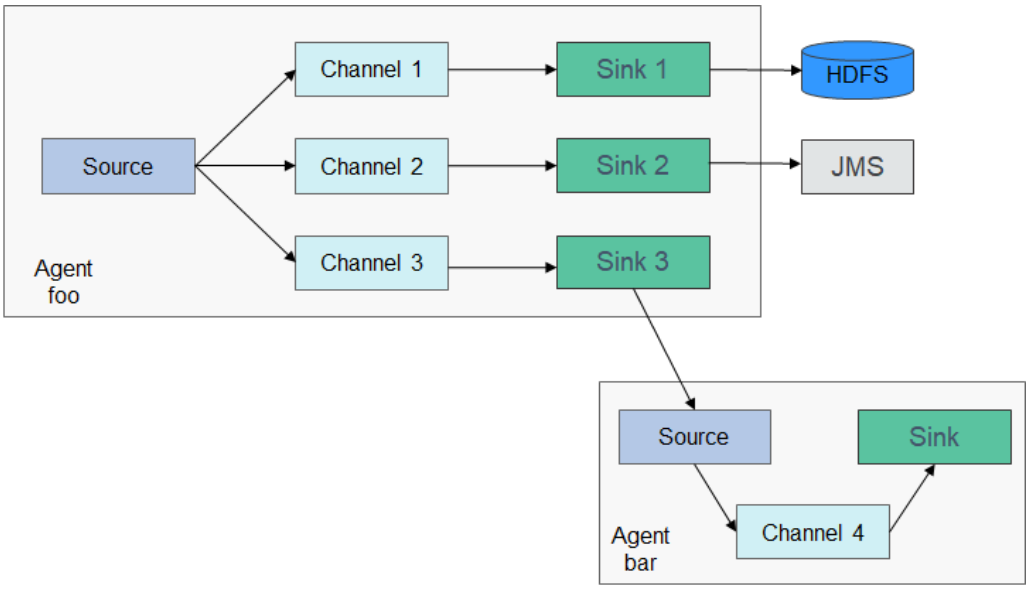
表 1-4 模块说明

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

名称	说明
Sink	Sink负责将数据传输到下一跳或最终目的，成功完成后将数据从Channel移除。 典型的Sink类型如下： ● 存储数据到最终目的终端Sink，比如：HDFS、HBase。 ● 自动消耗的Sink，比如：Null Sink。 ● 用于Agent间通信的IPC sink：Avro。 Sink必须作用于一个确切的Channel。

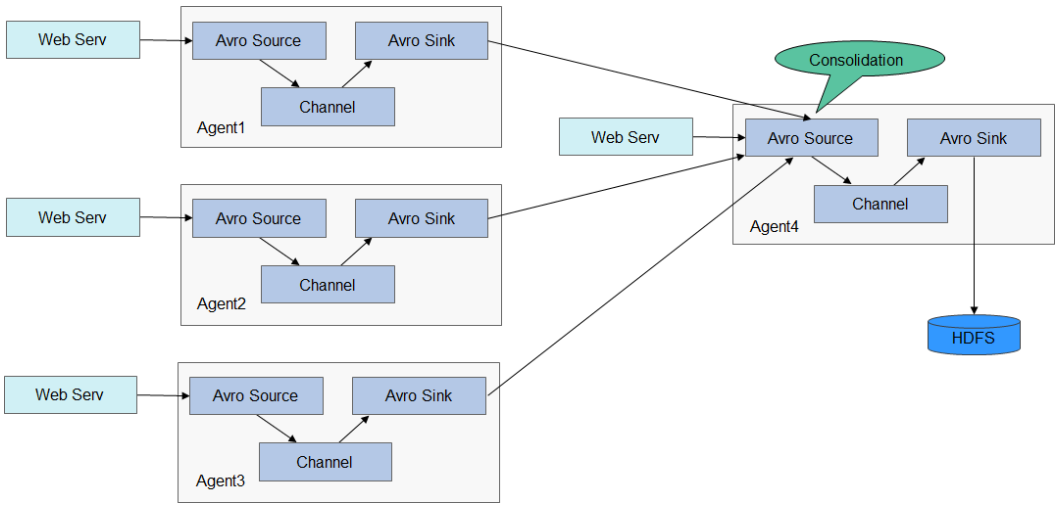
Flume也可以配置成多个Source、Channel、Sink，如图1-27所示：

图 1-27 Flume 结构图



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

图 1-28 Flume 级联结构图



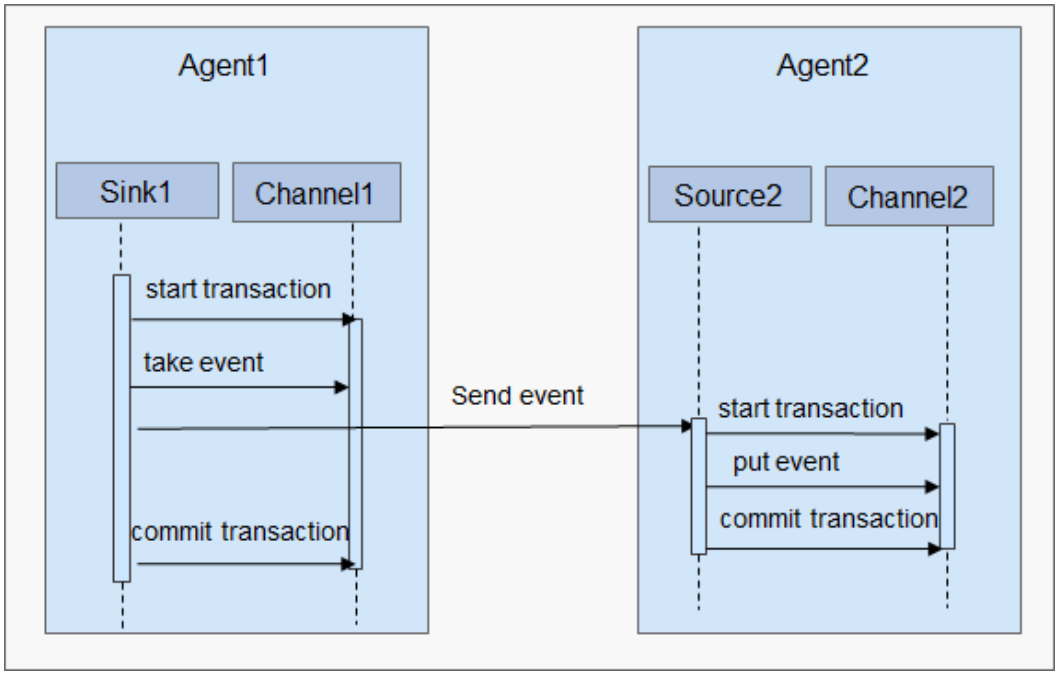
Flume的架构和详细原理介绍，请参见：**https://flume.apache.org/releases/1.9.0.html**。

Flume 原理

Agent之间的可靠性

Agent之间数据交换流程如**图1-29**所示。

图 1-29 Agent 数据传输流程



1. Flume采用基于Transactions的方式保证数据传输的可靠性，当数据从一个Agent流向另外一个Agent时，两个Transactions已经开始生效。发送Agent的Sink首先从Channel取出一条消息，并且将该消息发送给另外一个Agent。如果接受消息的

Agent成功地接受并处理消息，那么发送Agent将会提交Transactions，标识一次数据传输成功可靠地完成。

2. 当接收Agent接受到发送Agent发送的消息时，开始一个新的Transactions，当该数据被成功处理（写入Channel中），那么接收Agent提交该Transactions，并向发送Agent发送成功响应。
3. 如果在某次提交（commit）之前，数据传输出现了失败，将会再次开始上一次Transactions，并将上次发送失败的数据重新传输。因为commit操作已经将Transactions写入了磁盘，那么在进程故障退出并恢复业务之后，仍然可以继续上次的Transactions。

1.4.6.2 Flume 与其他组件的关系

Flume 与 HDFS 的关系

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

Flume 与 HBase 的关系

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

1.4.6.3 Flume 开源增强特性

Flume 开源增强特性

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

1.4.7 HBase

1.4.7.1 HBase 基本原理

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

存储在HBase中的表的典型特征：

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

MRS服务支持HBase组件的二级索引，支持为列值添加索引，提供使用原生的HBase接口的高性能基于列过滤查询的能力。

HBase 结构

HBase集群由主备Master进程和多个RegionServer进程组成。如图 HBase结构所示。

图 1-30 HBase 结构

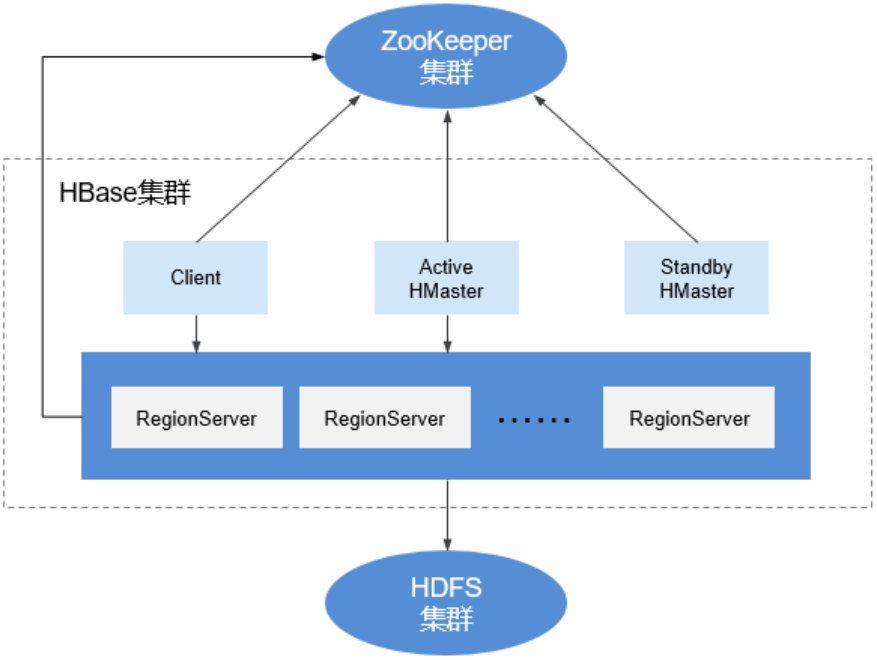


表 1-5 模块说明

名称	描述
Master	又叫HMaster，在HA模式下，包含主用Master和备用Master。 • 主用Master：负责HBase中RegionServer的管理，包括表的增删改查；RegionServer的负载均衡，Region分布调整；Region分裂以及分裂后的Region分配；RegionServer失效后的Region迁移等。 • 备用Master：当主用Master故障时，备用Master将取代主用Master对外提供服务。故障恢复后，原主用Master降为备用。
Client	Client使用HBase的RPC机制与Master、RegionServer进行通信。Client与Master进行管理类通信，与RegionServer进行数据操作类通信。
RegionServer	RegionServer负责提供表数据读写等服务，是HBase的数据处理和计算单元。 RegionServer一般与HDFS集群的DataNode部署在一起，实现数据的存储功能。

名称	描述
ZooKeeper集群	ZooKeeper为HBase集群中各进程提供分布式协作服务。各RegionServer将自己的信息注册到ZooKeeper中，主用Master据此感知各个RegionServer的健康状态。
HDFS集群	HDFS为HBase提供高可靠的文件存储服务，HBase的数据全部存储在HDFS中。

HBase 原理

- **HBase数据模型**

  HBase以表的形式存储数据，数据模型如图1-31所示。表中的数据划分为多个Region，并由Master分配给对应的RegionServer进行管理。

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

  图 1-31 HBase 数据模型

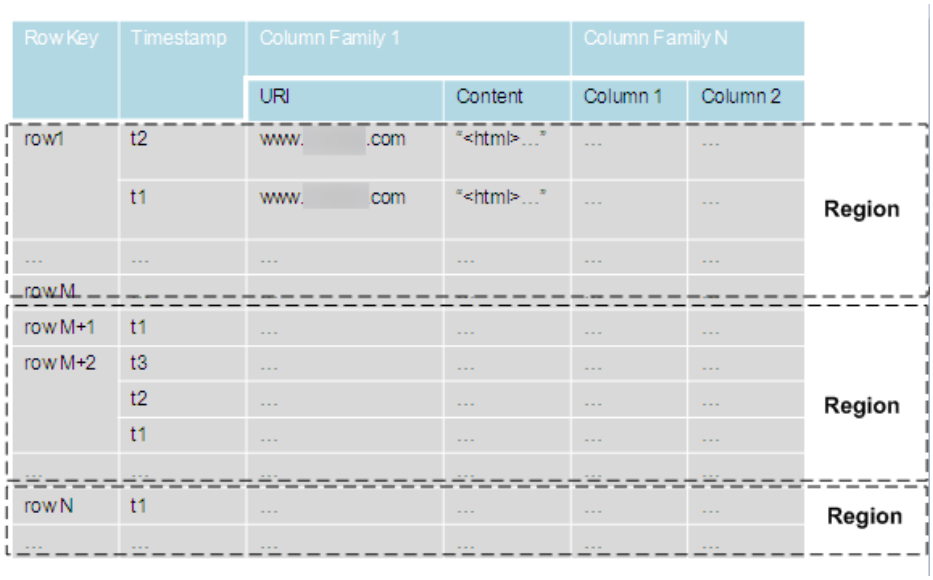


  表 1-6 概念介绍

名称	描述
RowKey	行键，相当于关系表的主键，每一行数据的唯一标识。字符串、整数、二进制串都可以作为RowKey。所有记录按照RowKey排序后存储。
Timestamp	每次数据操作对应的时间戳，数据按时间戳区分版本，每个Cell的多个版本的数据按时间倒序存储。

名称	描述
Cell	HBase最小的存储单元，由Key和Value组成。Key由row、column family、column qualifier、timestamp、type、MVCC version这6个字段组成。Value就是对应存储的二进制数据对象。
Column Family	列族，一个表在水平方向上由一个或多个Column Family组成。一个CF（Column Family）可以由任意多个Column组成。Column是CF下的一个标签，可以在写入数据时任意添加，因此CF支持动态扩展，无需预先定义Column的数量和类型。HBase中表的列非常稀疏，不同行的列的个数和类型都可以不同。此外，每个CF都有独立的生存周期（TTL）。可以只对行上锁，对行的操作始终是原始的。
Column	列，与传统的数据库类似，HBase的表中也有列的概念，列用于表示相同类型的数据。

- **RegionServer数据存储**

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

  图 1-32 RegionServer 的数据存储结构

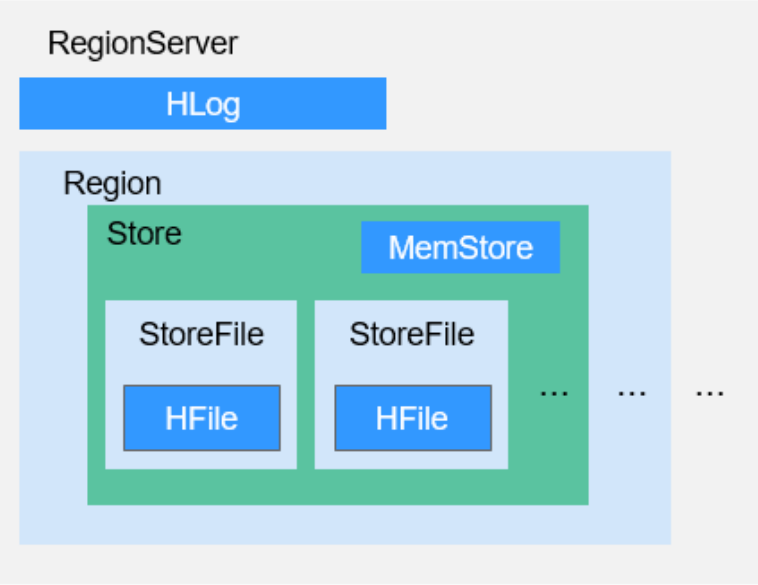
  

  图 RegionServer的数据存储结构中Region的各部分的说明如表 Region结构说明所示。

  表 1-7 Region 结构说明

名称	描述
Store	一个Region由一个或多个Store组成，每个Store对应图 HBase数据模型中的一个Column Family。

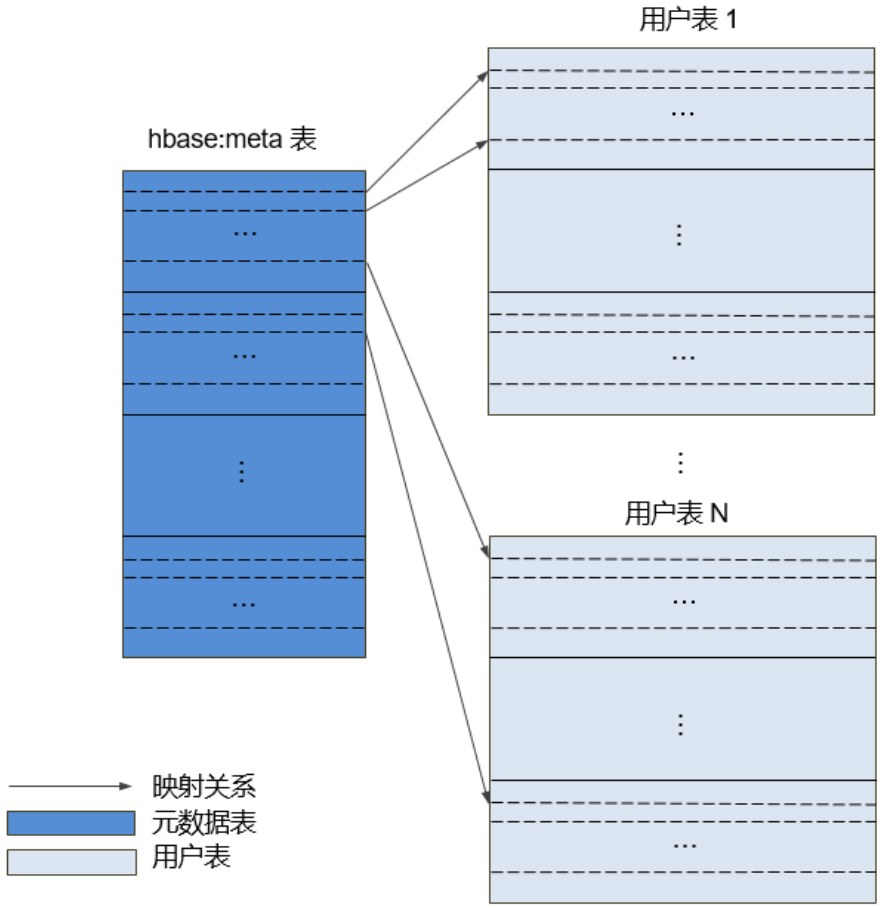
名称	描述
MemStore	一个Store包含一个MemStore，MemStore缓存客户端向Region插入的数据，当RegionServer中的MemStore大小达到配置的容量上限时，RegionServer会将MemStore中的数据“flush”到HDFS中。
StoreFile	MemStore的数据flush到HDFS后成为StoreFile，随着数据的插入，一个Store会产生多个StoreFile，当StoreFile的个数达到配置的最大值时，RegionServer会将多个StoreFile合并为一个大的StoreFile。
HFile	HFile定义了StoreFile在文件系统中的存储格式，它是当前HBase系统中StoreFile的具体实现。
HLog	HLog日志保证了当RegionServer故障的情况下用户写入的数据不丢失，RegionServer的多个Region共享一个相同的HLog。

- **元数据表**

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

  元数据表和用户表的映射关系如图 元数据表和用户表的映射关系所示。

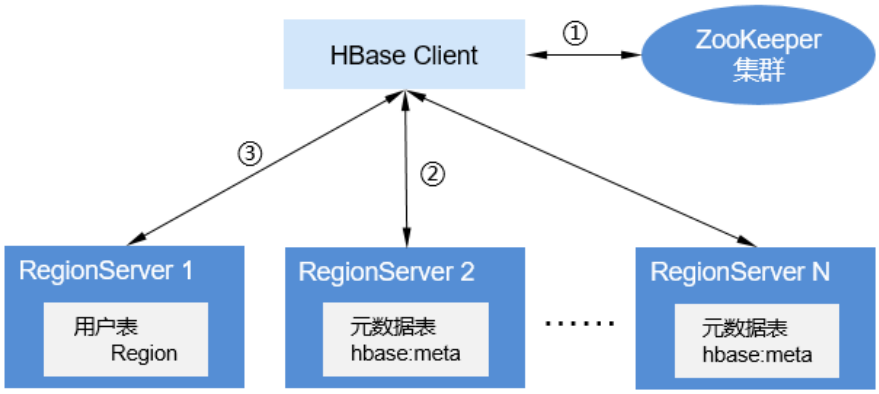
  图 1-33 元数据表和用户表的映射关系



- **数据操作流程**

  HBase数据操作流程如**图 数据操作流程**所示。

  图 1-34 数据操作流程



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

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

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

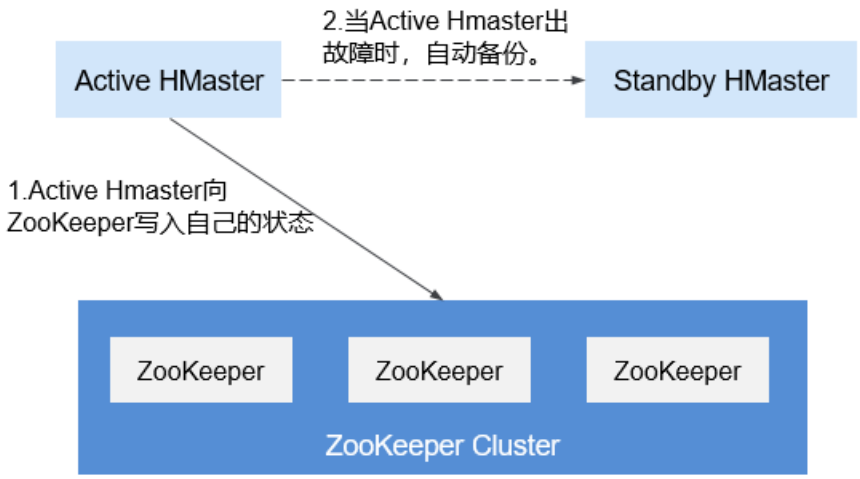
  为了提升数据操作的效率，HBase Client会在内存中缓存“hbase:meta”和用户表Region的信息，当应用程序发起下一次数据操作时，HBase Client会首先从内存中获取这些信息；当未在内存缓存中找到对应数据信息时，HBase Client会重复上述操作。

1.4.7.2 HBase HA 方案介绍

HBase HA 原理与实现方案

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

图 1-35 HMaster 高可用性实现架构



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

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

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

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

1.4.7.3 HBase 与其他组件的关系

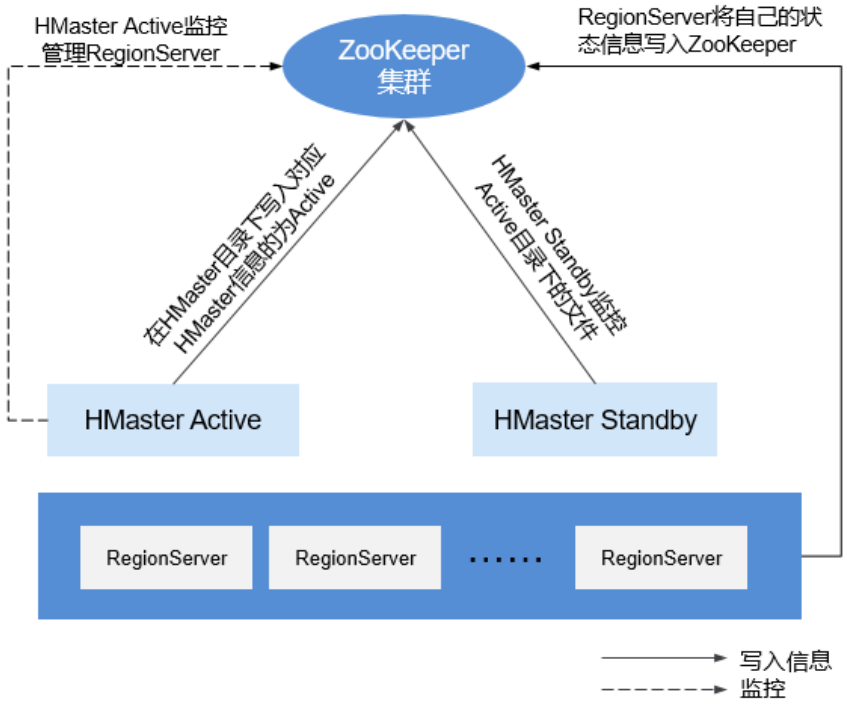
HBase 和 HDFS 的关系

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

HBase 和 ZooKeeper 的关系

HBase和ZooKeeper的关系如图 ZooKeeper和HBase的关系所示。

图 1-36 HBase 和 ZooKeeper 的关系



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

1.4.7.4 HBase 开源增强特性

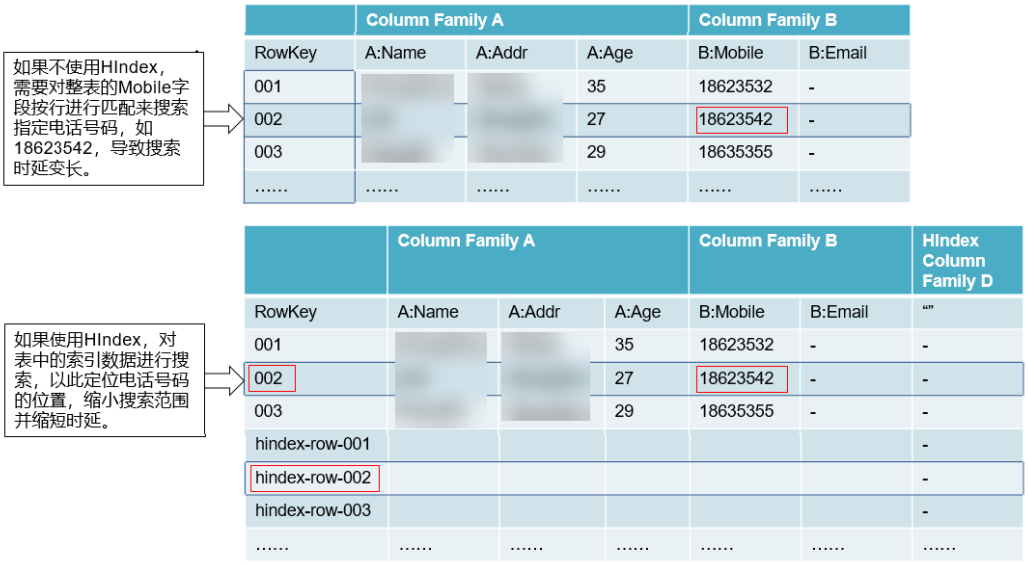
HBase 开源增强特性：HIndex

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

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

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

图 1-37 HIndex



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

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

    正确的写操作：

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

    错误的写操作：

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

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

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

    正确的查询操作：

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

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

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

错误的查询操作：

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

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

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

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

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

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

  索引2：IDX2=>cf2:[q2->datatype]

  正确的更新操作：

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

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

  错误的更新操作：

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

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

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

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

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

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

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

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

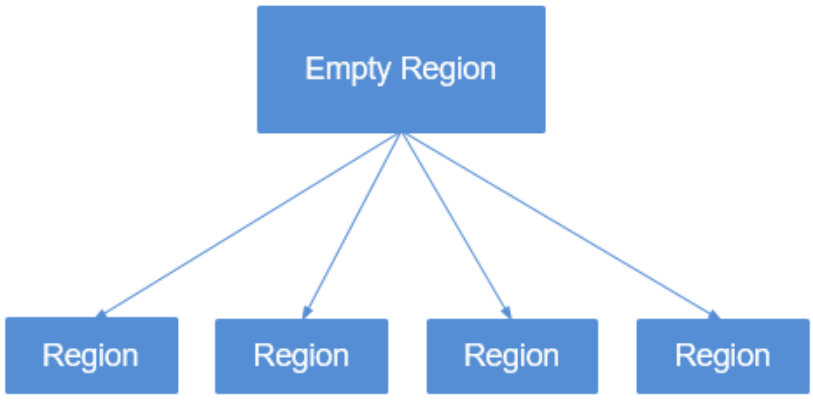
HBase 开源增强特性：支持多点分割

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

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

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

图 1-38 多点分割



HBase 开源增强特性：连接数限制

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

HBase 开源增强特性：容灾增强

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

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

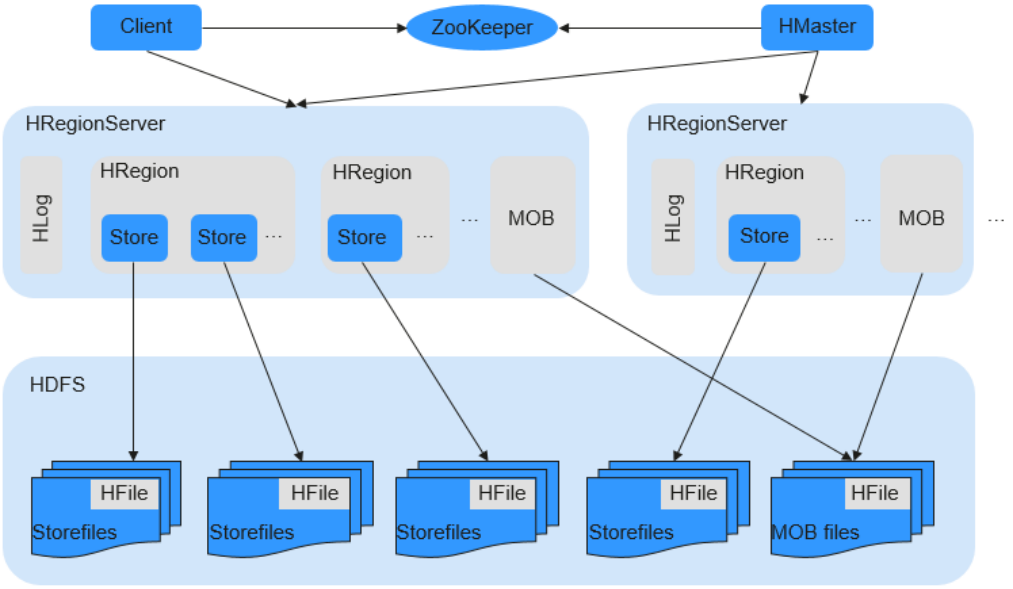
HBase 开源增强特性：HBase MOB

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

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

如**图1-39**所示，图中MOB模块表示存储在HRegion上的mobstore，mobstore存储的是key-value，key即为HBase中对应的key，value对应的就是存储在文件系统上的引用地址以及数据偏移量。读取数据时，mobstore会用自己的scanner，先读取mobstore中的key-value数据对象，然后通过value中的地址及数据大小信息，从文件系统中读取真正的数据。

图 1-39 MOB 数据存储原理



HBase 开源增强特性：HFS

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

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

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

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

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

HBase 开源增强特性：多 RegionServer 共机部署

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

单RegionServer资源利用率低：

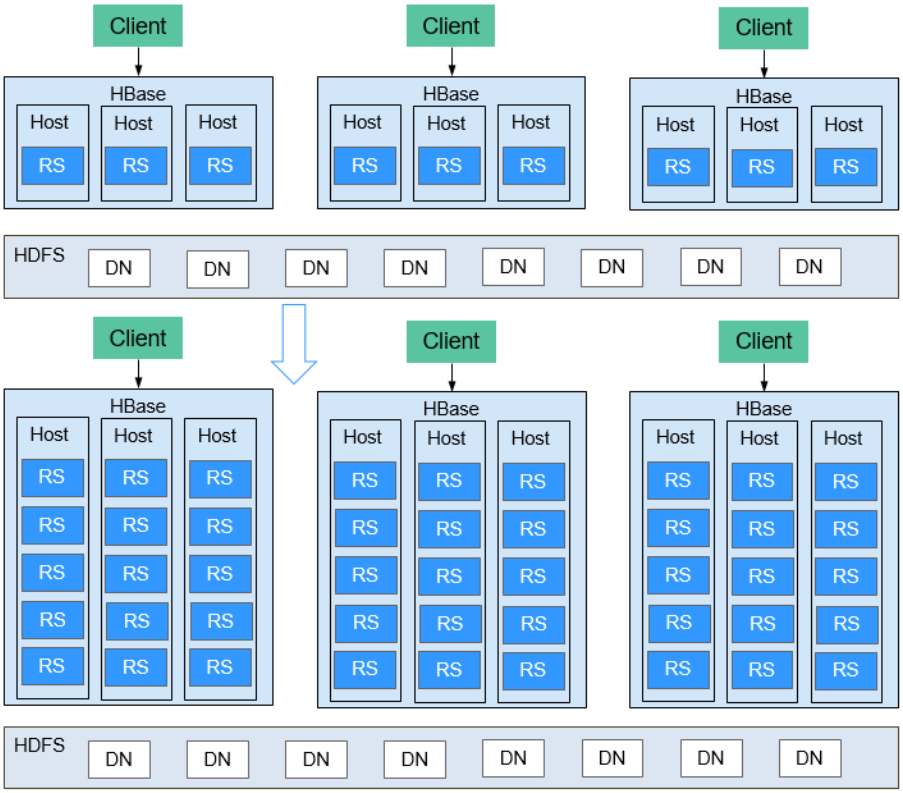
1. 单个RegionServer支持的Region数量有限，无法充分利用内存、CPU资源。
2. 单个RegionServer数据量为20T，两副本为40T，三副本60T，无法用完96T的磁盘。
3. 写入性能差：一台物理机一个RegionServer，只有一个HLog，只能同时写三块盘。

多RegionServer共机部署，提升HBase资源利用率：

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

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

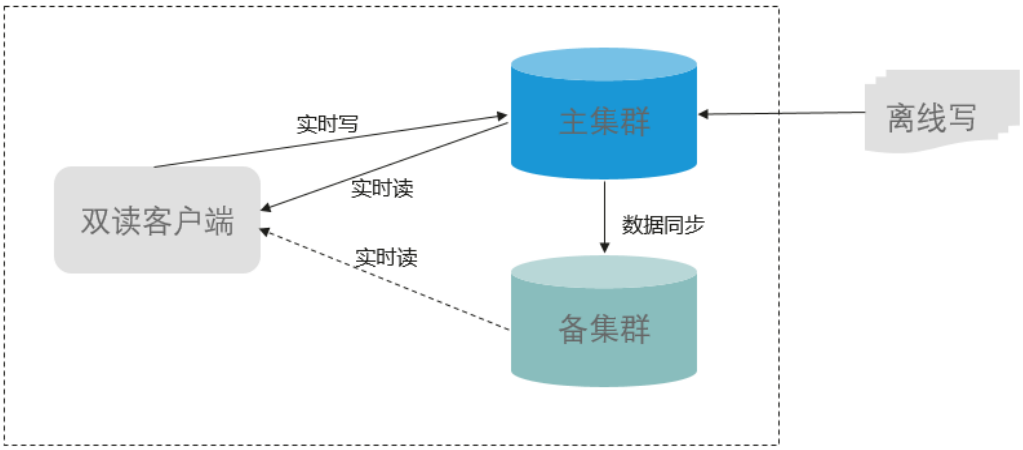
图 1-40 HBase 资源利用率提升



HBase 开源增强特性：HBase 双读

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

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



1.4.8 HDFS

1.4.8.1 HDFS 基本原理

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

HDFS 结构

HDFS包含主、备NameNode和多个DataNode，如图1-41所示。

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

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

图 1-41 HA HDFS 结构

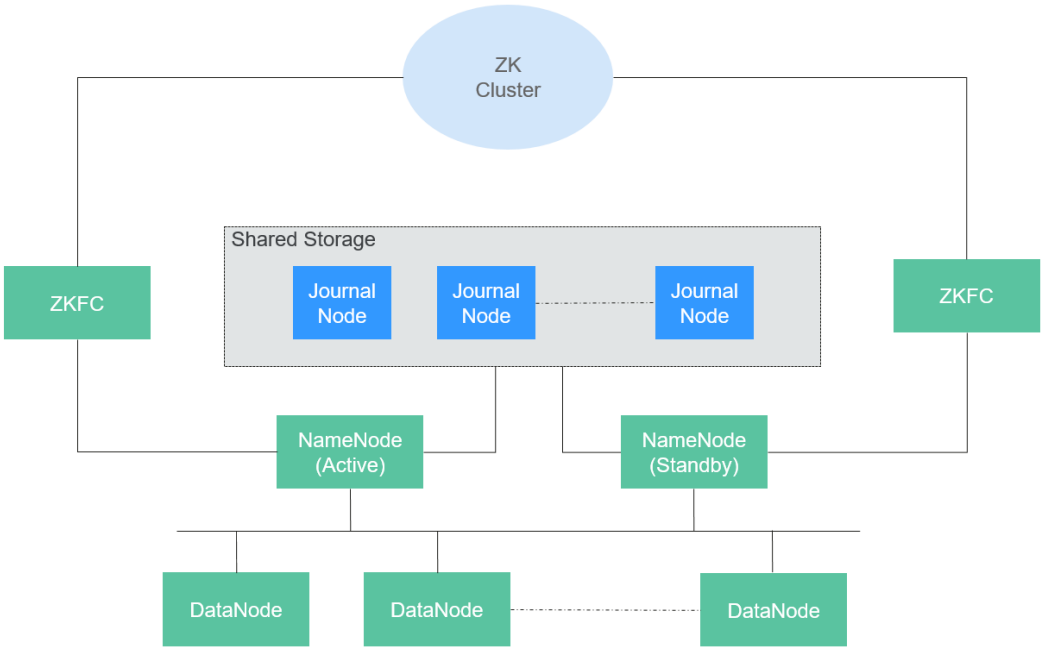


图1-41中各模块的功能说明如表1-8所示。

表 1-8 模块说明

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

- **HDFS HA架构**

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

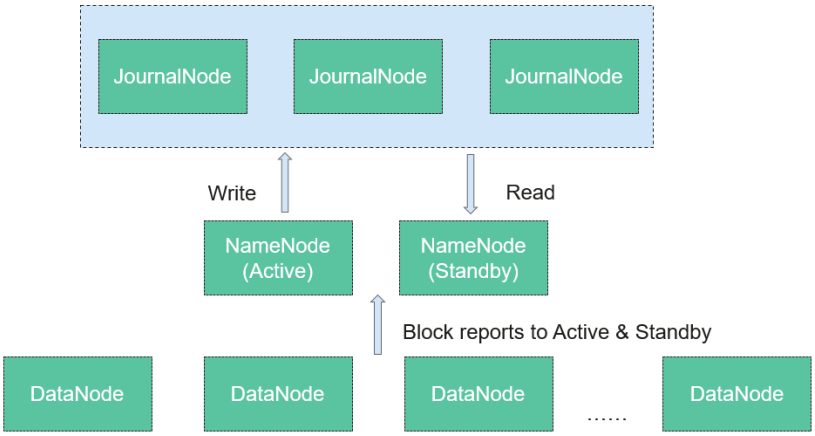
  在一个典型HDFS HA场景中，通常由两个NameNode组成，一个处于Active状态，另一个处于Standby状态。

  为了能实现Active和Standby两个NameNode的元数据信息同步，需提供一个共享存储系统。本版本提供基于QJM（Quorum Journal Manager）的HA解决方案，如图1-42所示。主备NameNode之间通过一组JournalNode同步元数据信息。

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

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

图 1-42 基于 QJM 的 HDFS 架构

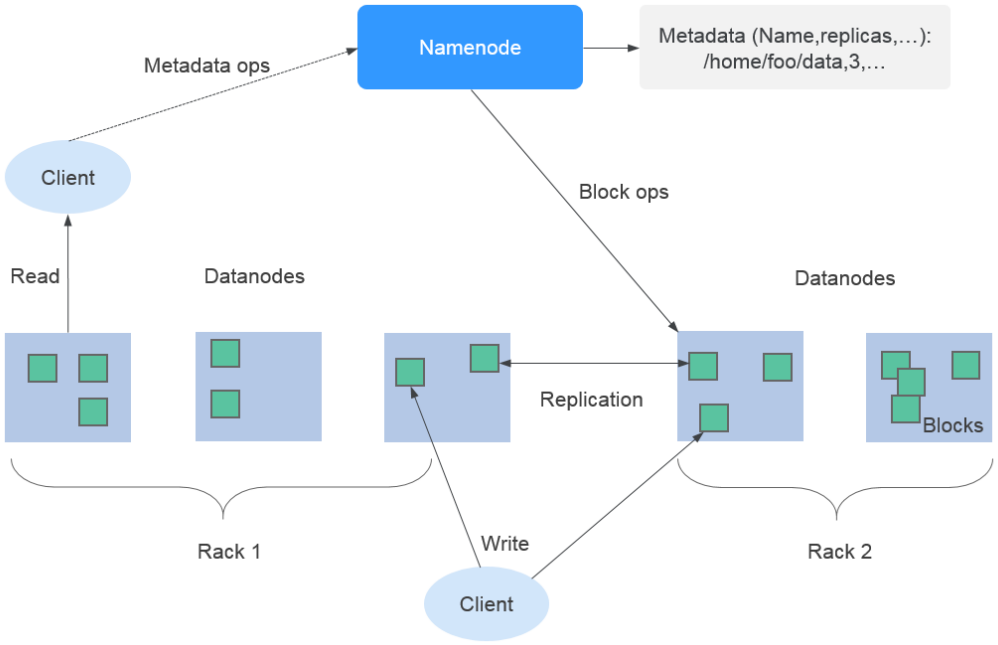


HDFS 原理

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

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

图 1-43 HDFS 架构



MRS服务的HDFS组件支持以下部分特性：

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

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

1.4.8.2 HDFS HA 方案介绍

HDFS HA 方案背景

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

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

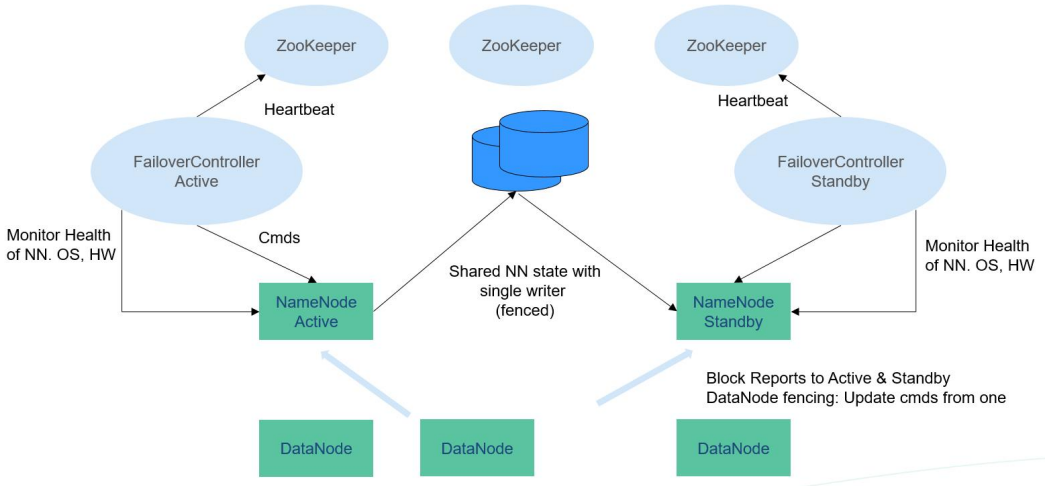
针对以上问题，HDFS高可用性方案通过自动或手动（可配置）的方式，在一个集群中为NameNode启动一个热替换的NameNode备份。当一台机器故障时，可以迅速地自动进行NameNode主备切换。或者当主NameNode节点需要进行维护时，通过MRS集群管理员控制，可以手动进行NameNode主备切换，从而保证集群在维护期间的可用性。

有关HDFS自动故障转移功能，请参阅：

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

HDFS HA 实现方案

图 1-44 典型的 HA 部署方式



在一个典型的HA集群中（如**图1-44**），需要把两个NameNodes配置在两台独立的机器上。在任何一个时间点，只有一个NameNode处于Active状态，另一个处于Standby状态。Active节点负责处理所有客户端操作，Standby节点时刻保持与Active节点同步的状态以便在必要时进行快速主备切换。

为保持Active和Standby节点的数据一致性，两个节点都要与一组称为JournalNode的节点通信。当Active对文件系统元数据进行修改时，会将其修改日志保存到大多数的JournalNode节点中，例如有3个JournalNode，则日志会保存在至少2个节点中。Standby节点监控JournalNodes的变化，并同步来自Active节点的修改。根据修改日志，Standby节点将变动应用到本地文件系统元数据中。一旦发生故障转移，Standby节点能够确保与Active节点的状态是一致的。这保证了文件系统元数据在故障转移时在Active和Standby之间是完全同步的。

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

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

关于HDFS高可用性方案的更多信息，可参考如下链接：

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

1.4.8.3 HDFS 与其他组件的关系

HDFS 和 HBase 的关系

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

HDFS 和 MapReduce 的关系

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

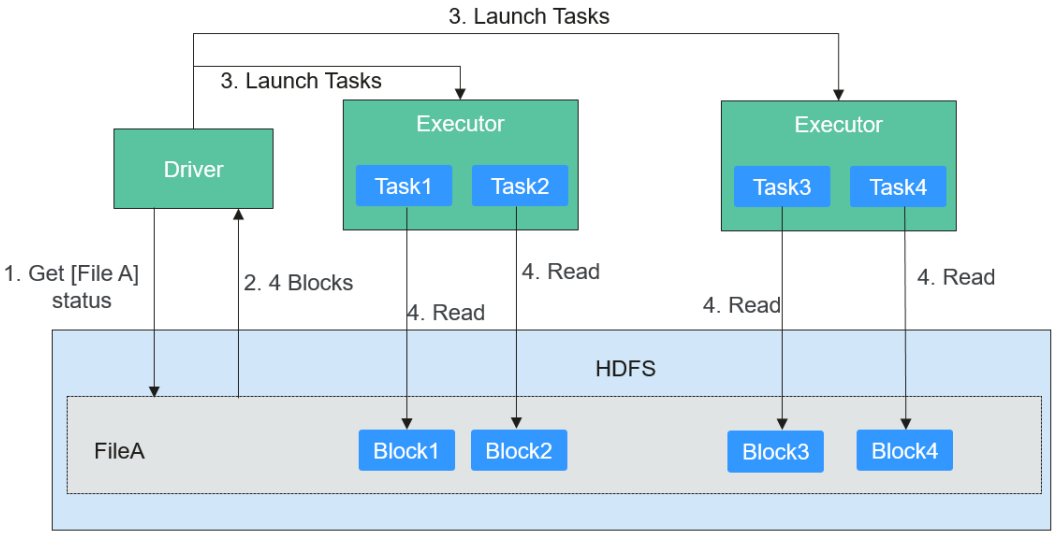
HDFS 和 Spark 的关系

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

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

读取文件的过程如**图1-45**所示。

图 1-45 读取文件过程

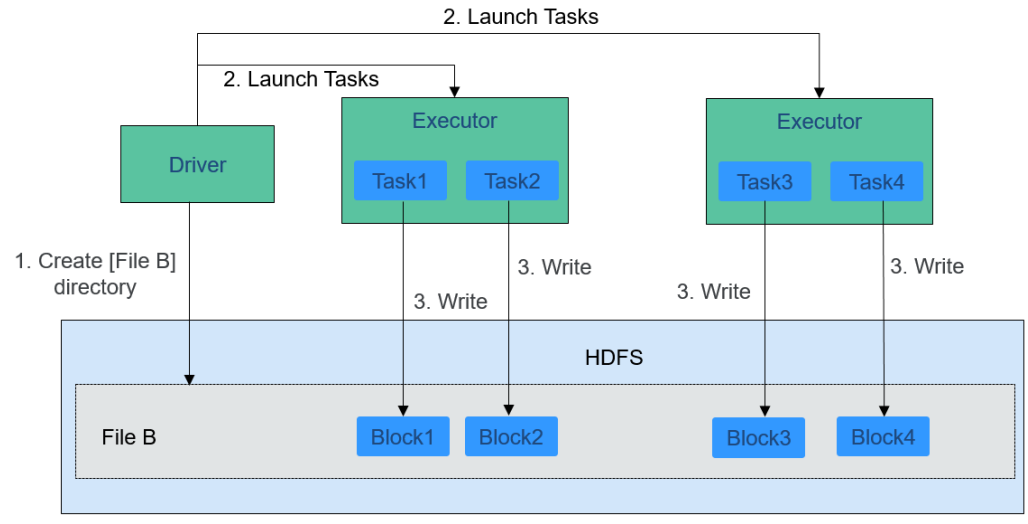


读取文件步骤的详细描述如下所示：

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

写入文件的过程如图1-46所示。

图 1-46 写入文件过程



HDFS文件写入的详细步骤如下所示：

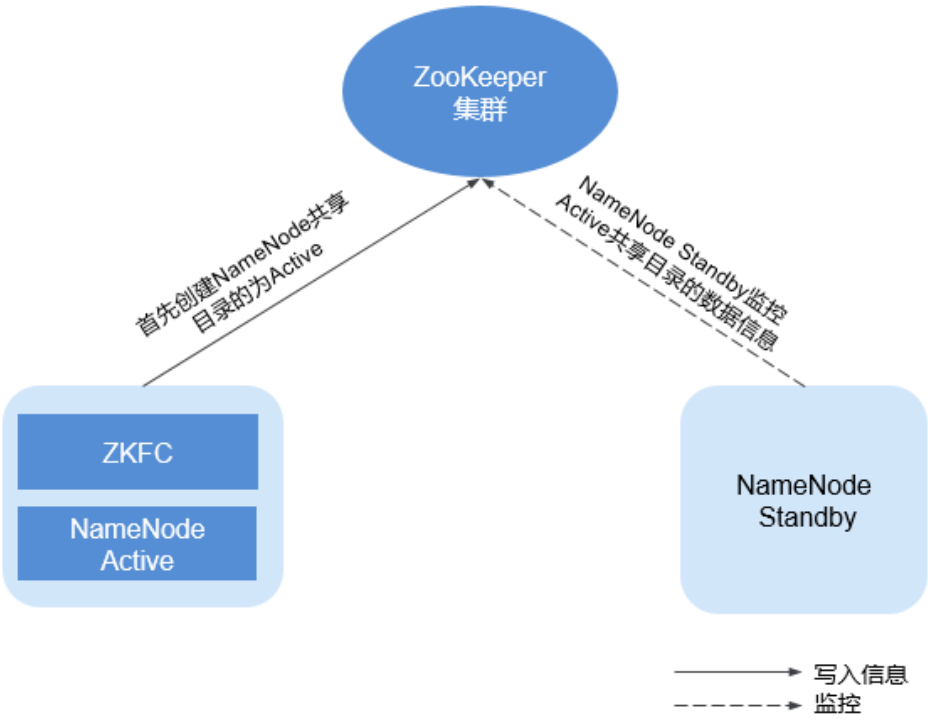
1. Driver创建要写入文件的目录。
2. 根据RDD分区分块情况，计算出写数据的Task数，并下发这些任务到Executor。

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

HDFS 和 ZooKeeper 的关系

ZooKeeper与HDFS的关系如图1-47所示。

图 1-47 ZooKeeper 和 HDFS 的关系



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

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

1.4.8.4 HDFS 开源增强特性

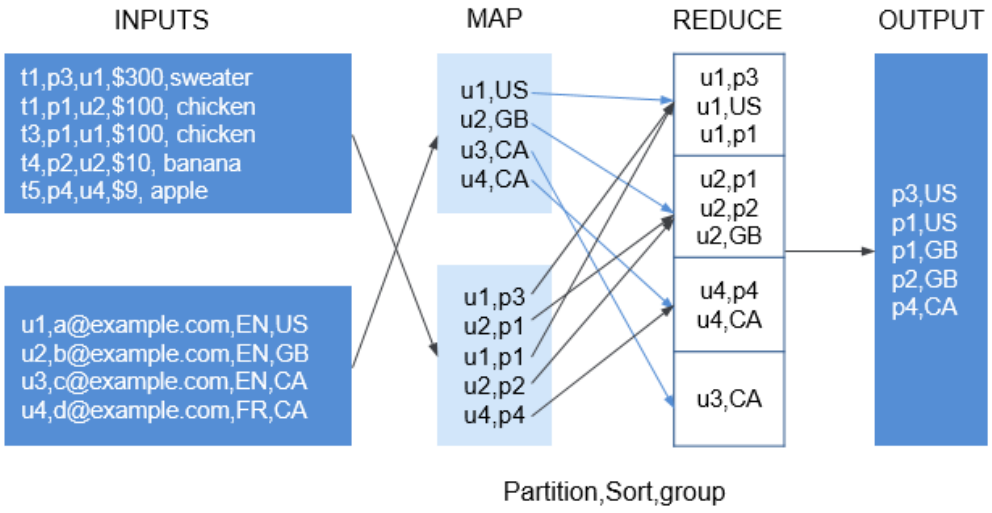
HDFS 开源增强特性：文件块同分布（Colocation）

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

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

以上方式的最大问题在于，由于数据分散在各节点上，所以在Map到Reduce过程中，需要大量的网络数据传输，使得Join计算的性能大大降低，该过程如图1-48所示：

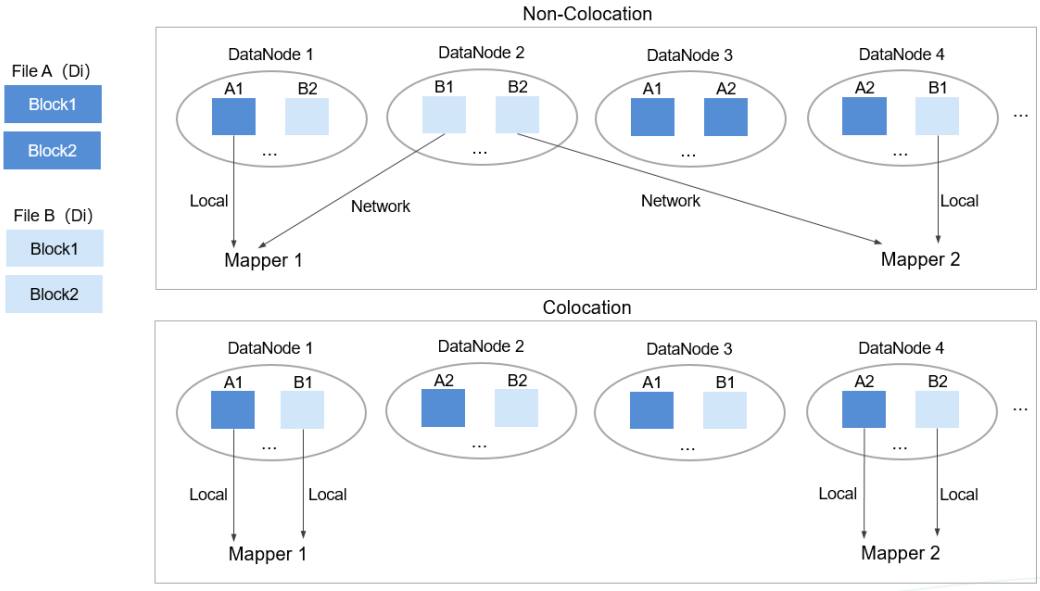
图 1-48 无同分布数据传输流程



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

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

图 1-49 无同分布与同分布数据块分布对比

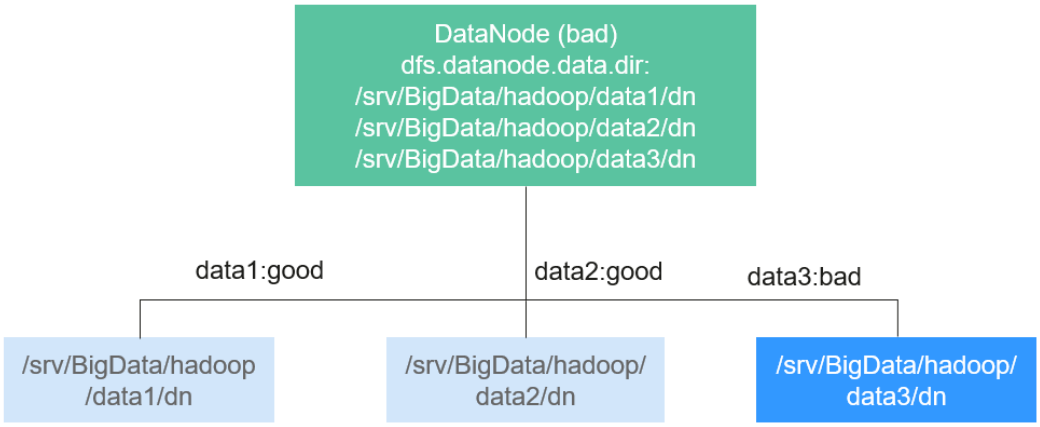


HDFS 开源增强特性：硬盘坏卷设置

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

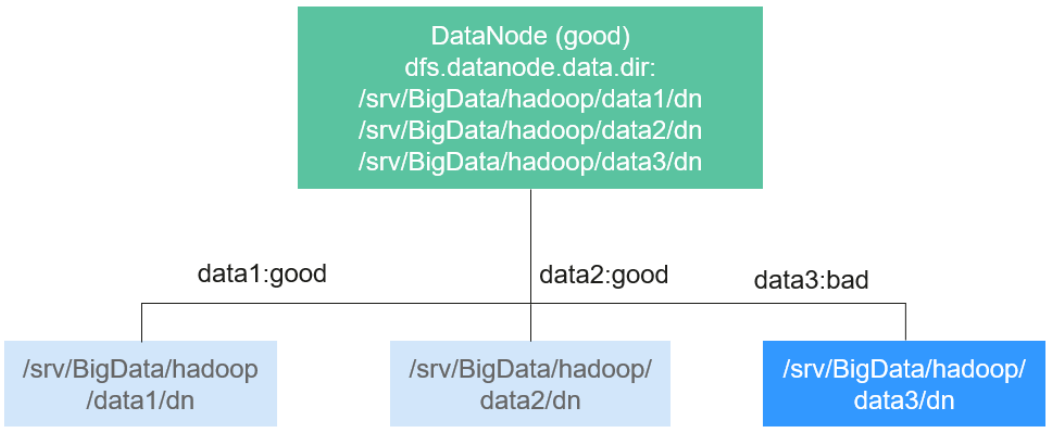
“dfs.datanode.failed.volumes.tolerated”取值范围为-1~DataNode上配置的磁盘卷数，默认值为-1，效果如图1-50所示。

图 1-50 选项设置为 0



例如：某个DataNode中挂载了3个数据存放卷，
“dfs.datanode.failed.volumes.tolerated”配置为1，则当该DataNode中的其中一个数据存放卷不能使用的时候，该DataNode会继续提供服务。如图1-51所示。

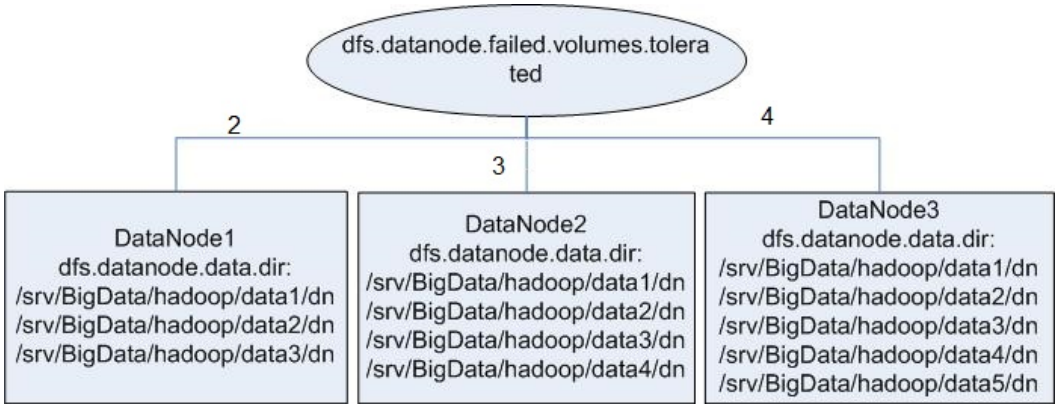
图 1-51 选项设置为 1



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

例如：集群中存在3个DataNode节点，第一个节点有3个数据目录，第二个节点有4个数据目录，第三个节点有5个数据目录，如果需要实现当节点有一个目录还可用的时候DataNode服务依然可用的效果，就需要如图1-52所示进行设置。

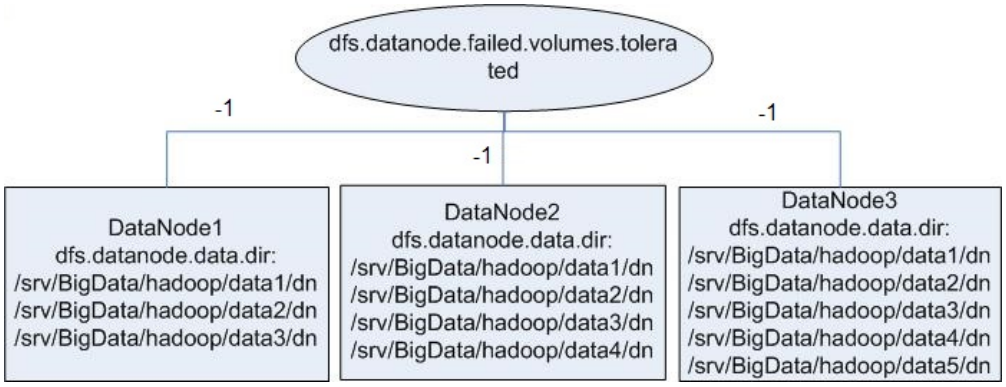
图 1-52 未增强前属性设置



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

所以对于上面提到的例子，该属性的配置将统一成-1，如图1-53所示。

图 1-53 增强后属性配置



HDFS 开源增强特性：HDFS 启动加速

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

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

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

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

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

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

如图1-54所示。

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

图 1-54 基于标签的数据块摆放策略样例

HDFS 开源增强特性：HDFS Load Balance

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

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

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

HDFS 开源增强特性：HDFS 冷热数据迁移

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

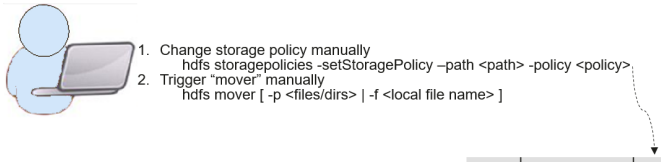
然而，随着技术的发展，Hadoop逐渐被用于以随机I/O访问模式的操作为主的上层应用上，如Hive、HBase等，而这种时延要求较高的场景中，低时延的高速磁盘（如SSD磁盘）可以得到广泛的应用。为了支持这种特性，HDFS现在支持了异构存储类型，这样用户就可以根据自己不同的业务需求场景来选择不同的数据存储类型。

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

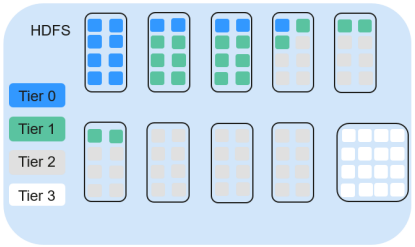
但是，这些高速低时延磁盘，例如SSD磁盘，通常比机械磁盘贵很多。大部分用户希望只有那些经常被访问的热数据才能一直被存储在昂贵的高速磁盘上，而随着数据的访问热度下降以及时间的老化，这些数据应该被迁移到价格低廉的存储介质上。

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

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



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



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

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

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

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

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

1.4.9 Hive

1.4.9.1 Hive 基本原理

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

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

Hive主要特点如下：

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

Hive体系结构：

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

Hive 结构

Hive为单实例的服务进程，提供服务的原理是将HQL编译解析成相应的MapReduce或者HDFS任务，图1-55为Hive的结构概图。

图 1-55 Hive 结构

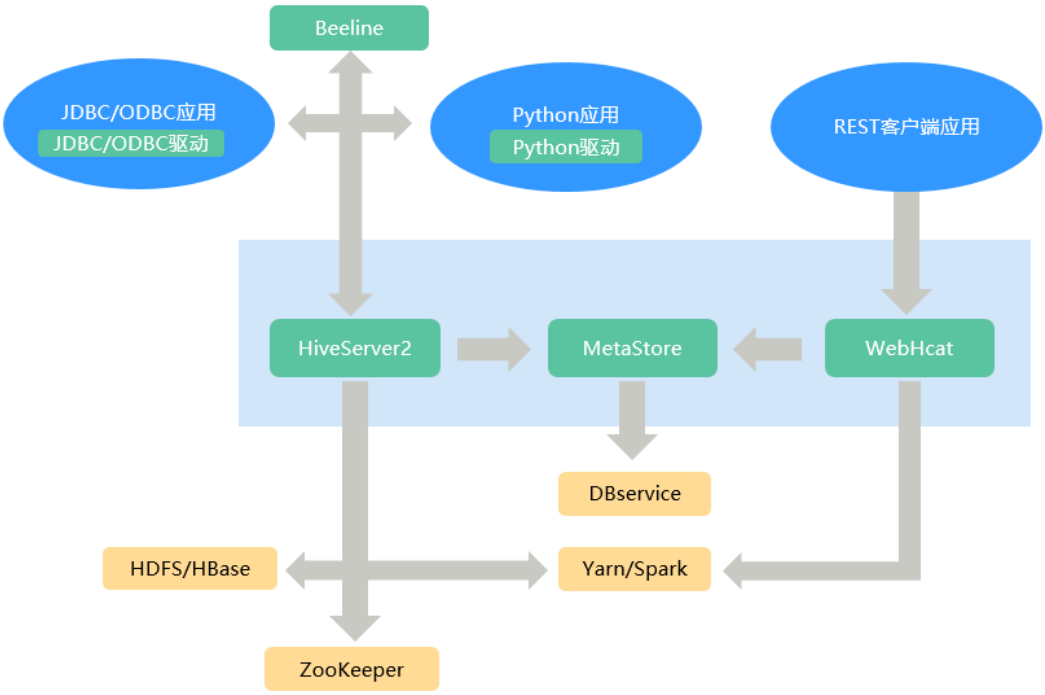
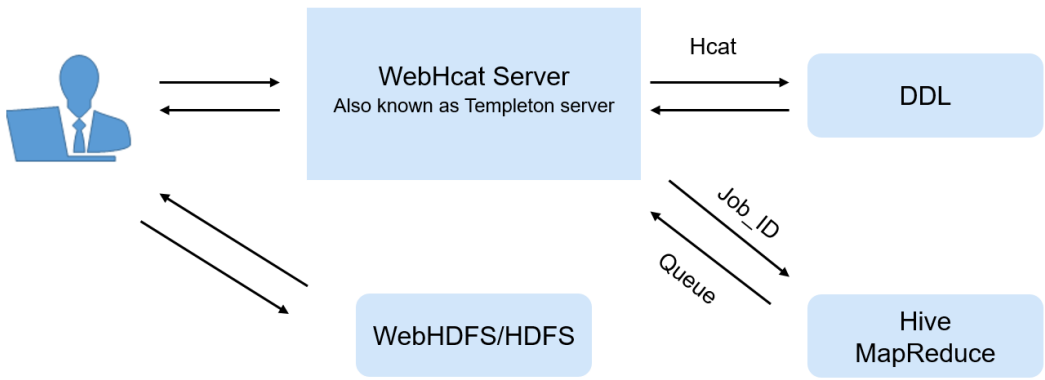


表 1-9 模块说明

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

HCatalog建立在Hive Metastore之上，具有Hive的DDL能力。从另外一种意义上说，HCatalog还是Hadoop的表和存储管理层，它使用户能够通过使用不同的数据处理工具（比如MapReduce），更轻松地在网格上读写HDFS上的数据，HCatalog还能为这些数据处理工具提供读写接口，并使用Hive的命令行接口发布数据定义和元数据探索命令。此外，经过封装这些命令，WebHcat Server还对外提供了RESTful接口，如图1-56所示。

图 1-56 WebHCat 的逻辑架构图



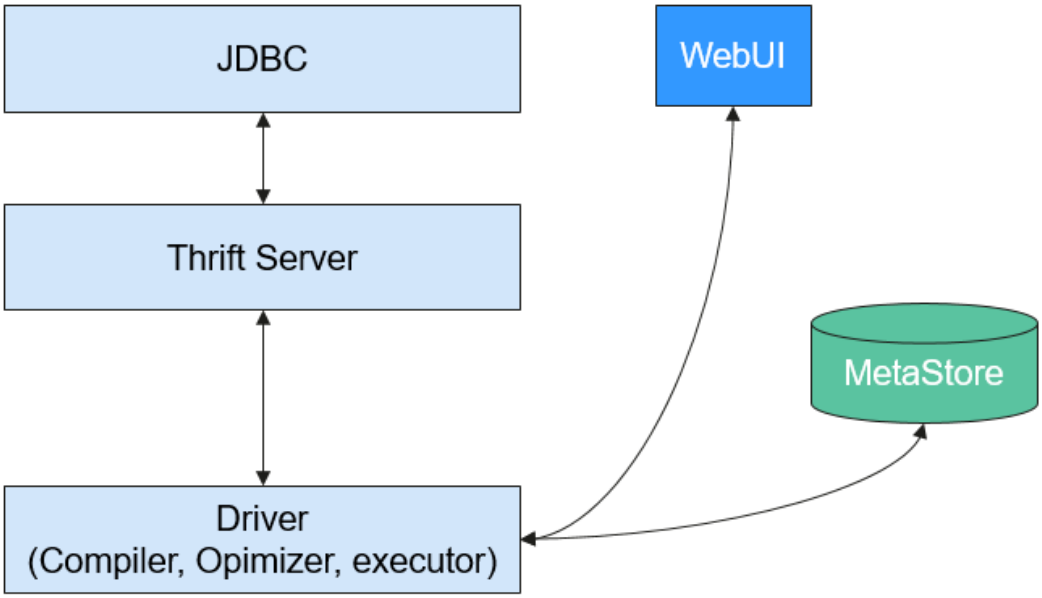
Hive 原理

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

图1-57为Hive的结构简图。

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

图 1-57 Hive 结构



1.4.9.2 Hive CBO 原理介绍

Hive CBO 原理介绍

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

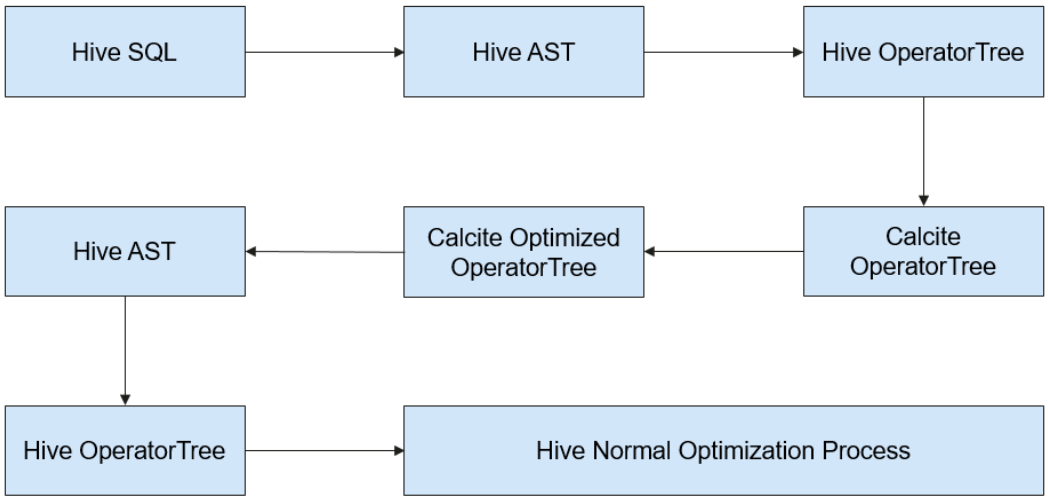
其优化目标是：

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

Hive中实现CBO的总体过程如下：

Hive使用开源组件Apache Calcite实现CBO。首先SQL语句转化成Hive的AST，然后转成Calcite可以识别的RelNodes。Calcite将RelNode中的Join顺序调整后，再由Hive将RelNode转成AST，继续Hive的逻辑优化和物理优化过程。流程图如图1-58所示：

图 1-58 实现流程图



Calcite调整Join顺序的具体过程如下：

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

代价的具体计算方法：

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

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

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

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

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

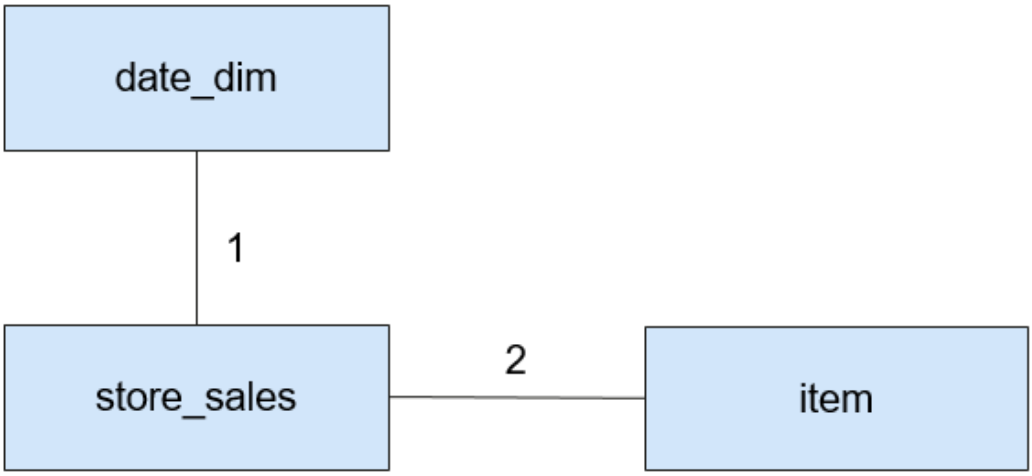
以下以TPC-DS Q3为例来介绍CBO是如何调整Join顺序的。

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

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

语句解释：这个语句由三张表来做Inner join，其中store_sales是事实表，有约2900000000条数据，date_dim是维度表，有约73000条数据，item是维度表，有约18000条数据。每一个表上都有过滤条件，其Join关系如所图1-59示：

图 1-59 Join 关系



CBO应该先选择能起到最好过滤效果的表来join。

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

表 1-10 数据过滤

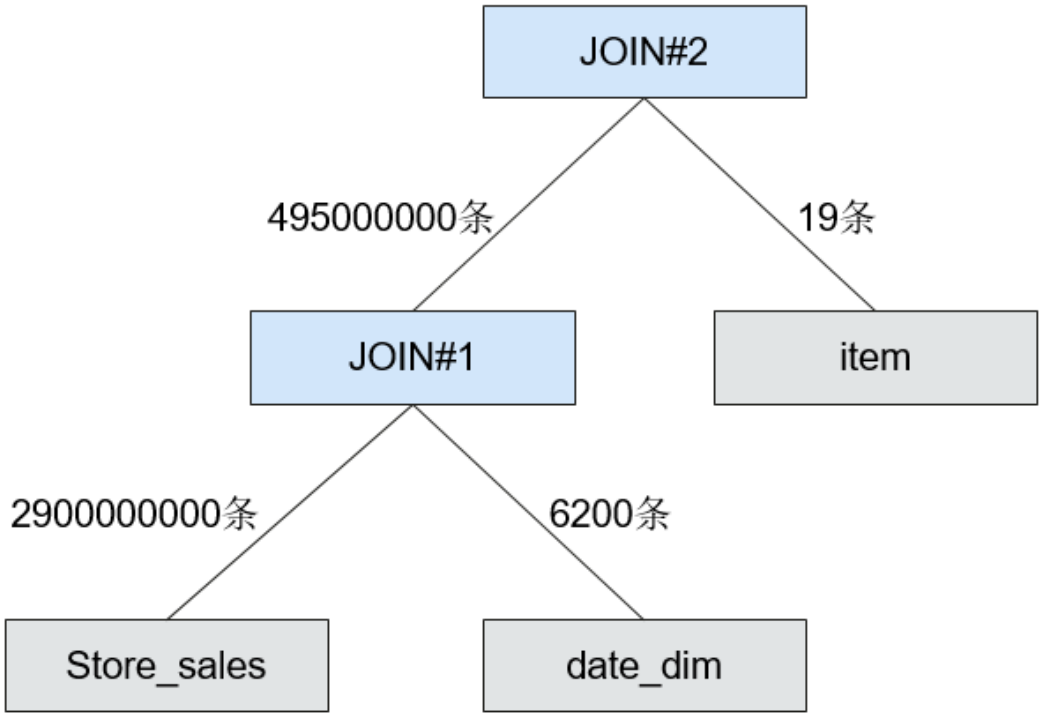
表名	原始数据条数	过滤后数据条数	选择率
date_dim	73000	6200	8.5%
item	18000	19	0.1%

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

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

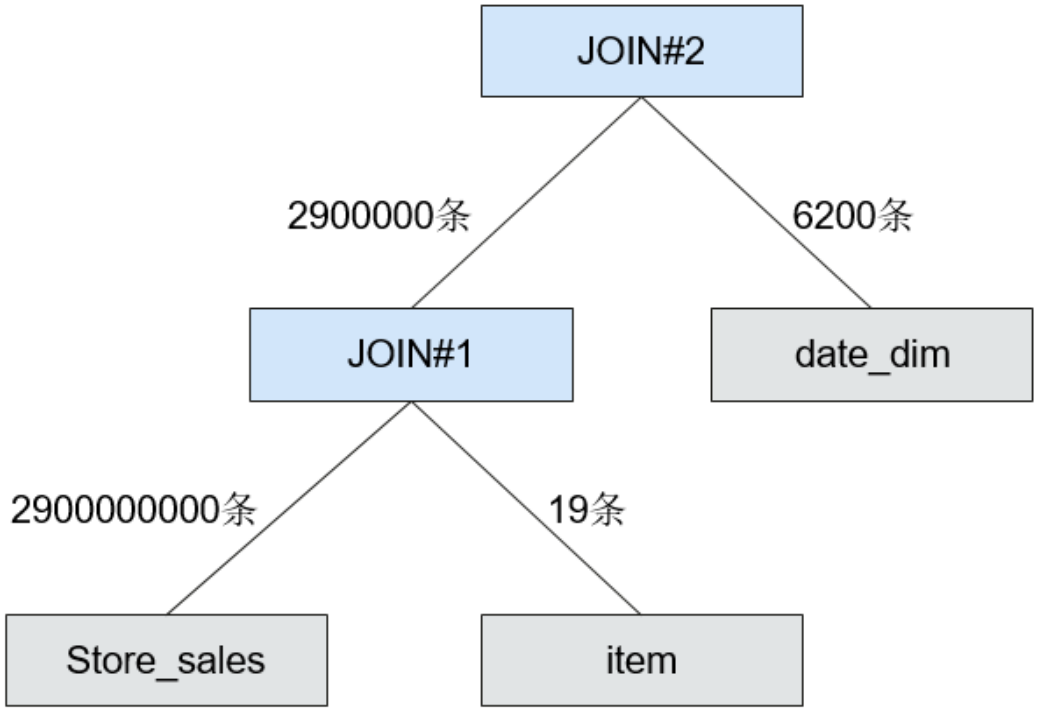
CBO未开启时的Join示意图如图1-60所示：

图 1-60 未开启 CBO



CBO开启后的Join示意图如**图1-61**所示：

图 1-61 开启 CBO



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

1.4.9.3 Hive 与其他组件的关系

Hive 与 HDFS 组件的关系

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

Hive 与 MapReduce 组件的关系

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

Hive 与 Tez 的关系

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

Hive 与 DBService 的关系

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

1.4.9.4 Hive 开源增强特性

Hive 开源增强特性：支持 HDFS Colocation

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

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

Hive 开源增强特性：支持列加密功能

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

Hive列加密机制目前支持的加密算法有两种，具体使用的算法在建表时指定。

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

Hive 开源增强特性：支持 HBase 删除功能

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

Hive 开源增强特性：支持行分隔符

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

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

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

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

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

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

Hive 开源增强特性：支持开启 Transform 功能

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

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

Hive 开源增强特性：支持创建临时函数不需要 ADMIN 权限的功能

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

用户可修改配置开关，实现创建临时函数不需要ADMIN权限。

Hive 开源增强特性：支持数据库授权

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

Hive 开源增强特性：支持列授权

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

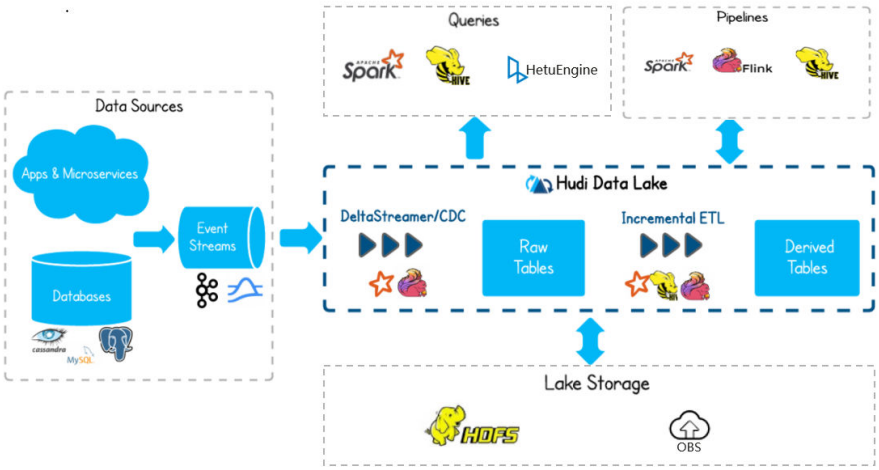
1.4.10 Hudi

Hudi是一种数据湖的存储格式，在Hadoop文件系统之上提供了更新数据和删除数据的能力以及消费变化数据的能力。支持多种计算引擎，提供IUD接口，在 HDFS的数据集上提供了插入更新和增量拉取的流原语。

说明

如需使用Hudi，请确保MRS集群内已安装Spark2x服务。

图 1-62 Hudi 基本架构



Hudi 特性

- ACID事务能力，支持实时入湖和批量入湖。
- 多种视图能力（读优化视图/增量视图/实时视图），支持快速数据分析。
- MVCC设计，支持数据版本回溯。
- 自动管理文件大小和布局，以优化查询性能准实时摄取，为查询提供最新数据。
- 支持并发读写，基于snapshot的隔离机制实现写入时可读取。
- 支持原地转表，将存量的历史表转换为Hudi数据集。

Hudi 关键技术和优势

- 可插拔索引机制：Hudi提供多种索引机制，可以快速完成对海量数据的更新和删除操作。
- 良好的生态支持：Hudi支持多种数据引擎接入包括Hive、Spark、Flink。

Hudi 支持两种表类型

- Copy On Write

  写时复制表也简称cow表，使用parquet文件存储数据，内部的更新操作需要通过重写原始parquet文件完成。
  - 优点 读取时，只读取对应分区的一个数据文件即可，较为高效
  - 缺点 数据写入的时候，需要复制一个先前的副本再在其基础上生成新的数据文件，这个过程比较耗时。且由于耗时，读请求读取到的数据相对就会滞后
- Merge On Read

  读时合并表也简称mor表，使用列格式parquet和行格式Avro两种方式混合存储数据。其中parquet格式文件用于存储基础数据，Avro格式文件（也可叫做log文件）用于存储增量数据。
  - 优点 由于写入数据先写delta log，且delta log较小，所以写入成本较低

  - 缺点 需要定期合并整理compact，否则碎片文件较多。读取性能较差，因为需要将delta log 和 老数据文件合并

Hudi 支持三种视图，针对不同场景提供相应的读能力

- Snapshot View

  实时视图：该视图提供当前hudi表最新的快照数据，即一旦有最新的数据写入hudi表，通过该视图就可以查出刚写入的新数据。

  cow表和mor均支持这种视图能力。

- Incremental View

  增量视图：该视图提供增量查询的能力，可以查询指定COMMIT之后的增量数据，可用于快速拉取增量数据。

  cow表支持该种视图能力， mor表也可以支持该视图，但是一旦mor表完成compact操作其增量视图能力消失。

- Read Optimized View

  读优化视图：该视图只会提供最新版本的parquet文件中存储的数据。

  该视图在cow表和mor表上表现不同：

  对于cow表，该视图能力和实时视图能力是一样的（ cow表只用parquet文件存数据 ）。

  对于mor表，仅访问基本文件，提供给定文件片自上次执行compact操作以来的数据， 可简单理解为该视图只会提供mor表parquet文件存储的数据，log文件里面的数据将被忽略。 该视图数据并不一定是最新的，但是mor表一旦完成compact操作，增量log数据被合入到了base数据里面，这个时候该视图和实时视图能力一样。

1.4.11 Hue

1.4.11.1 Hue 基本原理

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

Hue主要包括了文件浏览器和查询编辑器的功能：

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

通过Hue可以在界面针对组件进行以下操作：

- HDFS：
  - 查看、创建、管理、重命名、移动、删除文件/目录。
  - 上传、下载文件。
  - 搜索文件、目录、文件所有人、所属用户组；修改文件以及目录的属主和权限。
  - 手动配置HDFS目录存储策略，配置动态存储策略等操作。

- Hive：
  - 编辑、执行SQL/HQL语句；保存、复制、编辑SQL/HQL模板；解释SQL/HQL语句；保存SQL/HQL语句并进行查询。
  - 数据库展示，数据表展示。
  - 支持多种Hadoop存储。
  - 通过metastore对数据库及表和视图进行增删改查等操作。

    说明

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

    说明

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

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

Hue 结构

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

Hue由“Supervisor Process”和“WebServer”构成，“Supervisor Process”是Hue的核心进程，负责应用进程管理。“Supervisor Process”和“WebServer”通过“THRIFT/REST”接口与WebServer上的应用进行交互，如**图1-63**所示。

图 1-63 Hue 架构示意图

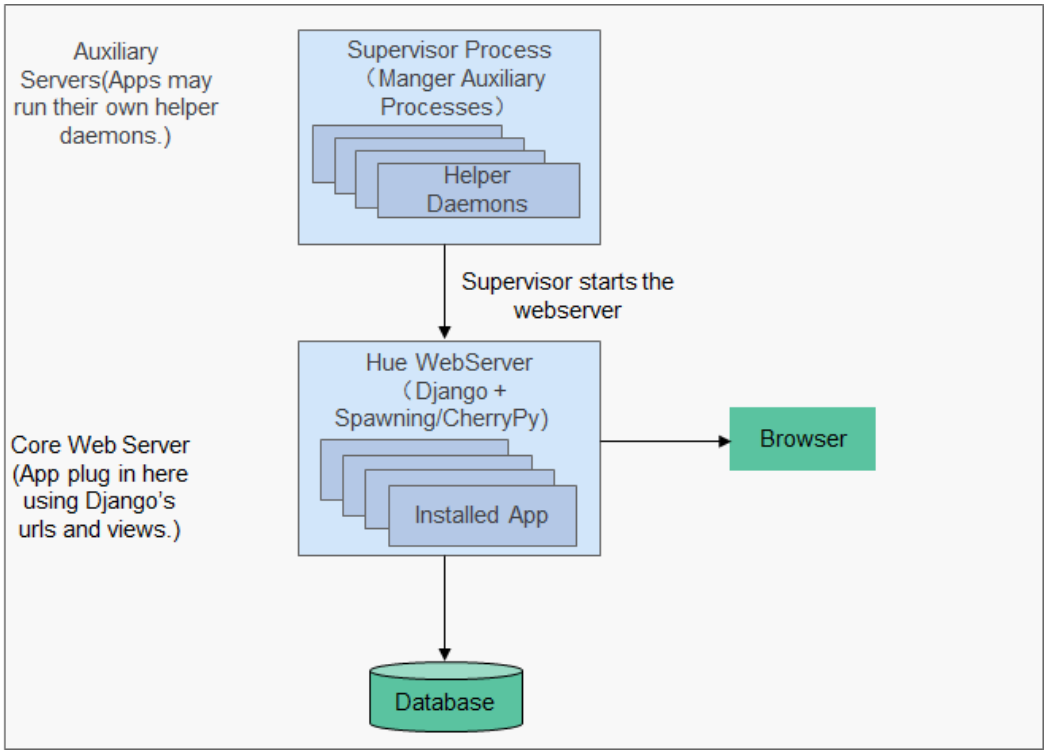


图1-63中各部分的功能说明如表1-11所示。

表 1-11 结构图说明

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

1.4.11.2 Hue 与其他组件的关系

Hue 与 Hadoop 集群的关系

Hue与Hadoop集群的交互关系如图1-64所示。

图 1-64 Hue 与 Hadoop 集群

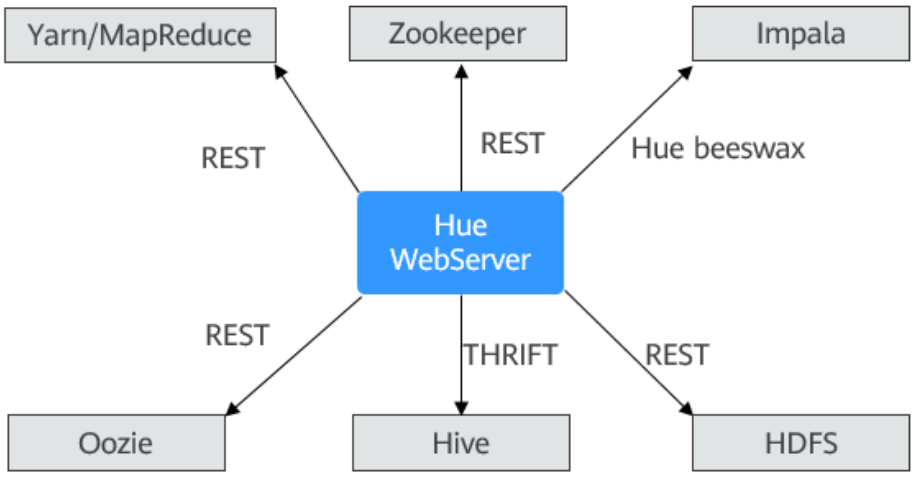


表 1-12 Hue 与其它组件的关系

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

1.4.11.3 Hue 开源增强特性

Hue 开源增强特性

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

1.4.12 Impala

Impala

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

Impala主要特点如下：

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

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

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

Impala由Impalad、Statestore、Catalog 3个角色组成。

Impala Daemon

impala daemon的进程名为 impalad，是Impala的核心进程。

impalad关键功能如下：

- 运行在所有的数据节点上。
- 读写数据文件。
- 接收来自于impala-shell命令、Hue、JDBC或者ODBC等客户端的查询请求。
- 可以并行执行来自集群中其他节点的查询请求，将中间结果返回给调度节点。
- 可以调用节点将结果返回给客户端。

impalad 进程通过持续的和statestore 通信来确认自己所在的节点是否健康 和是否可以接受新的任务请求。

Impala StateStore

负责检查impala的所有进程健康状态管理进程，进程名为statestored，当有Impalad的进程因硬件失败、网络错误、软件原因或者其他原因下线时，StateStore负责通知到其他的impalad进程，避免请求分发到不可用的节点上。

Impala Catalog Service

负责impala的元数据管理，进程名为catalogd，将元数据的变化发送到所有的Impalad进程。当创建表、加载数据或者其他的一些从Hive发起的操作后，impala查询之前需要在impalad上执行 REFRESH 或者 INVALIDATE METADATA 刷新 Catalog 上缓存的元数据信息。如果元数据变化是通过impala执行的，则不需要执行刷新。

1.4.13 Kafka

1.4.13.1 Kafka 基本原理

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

Kafka 结构

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

图 1-65 Kafka 结构

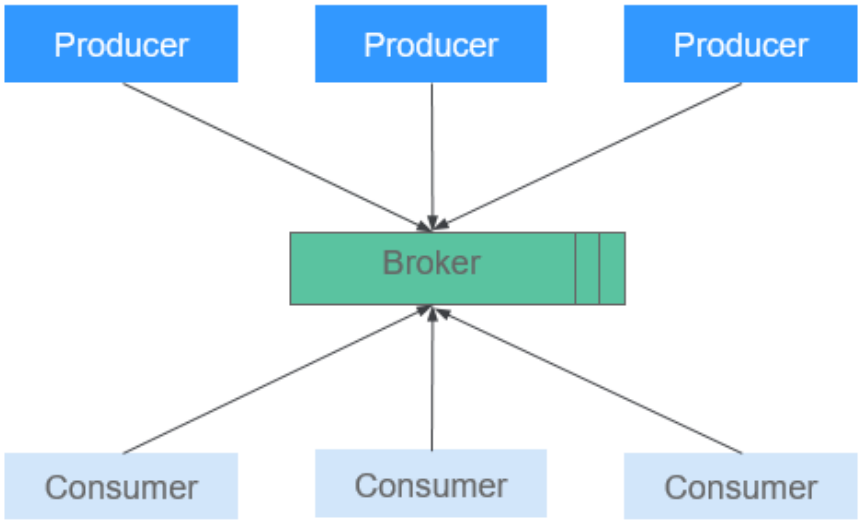
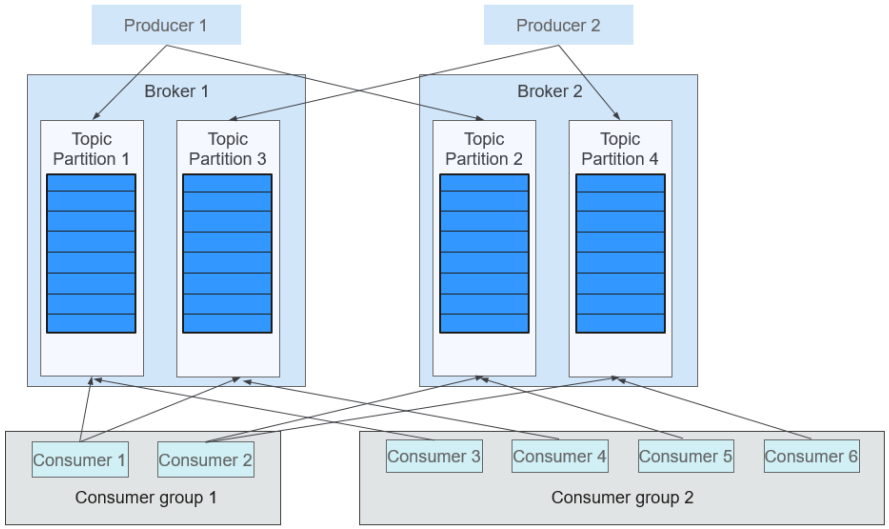


表 1-13 Kafka 结构图说明

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

各模块间关系如图1-66所示。

图 1-66 Kafka 模块间关系



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

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

Kafka 原理

- **消息可靠性**

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

- **高吞吐量**

  Kafka通过以下方式提供系统高吞吐量：

  - 数据磁盘持久化：消息不在内存中cache，直接写入到磁盘，充分利用磁盘的顺序读写性能。
  - Zero-copy：减少IO操作步骤。
  - 数据批量发送：提高网络利用率。
  - Topic划分为多个Partition，提高并发度，可以由多个Producer、Consumer数目之间的关系并发来读、写消息。Producer根据用户指定的算法，将消息发送到指定的Partition。

- **消息订阅-通知机制**

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

- **可扩展性**

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

Kafka 开源特性

- 可靠性

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

- 高吞吐

  同时为发布和订阅提供高吞吐量。

- 持久化

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

- 分布式

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

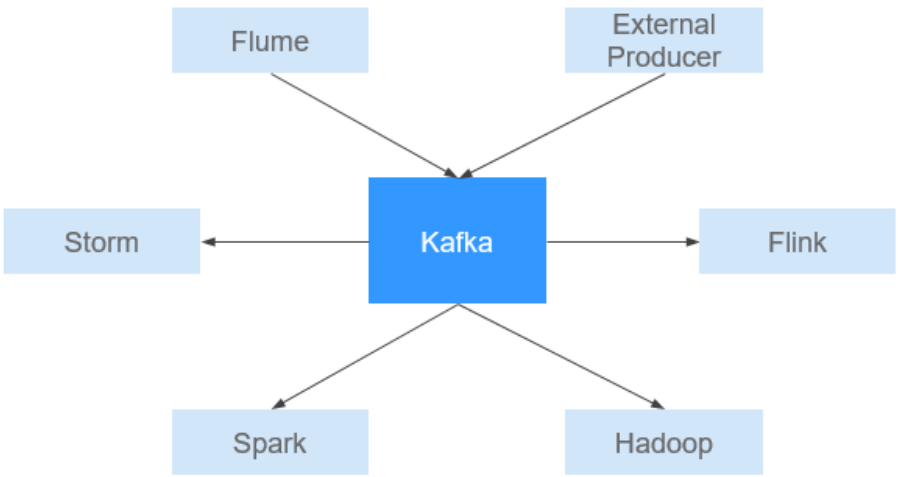
1.4.13.2 Kafka 与其他组件的关系

Kafka作为一个消息发布-订阅系统，为整个大数据平台多个子系统之间数据的传递提供了高速数据流转方式。

可以实时接受来自外部的消息，并提供给在线以及离线业务进行处理。

具体的关系如下图所示：

图 1-67 与其他组件关系



1.4.13.3 Kafka 开源增强特性

Kafka 开源增强特性

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

1.4.14 KafkaManager

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

本章节仅适用于MRS 1.8.9版本集群。

通过KafkaManager进行以下操作：

- 支持管理多个Kafka集群
- 支持界面检查集群状态（主题，消费者，偏移量，分区，副本，节点）
- 支持界面执行副本的leader选举
- 使用选择生成分区分配以选择要使用的分区方案
- 支持界面执行分区重新分配（基于生成的分区方案）
- 支持界面选择配置创建主题（支持多种Kafka版本集群）
- 支持界面删除主题（仅支持0.8.2+并设置了delete.topic.enable = true）
- 支持批量生成多个主题的分区分配，并可选择要使用的分区方案
- 支持批量运行重新分配多个主题的分区
- 支持为已有主题增加分区
- 支持更新现有主题的配置
- 可以为分区级别和主题级别度量标准启用JMX查询
- 可以过滤掉zookeeper中没有ids / owner /＆offsets /目录的使用者。

1.4.15 KrbServer 及 LdapServer

1.4.15.1 KrbServer 及 LdapServer 基本原理

KrbServer 及 LdapServer 简介

为了管理集群中数据与资源的访问控制权限，推荐以安全模式安装集群。在安全模式下，客户端应用程序在访问集群中的任意资源之前均需要通过身份认证，建立安全会

话链接。MRS通过KrbServer为所有组件提供Kerberos认证功能，实现了可靠的认证机制。

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

KrbServer 及 LdapServer 结构

用户登录时安全认证功能主要依赖于Kerberos和LDAP。

图 1-68 安全认证场景架构

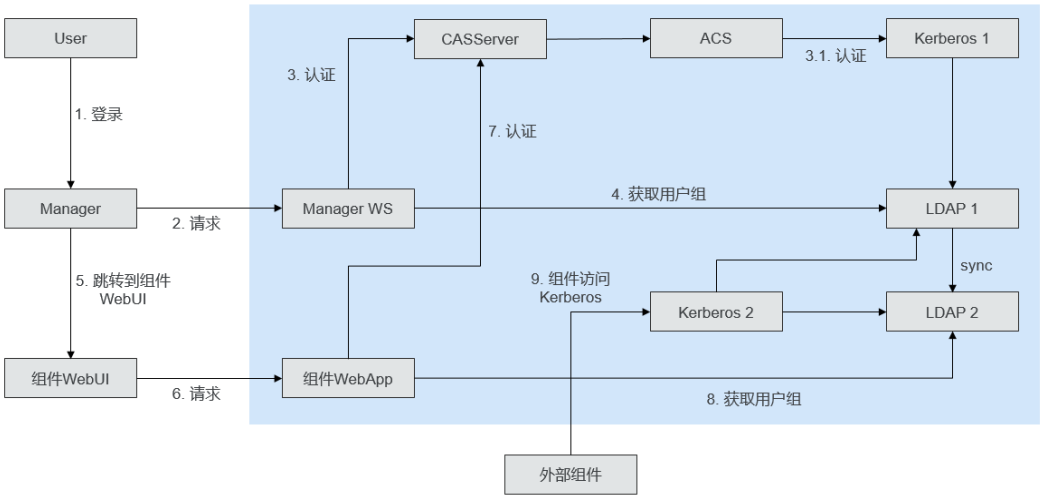


图1-68可分为三类场景：

- 登录Manager WebUI
  认证架构包含步骤1、2、3、4
- 登录组件Web UI
  认证架构包含步骤5、6、7、8
- 组件间访问
  认证架构为步骤9

表 1-14 关键模块解释

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

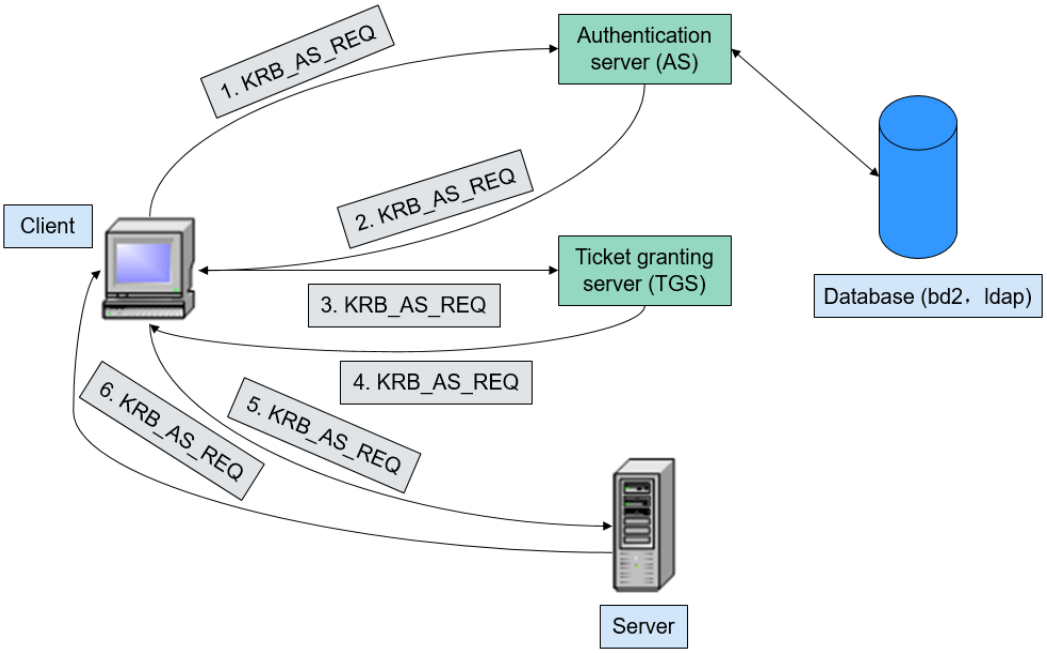
Kerberos1访问LDAP数据：以负载均衡方式访问主备LDAP1两个实例和双备LDAP2两个实例。只能在主LDAP1主实例上进行数据的写操作，可以在LDAP1或者LDAP2上进行数据的读操作。

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

KrbServer 及 LdapServer 原理

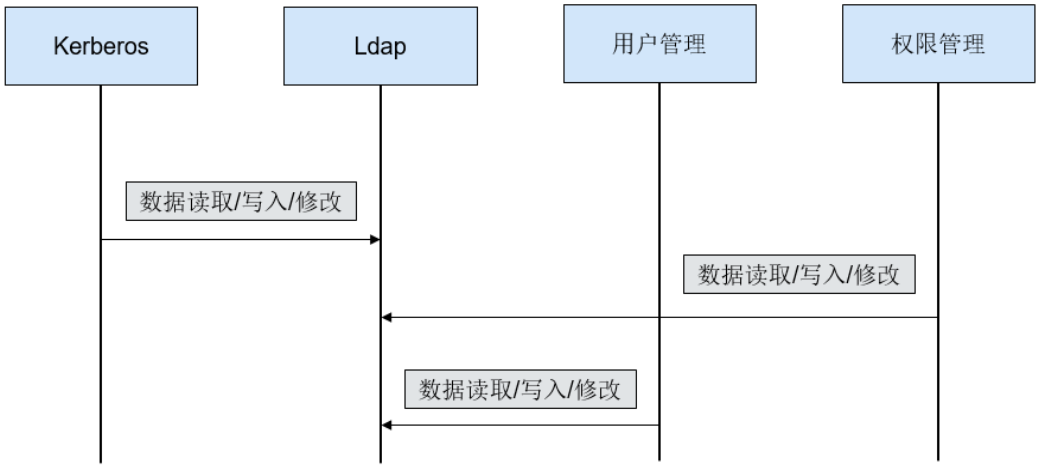
Kerberos认证

图 1-69 认证流程图



LDAP数据读写

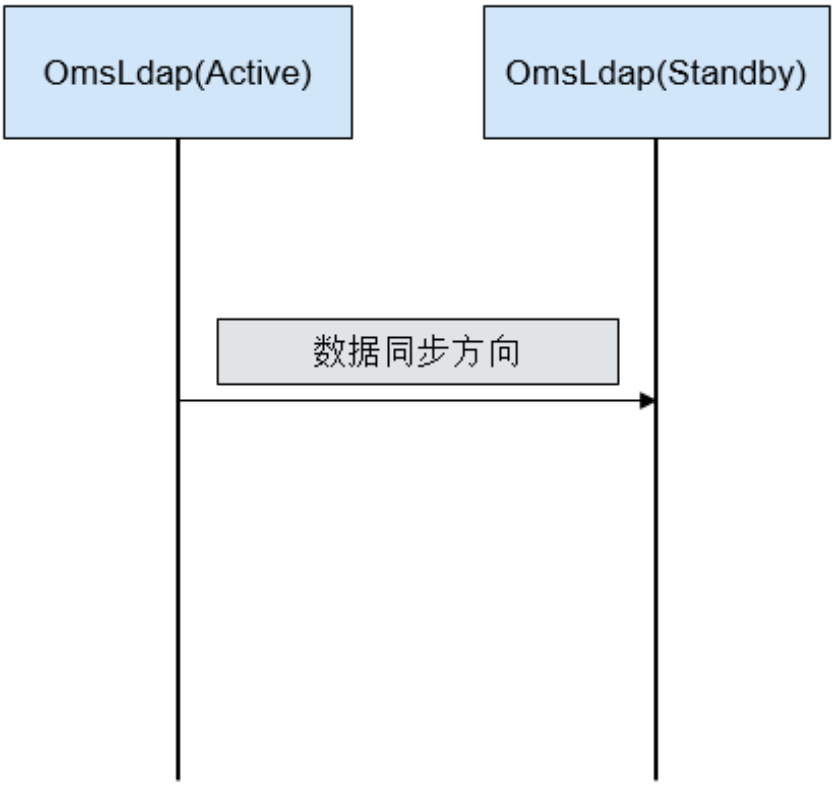
图 1-70 数据修改过程



LDAP数据同步

- 安装集群前OMS LDAP数据同步

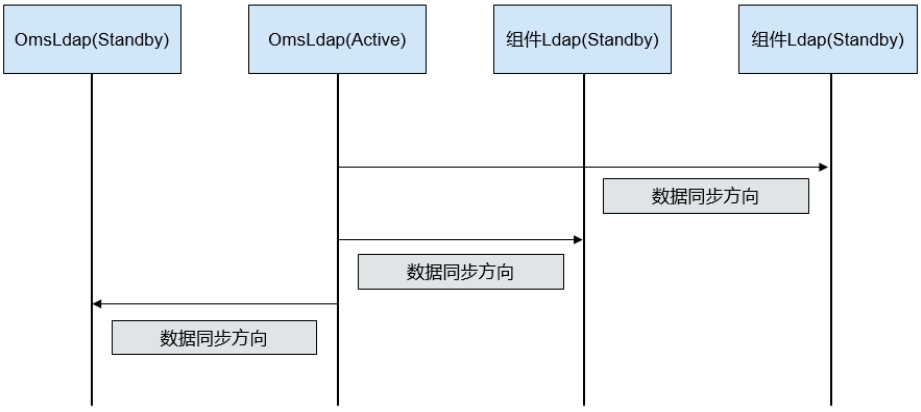
  图 1-71 OMS LDAP 数据同步

  安装集群前数据同步方向：主OMS LDAP同步到备OMS LDAP。

- 安装集群后LDAP数据同步

  图 1-72 LDAP 数据同步

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

1.4.15.2 KrbServer 及 LdapServer 开源增强特性

KrbServer 及 LdapServer 开源增强特性：集群内服务认证

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

KrbServer 及 LdapServer 开源增强特性：应用开发认证

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

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

KrbServer 及 LdapServer 开源增强特性：跨系统互信特性

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

1.4.16 Kudu

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

Kudu的设计具有以下优点：

- 能够快速处理OLAP工作负载
- 支持与MapReduce，Spark和其他Hadoop生态系统组件集成
- 与Apache Impala的紧密集成，使其成为将HDFS与Apache Parquet结合使用的更好选择
- 提供强大而灵活的一致性模型，允许您根据每个请求选择一致性要求，包括用于严格可序列化的一致性的选项
- 提供同时运行顺序读写和随机读写的良好性能
- 易于管理
- 高可用性。Master和TServer采用raft算法，该算法可确保只要副本总数的一半以上可用，tablet就可以进行读写操作。例如，如果3个副本中有2个副本或5个副本中有3个副本可用，则tablet可用。即使主tablet出现故障，也可以通过只读的副tablet提供读取服务
- 支持结构化数据模型

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

Kudu的应用场景有：

- 需要最终用户立即使用新到达数据的报告型应用

- 同时支持大量历史数据查询和细粒度查询的时序应用
- 使用预测模型并基于所有历史数据定期刷新预测模型来做出实时决策的应用

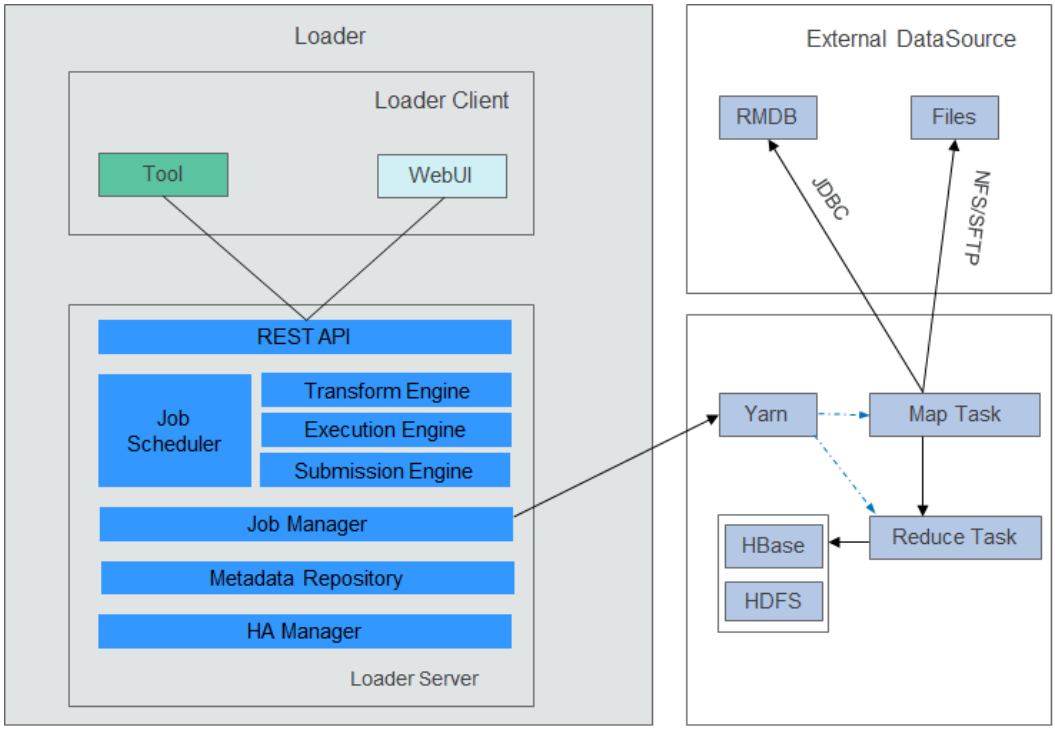
1.4.17 Loader

1.4.17.1 Loader 基本原理

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

Loader模型主要由Loader Client和Loader Server组成，如图1-73所示。

图 1-73 Loader 模型



上图中各部分的功能说明如表1-15所示。

表 1-15 Loader 模型组成

名称	描述
Loader Client	Loader的客户端，包括WebUI和CLI版本两种交互界面。
Loader Server	Loader的服务端，主要功能包括：处理客户端操作请求、管理连接器和元数据、提交MapReduce作业和监控MapReduce作业状态等。
REST API	实现RESTful（HTTP + JSON）接口，处理来自客户端的操作请求。

名称	描述
Job Scheduler	简单的作业调度模块，支持周期性的执行Loader作业。
Transform Engine	数据转换处理引擎，支持字段合并、字符串剪切、字符串反序等。
Execution Engine	Loader作业执行引擎，支持以MapReduce方式执行Loader作业。
Submission Engine	Loader作业提交引擎，支持将作业提交给MapReduce执行。
Job Manager	管理Loader作业，包括创建作业、查询作业、更新作业、删除作业、激活作业、去激活作业、启动作业、停止作业。
Metadata Repository	元数据仓库，存储和管理Loader的连接器、转换步骤、作业等数据。
HA Manager	管理Loader Server进程的主备状态，Loader Server包含2个节点，以主备方式部署。

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

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

- **数据导入到HBase**

  在MapReduce作业的Map阶段中从外部数据源抽取数据。

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

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

- **数据导入HDFS**

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

  在MapReduce作业的提交阶段，将文件从临时目录迁移到输出目录中。

- **数据导出到关系型数据库**

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

  在MapReduce作业的提交阶段，将数据从临时表迁移到正式表中。

- **数据导出到文件系统**

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

  在MapReduce作业的提交阶段，将文件从临时目录迁移到正式目录。

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

1.4.17.2 Loader 与其他组件的关系

与Loader有交互关系的组件有HDFS、HBase、Hive、Yarn、Mapreduce和ZooKeeper。

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

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

1.4.17.3 Loader 开源增强特性

Loader 开源增强特性：数据导入导出

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

- 提供数据转化功能
- 支持图形化配置转换步骤
- 支持从SFTP/FTP服务器导入数据到HDFS/OBS
- 支持从SFTP/FTP服务器导入数据到HBase表
- 支持从SFTP/FTP服务器导入数据到Phoenix表
- 支持从SFTP/FTP服务器导入数据到Hive表
- 支持从HDFS/OBS导出数据到SFTP/FTP服务器
- 支持从HBase表导出数据到SFTP/FTP服务器
- 支持从Phoenix表导出数据到SFTP/FTP服务器
- 支持从关系型数据库导入数据到HBase表
- 支持从关系型数据库导入数据到Phoenix表
- 支持从关系型数据库导入数据到Hive表
- 支持从HBase表导出数据到关系型数据库
- 支持从Phoenix表导出数据到关系型数据库
- 支持从Oracle分区表导入数据到HDFS/OBS
- 支持从Oracle分区表导入数据到HBase表
- 支持从Oracle分区表导入数据到Phoenix表
- 支持从Oracle分区表导入数据到Hive表
- 支持从HDFS/OBS导出数据到Oracle分区表
- 支持从HBase导出数据到Oracle分区表
- 支持从Phoenix表导出数据到Oracle分区表
- 在同一个集群内，支持从HDFS导数据到HBase、Phoenix表和Hive表
- 在同一个集群内，支持从HBase和Phoenix表导数据到HDFS/OBS
- 导入数据到HBase和Phoenix表时支持使用bulkload和put list两种方式
- 支持从SFTP/FTP导入所有类型的文件到HDFS，开源只支持导入文本文件

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

1.4.18 Manager

1.4.18.1 Manager 基本原理

Manager 功能

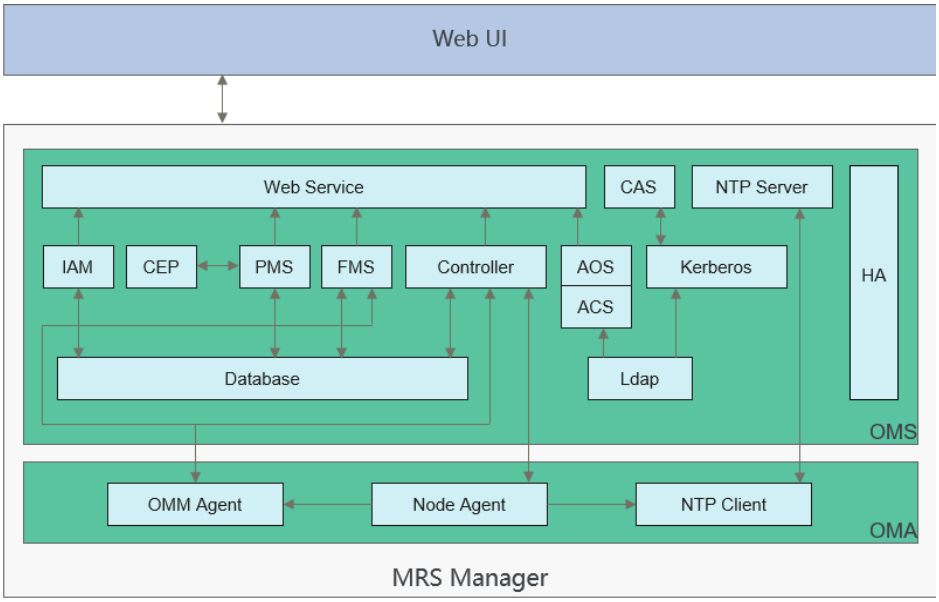
Manager是MRS的运维管理系统，为部署在集群内的服务提供统一的集群管理能力。

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

Manager 结构

Manager的整体逻辑架构如图1-74所示。

图 1-74 Manager 逻辑架构



Manager由OMS和OMA组成：

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

图1-74中各模块的说明如表1-16所示：

表 1-16 业务模块说明

模块名称	描述
Web Service	是一个部署在Tomcat下的Web服务，提供Manager的https接口，用于通过浏览器访问Manager。同时还提供基于Syslog和SNMP协议的北向接入能力。
OMS	操作维护系统的管理节点，OMS节点一般有两个，互为主备。
OMA	操作维护系统中的被管理节点，一般有多个。
Controller	Controller是Manager的控制中心，负责汇聚来自集群中所有节点的信息，统一向MRS集群管理员展示，以及负责接收来自MRS集群管理员的操作指令，并且依据操作指令所影响的范围，向集群的所有相关节点同步信息。 Manager的控制进程，负责各种管理动作的执行： 1. Web Service将各种管理动作（安装、启停服务、修改配置等）下发到Controller。 2. Controller将命令分解，分解后将动作下发到每一个Node Agent。例如启动一个服务，会涉及多个角色和实例。 3. Controller负责监控每一个动作的执行情况。
Node Agent	Node Agent存在于每一个集群节点，是Manager在单个节点的使能器。 • Node Agent代表本节点上部署的所有组件与Controller交互，实现整个集群多点到单点的汇聚。 • Node Agent是Controller对部署在该节点上组件做一切操作的使能器，其代表着Controller的功能。 Node Agent每隔3秒向Controller发送心跳信息，不支持配置时间间隔。
IAM	负责记录审计日志。在Manager的UI上每一个非查询类操作，都有对应的审计日志。
PMS	性能监控模块，搜集每一个OMA上的性能监控数据并提供查询。
CEP	汇聚功能模块。比如将所有OMA上的磁盘已用空间汇总成一个性能指标。
FMS	告警模块，搜集每一个OMA上的告警并提供查询。
OMM Agent	OMA上面性能监控和告警的Agent，负责收集该Agent Node上的性能监控数据和告警数据。
CAS	统一认证中心，登录Web Service时需要在CAS进行登录认证，浏览器通过URL自动跳转访问CAS。
AOS	权限管理模块，管理用户和用户组的权限。

模块名称	描述
ACS	用户和用户组管理模块，管理用户及用户归属的用户组。
Kerberos	在OMS与集群中各部署一个。 ● OMS Kerberos提供单点登录及Controller与Node Agent间认证的功能。 ● 集群中Kerberos提供组件用户安全认证功能，其服务名称为KrbServer，包含两种角色实例： – KerberosServer：认证服务器，为MRS提供安全认证使用。 – KerberosAdmin：管理Kerberos用户的进程。
Ldap	在OMS与集群中各部署一个。 ● OMS Ldap为用户认证提供数据存储。 ● 集群中的Ldap作为OMS Ldap的备份，其服务名称为LdapServer，角色实例为SlapdServer。
Database	Manager的数据库，负责存储日志、告警等信息。
HA	高可用性管理模块，主备OMS通过HA进行主备管理。
NTP Server NTP Client	负责同步集群内各节点的系统时钟。

1.4.18.2 Manager 关键特性

Manager 关键特性：统一监控告警

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

Manager 关键特性：统一用户权限管理

Manager提供系统中各组件的权限集中管理功能。

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

Manager 关键特性：单点登录

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

此功能统一了Manager系统用户和组件用户的管理及认证。整个系统使用LDAP管理用户，使用Kerberos进行认证，并在OMS和组件间各使用一套Kerberos和LDAP的管理机制，通过CAS实现单点登录（包括单点登入和单点登出）。用户只需要登录一次，即

可在Manager WebUI和组件Web UI之间，甚至第三方系统之间进行任务跳转操作，无需切换用户重新登录。

说明

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

Manager 关键特性：自动健康检查与巡检

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

Manager 关键特性：租户管理

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

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

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

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

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

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

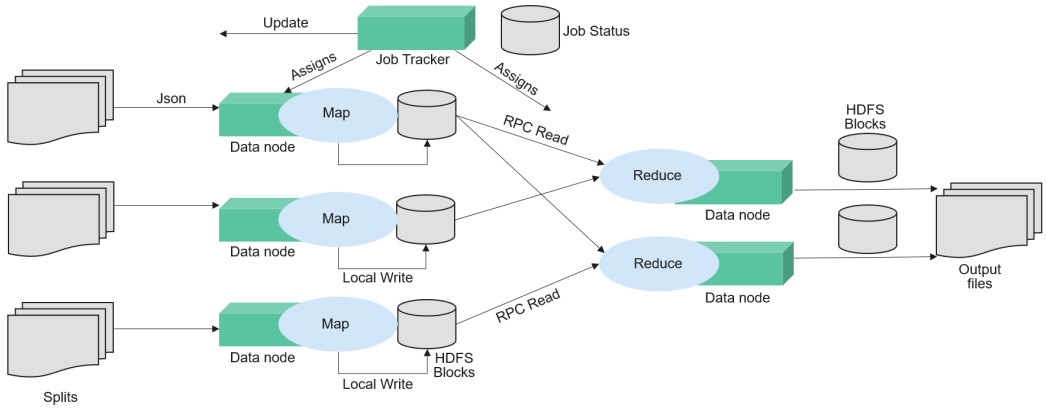
1.4.19 MapReduce

1.4.19.1 MapReduce 基本原理

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

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

图 1-75 分布式批处理引擎



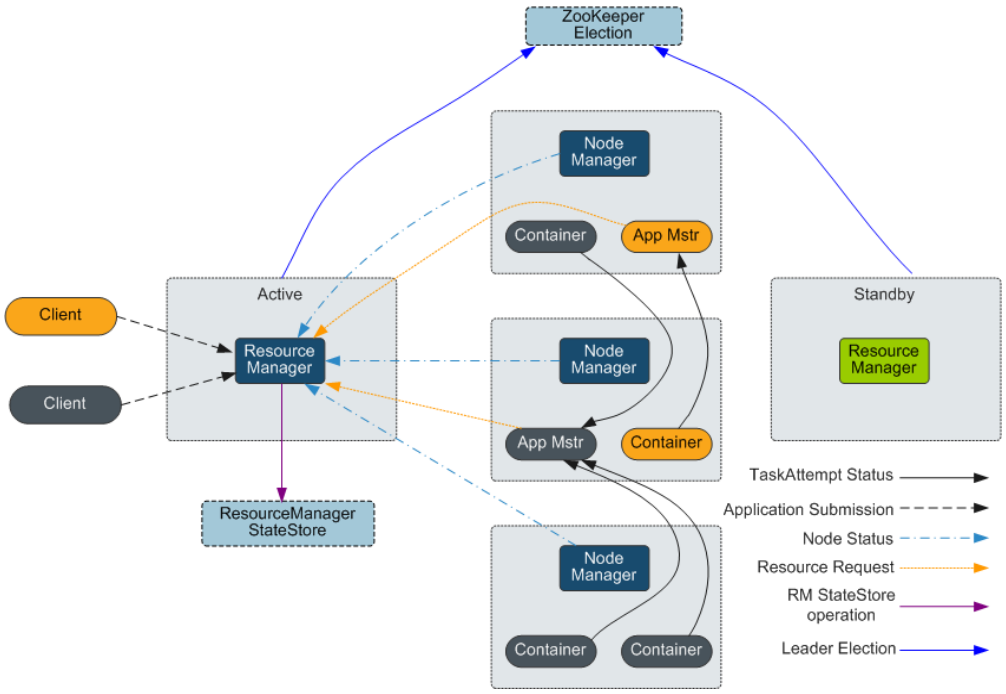
MapReduce是用于并行处理大数据集的软件框架。MapReduce的根源是函数性编程中的map和reduce函数。Map函数接受一组数据并将其转换为一个键/值对列表，输入域中的每个元素对应一个键/值对。Reduce函数接受Map函数生成的列表，然后根据它们的键缩小键/值对列表。MapReduce起到了将大事务分散到不同设备处理的能力，这样原本必须用单台较强服务器才能运行的任务，在分布式环境下也能完成。

更多信息，请参阅**MapReduce教程**。

MapReduce 结构

如**图1-76**所示，MapReduce通过实现YARN的Client和ApplicationMaster接口集成到YARN中，利用YARN申请计算所需资源。

图 1-76 Apache YARN&MapReduce 的基本架构



1.4.19.2 MapReduce 与其他组件的关系

MapReduce 和 HDFS 的关系

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

MapReduce 和 YARN 的关系

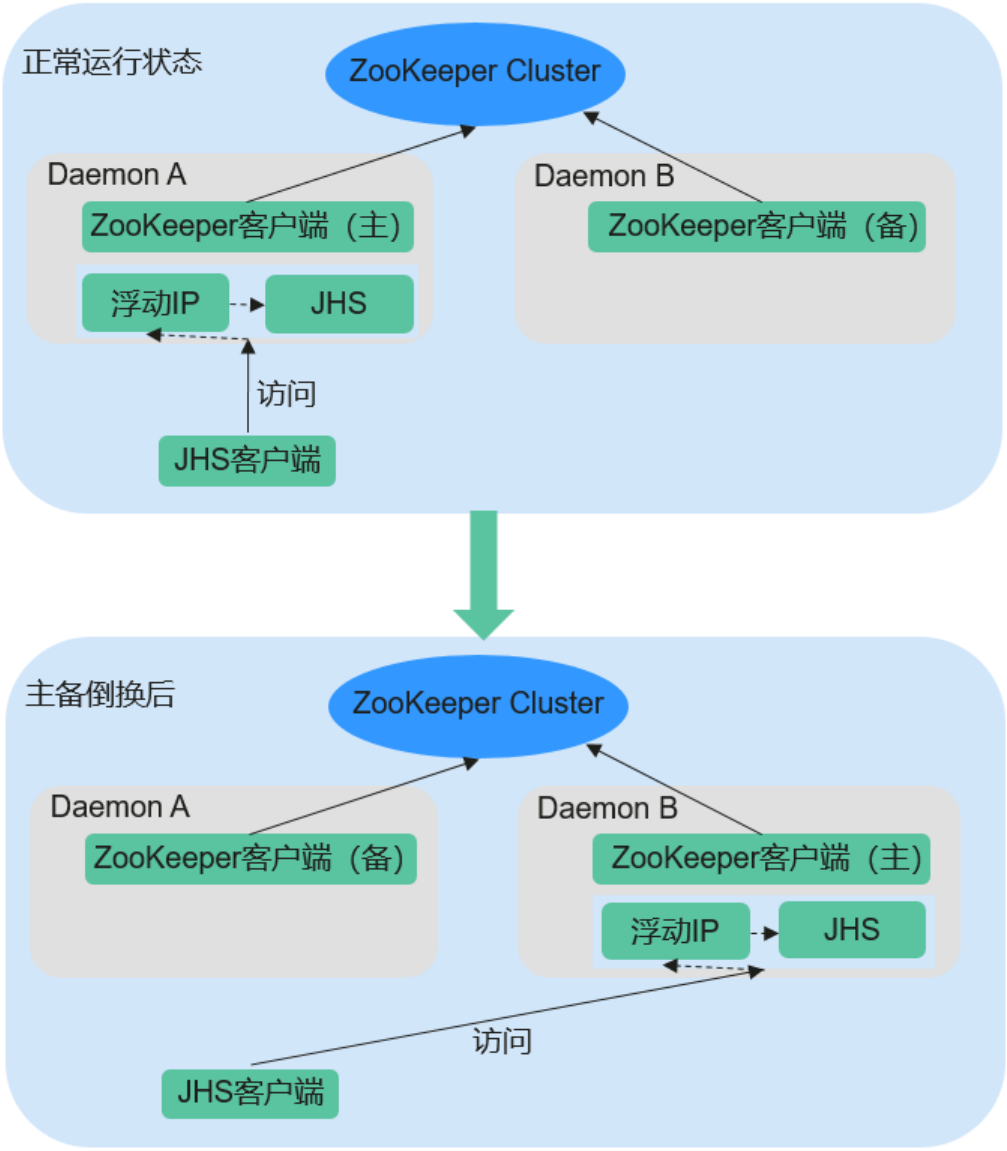
MapReduce是运行在YARN之上的一个批处理的计算框架。MRv1是Hadoop 1.0中的MapReduce实现，它由编程模型（新旧编程接口）、运行时环境（由JobTracker和TaskTracker组成）和数据处理引擎（MapTask和ReduceTask）三部分组成。该框架在扩展性、容错性（JobTracker单点）和多框架支持（仅支持MapReduce一种计算框架）等方面存在不足。MRv2是Hadoop 2.0中的MapReduce实现，它在源码级重用了MRv1的编程模型和数据处理引擎实现，但运行时环境由YARN的ResourceManager和ApplicationMaster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

1.4.19.3 MapReduce 开源增强特性

MapReduce 开源增强特性：JobHistoryServer HA 特性

JobHistoryServer（JHS）是用于查看MapReduce历史任务信息的服务器，当前开源JHS只支持单实例服务。JobHistoryServer HA能够解决JHS单点故障时，应用访问MapReduce接口无效，导致整体应用执行失败的场景，从而大大提升MapReduce服务的高可用性。

图 1-77 JobHistoryServer HA 主备倒换的状态转移过程



JobHistoryServer高可用性

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

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

下图展示了MapReduce任务的工作流程。

图 1-78 MapReduce job

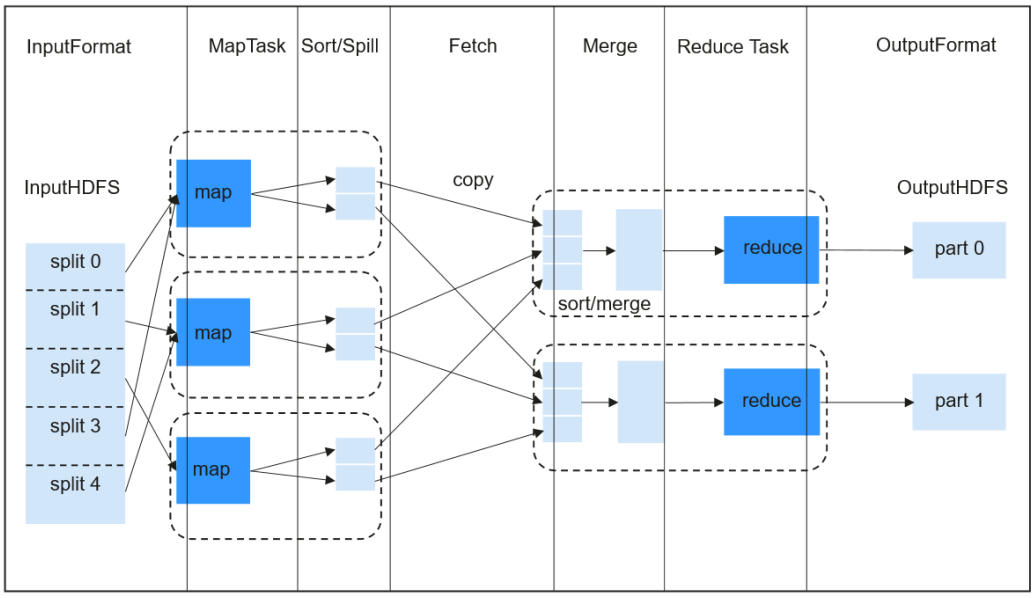
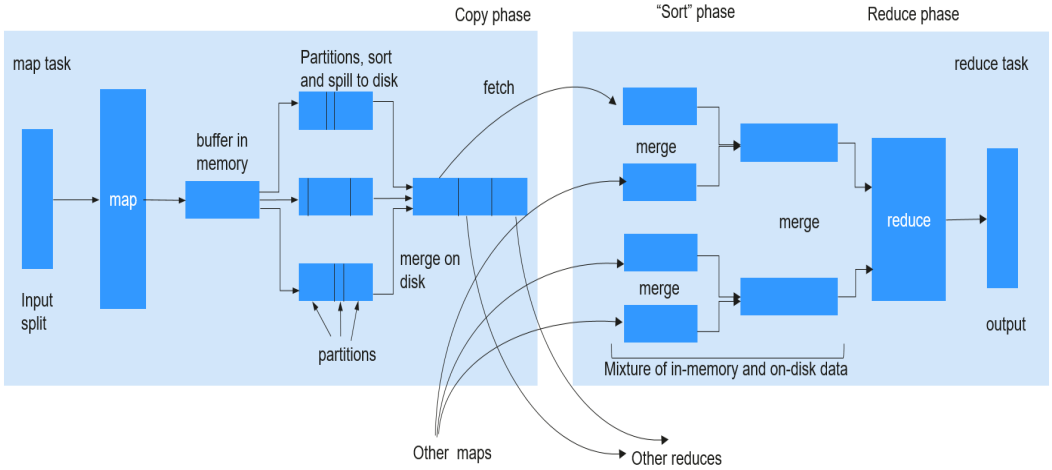


图 1-79 MapReduce job execution flow



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

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

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

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

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

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

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

日志归档

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

归档日志清理

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

归档日志浏览

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

说明

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

1.4.20 Oozie

1.4.20.1 Oozie 基本原理

Oozie 简介

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

Oozie 结构

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

基于Ext提供WEB Console，该Console仅提供对Oozie工作流的查看和监控功能。通过Oozie对外提REST方式的WS接口，Oozie client通过该接口控制（启动、停止等操作）Workflow流程，从而编排、运行Hadoop MapReduce任务，如图1-80所示。

图 1-80 Oozie 框架

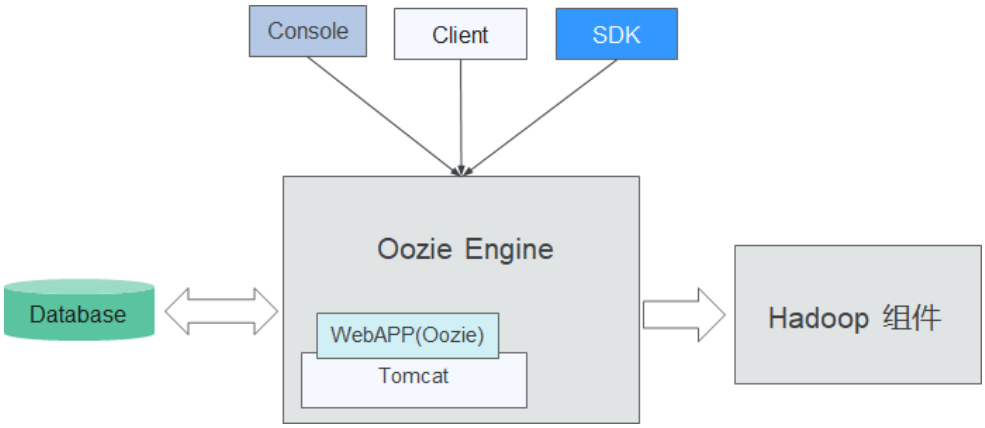


图1-80中各部分的功能说明如表1-17所示。

表 1-17 结构图说明

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

Oozie 原理

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

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

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

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

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

1.4.20.2 Oozie 开源增强特性

Oozie 开源增强特性：安全增强

支持Oozie权限管理，提供管理员与普通用户两种角色。

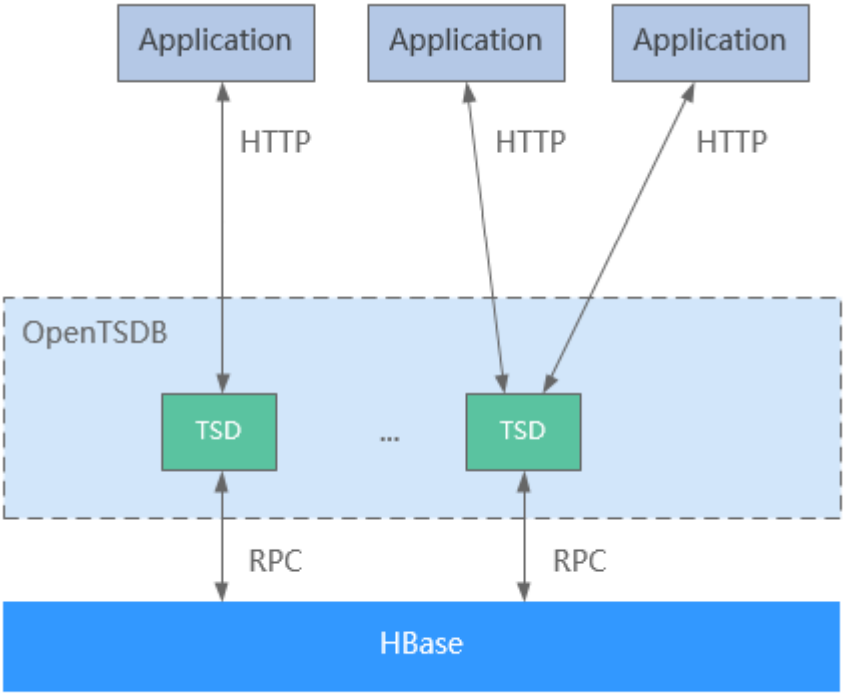
支持单点登录登出，HTTPS访问以及审计日志。

1.4.21 OpenTSDB

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

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

图 1-81 OpenTSDB 架构



OpenTSDB使用场景有如下几个特点：

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

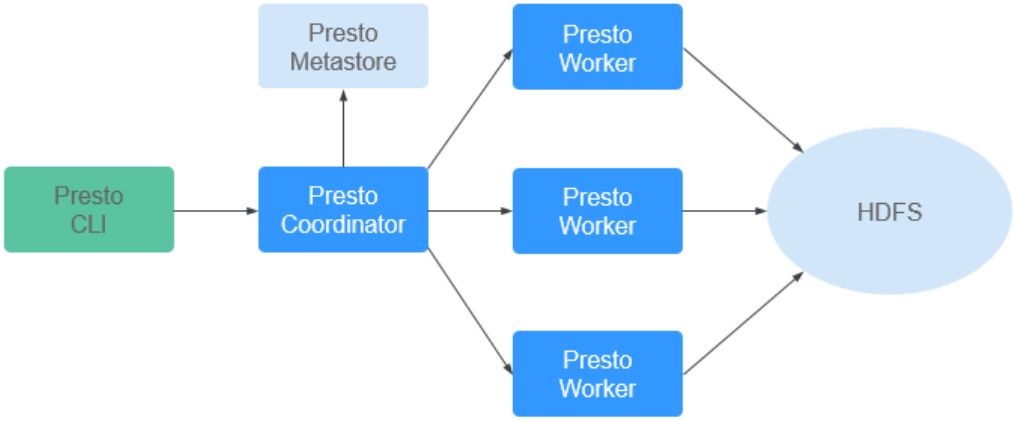
OpenTSDB提供基于HTTP的应用程序编程接口，以实现与外部系统的集成。 几乎所有OpenTSDB功能都可通过API访问，例如查询时间序列数据，管理元数据和存储数据点。详情请参见：**https://opentsdb.net/docs/build/html/api_http/index.html**。

1.4.22 Presto

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

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

图 1-82 Presto 架构



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

有关Presto的详细信息，请参见：**https://prestodb.github.io/**或**https://prestosql.io/**。

Presto 多实例

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

Presto多实例仅作用于ARM架构规格，当前单节点最多支持4个实例。

更多Presto部署信息请参考：**https://prestodb.io/docs/current/installation/deployment.html**或**https://trino.io/docs/current/installation/deployment.html**。

1.4.23 Ranger

1.4.23.1 Ranger 基本原理

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

Ranger架构如**图1-83**所示

图 1-83 Ranger 结构

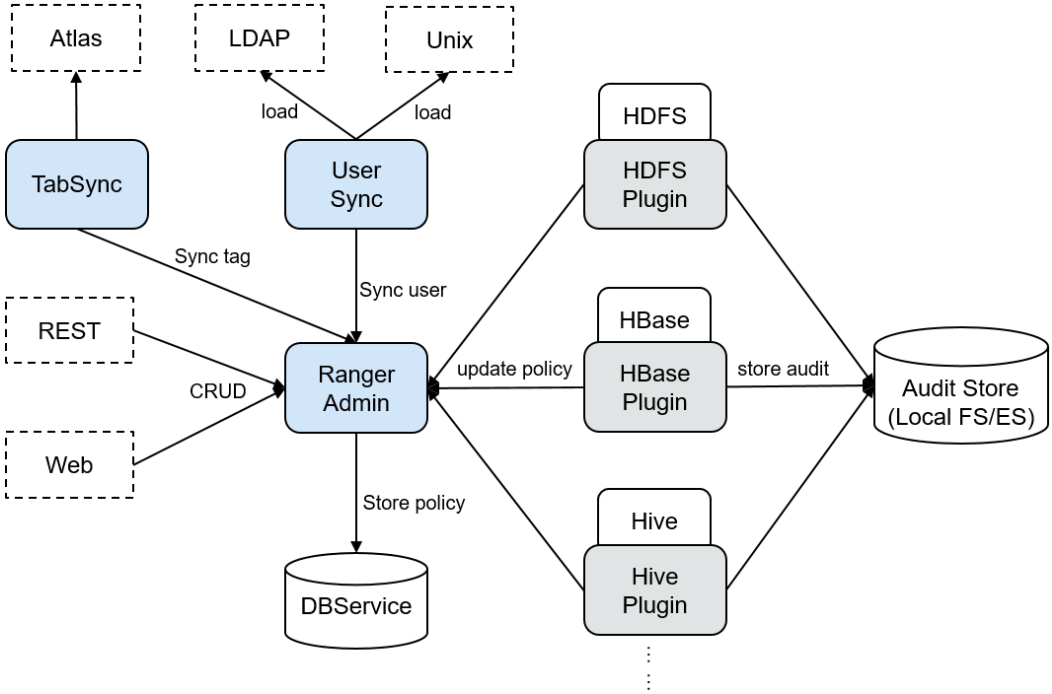


表 1-18 结构图说明

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

Ranger 原理

- 组件Ranger插件

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

- UserSync用户同步

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

- 统一审计

  Ranger插件支持记录审计日志，当前审计日志存储介质支持本地文件。

- 高可靠性

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

- 高性能

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

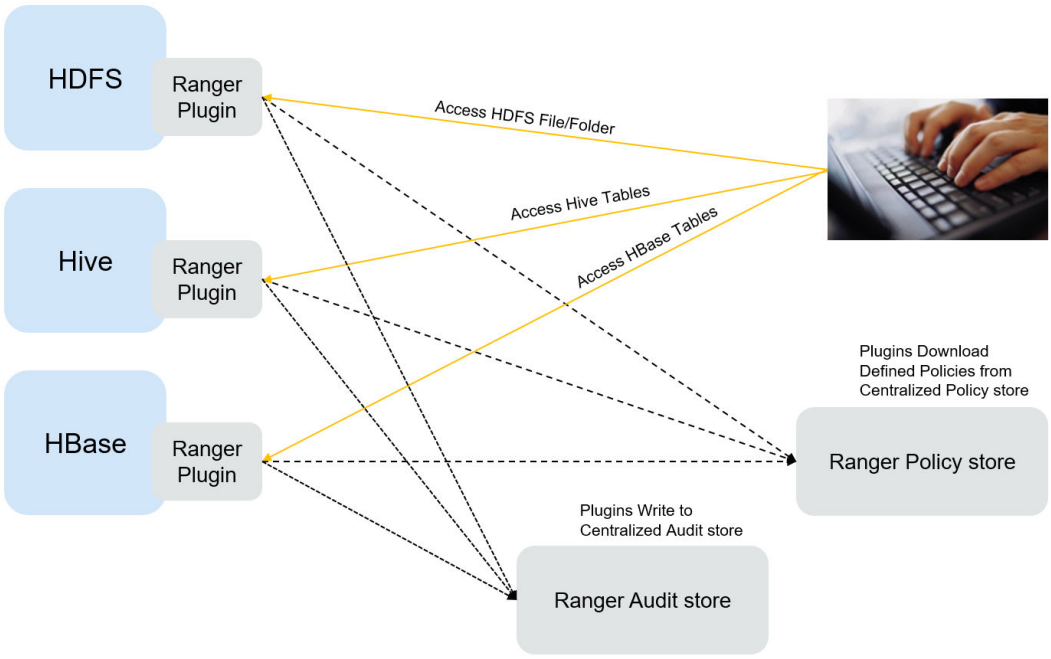
1.4.23.2 Ranger 与其他组件的关系

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

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

组件每次启动都会检查组件默认的Ranger Service是否存在，如果不存在则会创建以及为其添加默认Policy。如果用户在使用过程中误删了Service，可以重启或者滚动重启相应组件服务来恢复，如果是误删了默认Policy，可先手动删除Service，再重启组件服务。

图 1-84 Ranger 与组件的关系



1.4.24 Spark

1.4.24.1 Spark 基本原理

说明

Spark组件适用于MRS 3.x之前版本。

Spark 简介

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

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

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

Spark具有如下特点：

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

MRS服务的Spark组件具有以下优势：

- MRS服务的Spark Streaming组件支持数据实时处理能力而非定时触发。
- MRS服务的Spark组件支持Structured Streaming，支持DataSet API来构建流式应用，提供了exactly-once的语义支持，流和流的join操作支持内连接和外连接。
- MRS服务的Spark组件支持pandas_udf，可以利用pandas_udf替代pyspark中原来的udf对数据进行处理，可以减少60%-90%的处理时长（受具体操作影响）。

- MRS服务的Spark组件支持 Graph 功能，支持图计算作业使用图进行建模。
- MRS服务的SparkSQL兼容部分Hive语法（以Hive-Test-benchmark测试集上的64个SQL语句为准）和标准SQL语法（以tpc-ds测试集上的99个SQL语句为准）。

Spark 结构

Spark的结构如图1-85所示，各模块的说明如表 基本概念说明所示。

图 1-85 Spark 结构

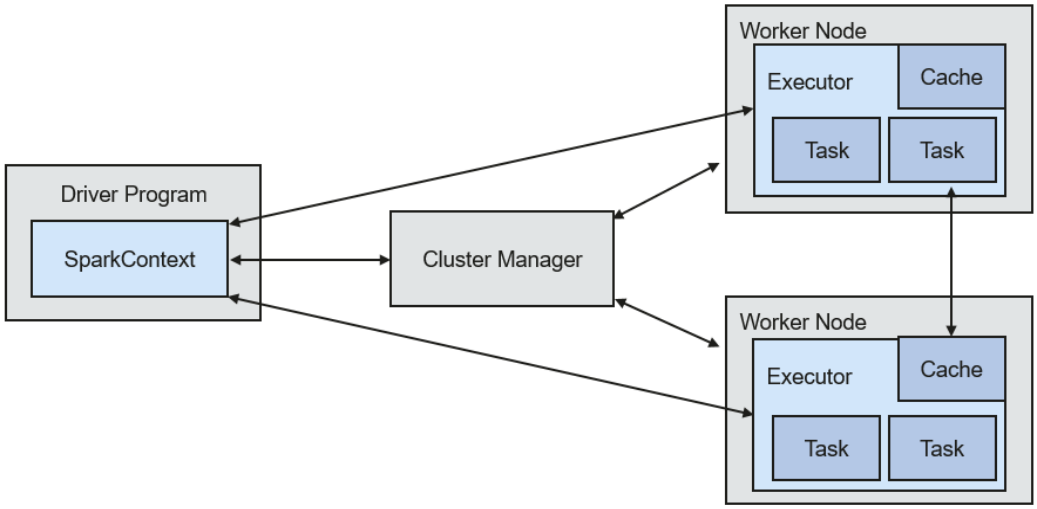


表 1-19 基本概念说明

模块	说明
Cluster Manager	集群管理器，管理集群中的资源。Spark支持多种集群管理器，Spark自带的Standalone集群管理器、Mesos或YARN，系统默认采用YARN模式。
Application	Spark应用，由一个Driver Program和多个Executor组成。
Deploy Mode	部署模式，分为cluster和client模式。cluster模式下，Driver会在集群内的节点运行；而在client模式下，Driver在客户端运行（集群外）。
Driver Program	是Spark应用程序的主进程，运行Application的main()函数并创建SparkContext。负责应用程序的解析、生成Stage并调度Task到Executor上。通常SparkContext代表Driver Program。
Executor	在Work Node上启动的进程，用来执行Task，管理并处理应用中使用到的数据。一个Spark应用一般包含多个Executor，每个Executor接收Driver的命令，并执行一到多个Task。
Worker Node	集群中负责启动并管理Executor以及资源的节点。
Job	一个Action算子（比如collect算子）对应一个Job，由并行计算的多个Task组成。

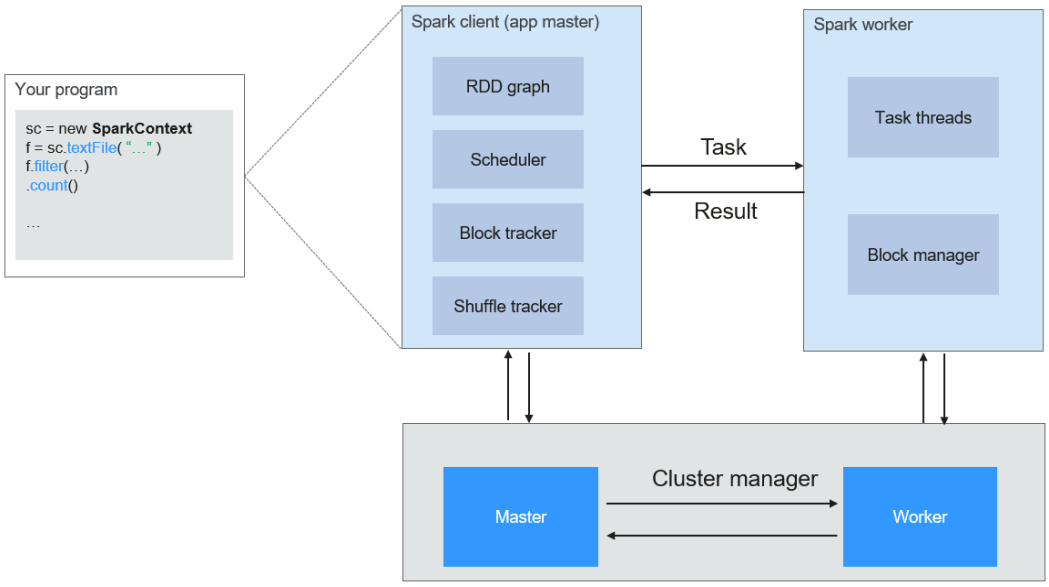
模块	说明
Stage	每个Job由多个Stage组成，每个Stage是一个Task集合，由DAG分割而成。
Task	承载业务逻辑的运算单元，是Spark平台上可执行的最小工作单元。一个应用根据执行计划以及计算量分为多个Task。

Spark 应用运行原理

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

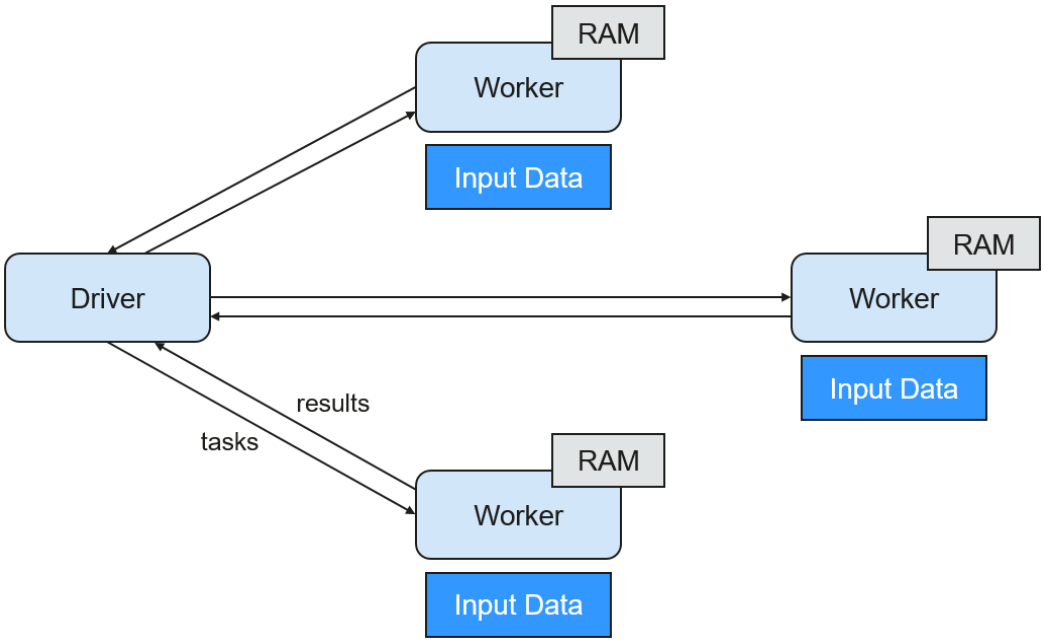
1. 应用程序（Application）是作为一个进程的集合运行在集群上的，由Driver进行协调。
2. 在运行一个应用时，Driver会去连接集群管理器（Standalone、Mesos、YARN）申请运行Executor资源，并启动ExecutorBackend。然后由集群管理器在不同的应用之间调度资源。Driver同时会启动应用程序DAG调度、Stage划分、Task生成。
3. 然后Spark会把应用的代码（传递给SparkContext的JAR或者Python定义的代码）发送到Executor上。
4. 所有的Task执行完成后，用户的应用程序运行结束。

图 1-86 Spark 应用运行架构



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

图 1-87 Spark 的 Master 和 Worker



在此结构中，有几个说明点：

- 应用之间是独立的。

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

Spark on YARN有两种部署模式：

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

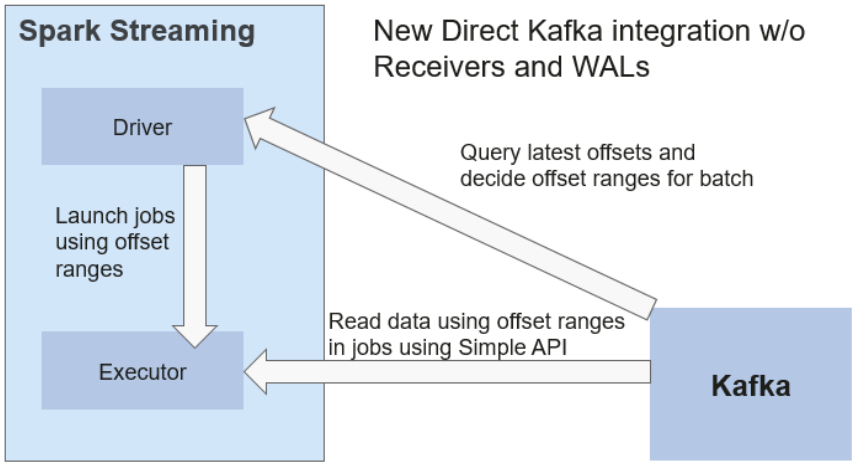
Spark Streaming 原理

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

- Direct Streaming

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

图 1-88 Direct Kafka 接口数据传输



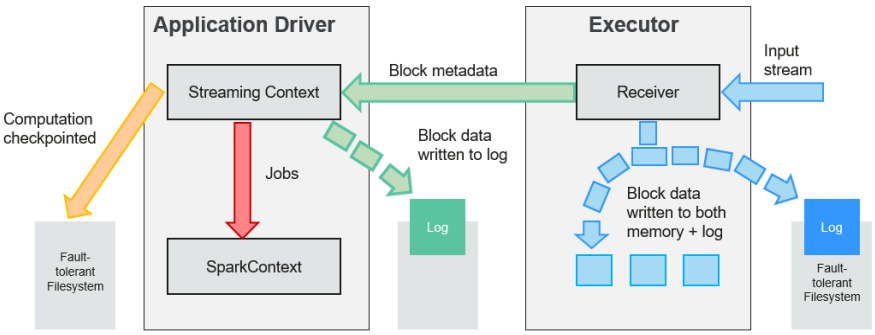
需要注意的是，Spark Streaming可以在失败后重新从Kafka中读取并处理数据段。然而，由于语义仅被处理一次，重新处理的结果和没有失败处理的结果是一致的。

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

- Receiver

  在一个Spark Streaming应用开始时（也就是Driver开始时），相关的StreamingContext（所有流功能的基础）使用SparkContext启动Receiver成为长驻运行任务。这些Receiver接收并保存流数据到Spark内存中以供处理。用户传送数据的生命周期如图1-89所示：

  图 1-89 数据传输生命周期

  a. 接收数据（蓝色箭头）

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

  b. 通知Driver（绿色箭头）

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

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

- 若启用了WAL，还包括块数据在日志中的偏移信息。

c. 处理数据（红色箭头）

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

d. 周期性地设置检查点（橙色箭头）

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

容错性

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

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

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

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

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

如果需要启用预写日志功能，可以通过如下动作实现：

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

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

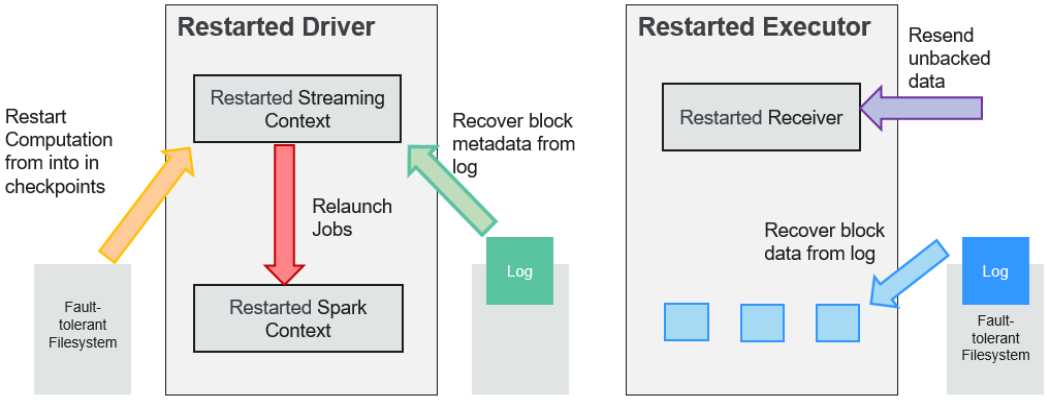
说明

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

恢复流程

当一个失败的Driver重启时，按如下流程启动：

图 1-90 计算恢复流程



1. 恢复计算（橙色箭头）

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

2. 恢复元数据块（绿色箭头）

   为了保证能够继续下去所必备的全部元数据块都被恢复。

3. 未完成作业的重新形成（红色箭头）

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

4. 读取保存在日志中的块数据（蓝色箭头）

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

5. 重发尚未确认的数据（紫色箭头）

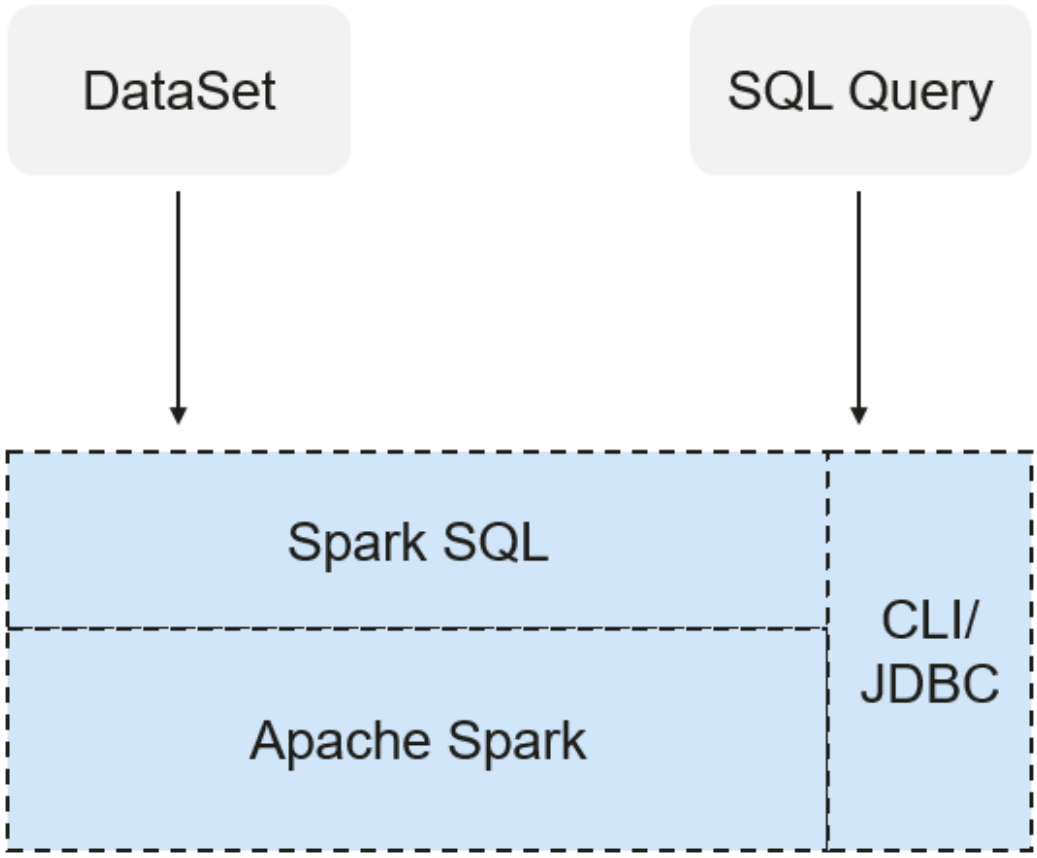
   失败时没有保存到日志中的缓存数据将由数据源再次发送。因为Receiver尚未对其确认。

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

SparkSQL 和 DataSet 原理

SparkSQL

图 1-91 SparkSQL 和 DataSet



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

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

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

Spark SQL Native DDL/DML

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

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

DataSet

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

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

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

可用于DataSet的操作分为Transformation和Action：

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

通常使用以下两种方法创建一个DataSet：

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

CLI和JDBCServer

除了API编程接口之外，Spark SQL还对外提供CLI/JDBC接口：

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

SparkSession 原理

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

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

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

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

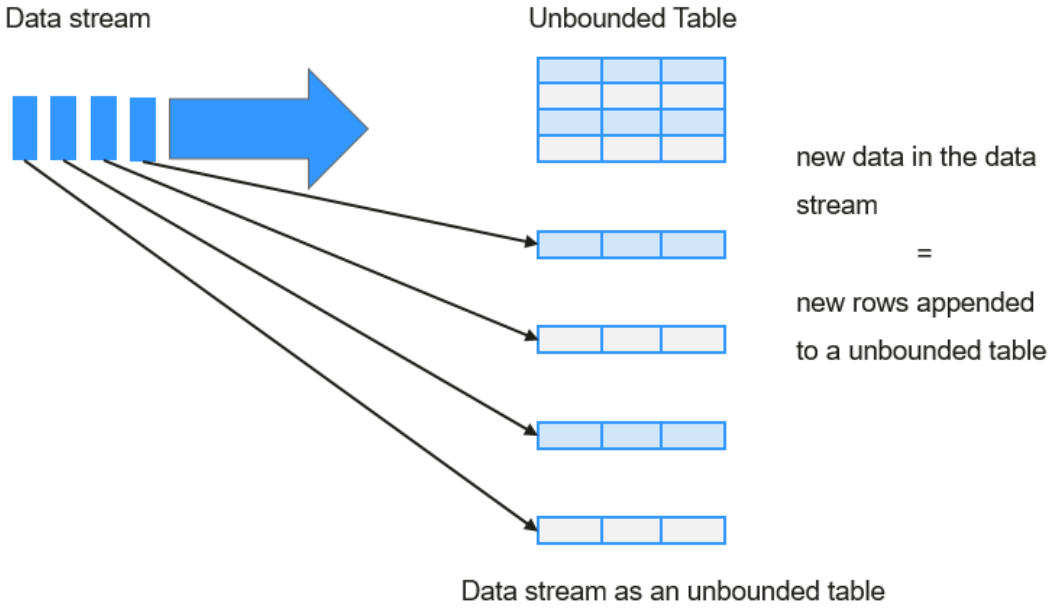
- 底层SparkContext可以通过SparkSession的SparkContext API访问。
  ```
  val sparkContext = sparkSession.sparkContext
  ```

Structured Streaming 原理

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

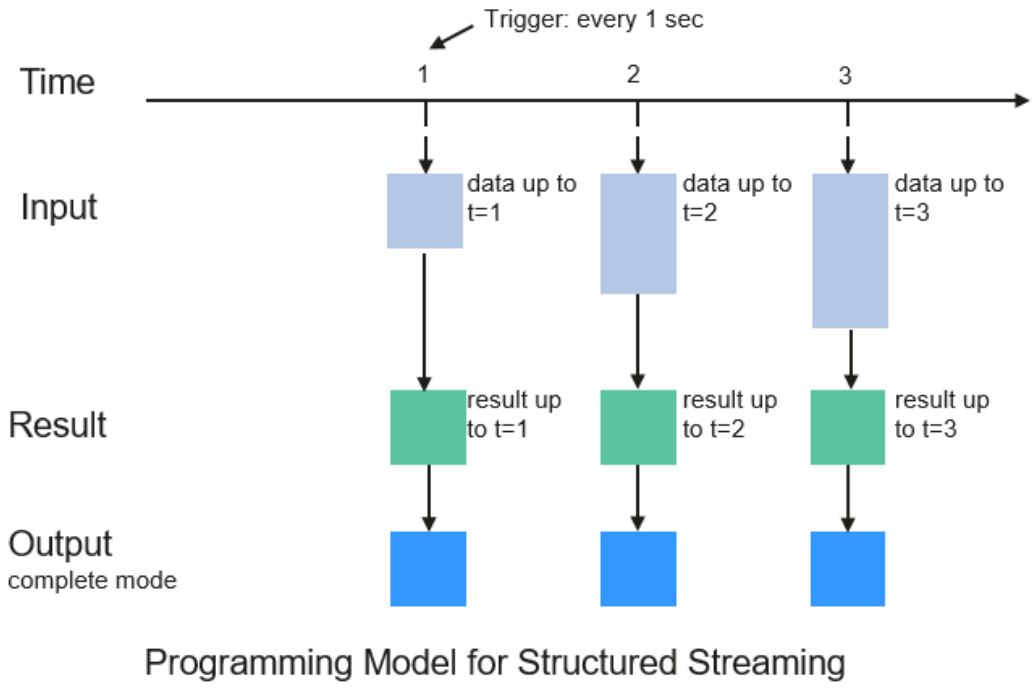
Structured Streaming的核心是将流式的数据看成一张不断增加的数据库表，这种流式的数据处理模型类似于数据块处理模型，可以把静态数据库表的一些查询操作应用在流式计算中，Spark执行标准的SQL查询，从不断增加的无边界表中获取数据。

图 1-92 Structured Streaming 无边界表



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

图 1-93 Structured Streaming 数据处理模型



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

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

Spark 常见基本概念

- **RDD**

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

  RDD的生成：

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

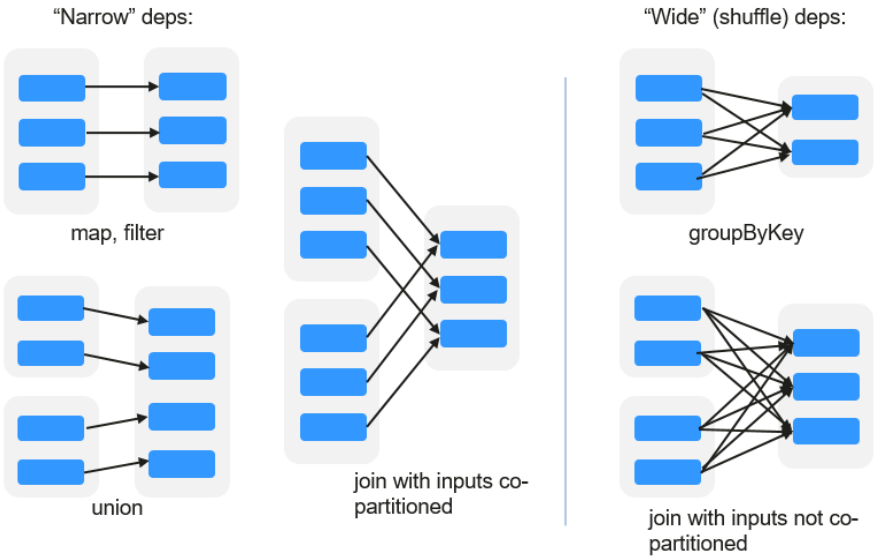
  RDD的存储：

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

- **Dependency（RDD的依赖）**

  RDD的依赖分别为：窄依赖和宽依赖。

图 1-94 RDD 的依赖



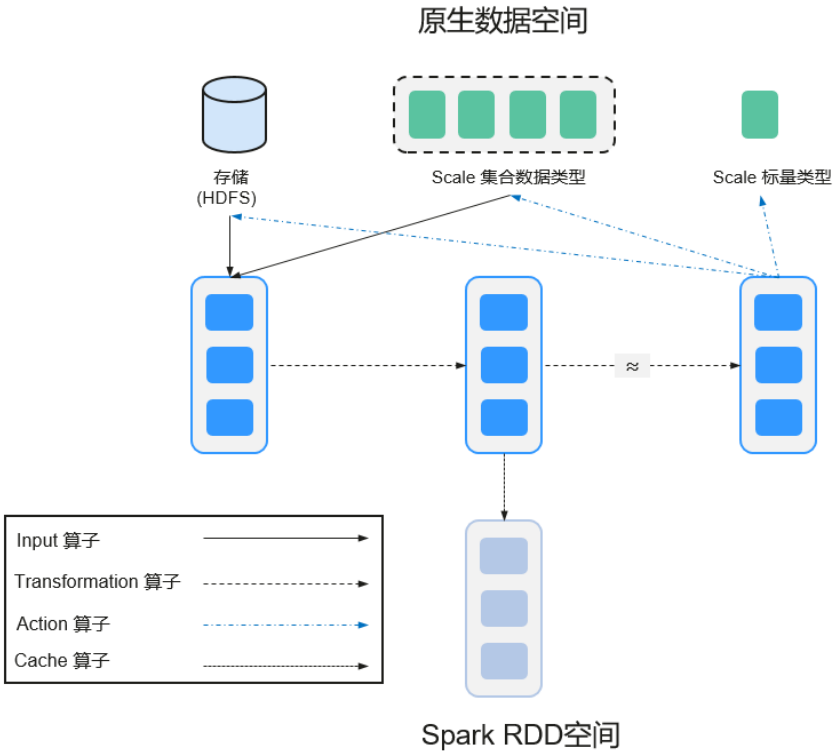
- **窄依赖：**指父RDD的每一个分区最多被一个子RDD的分区所用。
- **宽依赖：**指子RDD的分区依赖于父RDD的所有分区。

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

- **Transformation和Action（RDD的操作）**

  对RDD的操作包含Transformation（返回值还是一个RDD）和Action（返回值不是一个RDD）两种。RDD的操作流程如图1-95所示。其中Transformation操作是Lazy的，也就是说从一个RDD转换生成另一个RDD的操作不是马上执行，Spark在遇到Transformations操作时只会记录需要这样的操作，并不会去执行，需要等到有Actions操作的时候才会真正启动计算过程进行计算。Actions操作会返回结果或把RDD数据写到存储系统中。Actions是触发Spark启动计算的动因。

图 1-95 RDD 操作示例



RDD看起来与Scala集合类型没有太大差别，但数据和运行模型大相迥异。

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

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

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

c. cache算子缓存下来以备未来使用。

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

Transformation操作可以分为如下几种类型：

- 视RDD的元素为简单元素。

  输入输出一对一，且结果RDD的分区结构不变，主要是map。

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

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

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

- 视RDD的元素为Key-Value对。

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

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

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

对两个RDD基于key进行join和重组，如join、cogroup。

说明

后三种操作都涉及重排，称为shuffle类操作。

Action操作可以分为如下几种：

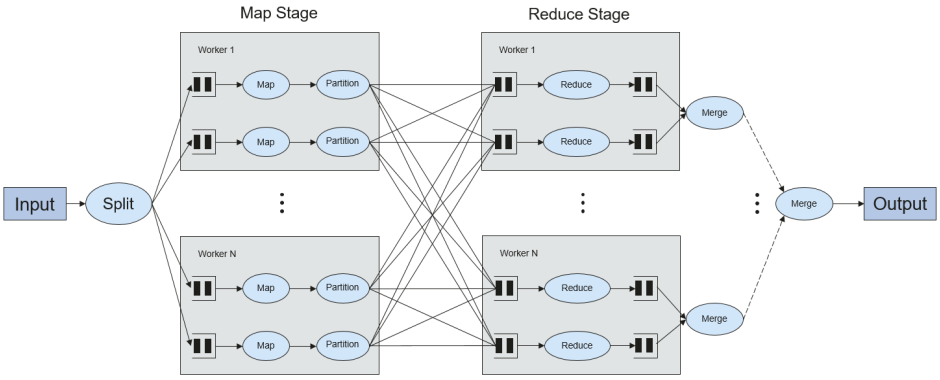
- 生成标量，如count（返回RDD中元素的个数）、reduce、fold/aggregate（返回几个标量）、take（返回前几个元素）。
- 生成Scala集合类型，如collect（把RDD中的所有元素倒入Scala集合类型）、lookup（查找对应key的所有值）。
- 写入存储，如与前文textFile对应的saveAsTextFile。
- 还有一个检查点算子checkpoint。当Lineage特别长时（这在图计算中时常发生），出错时重新执行整个序列要很长时间，可以主动调用checkpoint把当前数据写入稳定存储，作为检查点。

- **Shuffle**

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

  下图清晰地描述了MapReduce算法的整个流程。

  图 1-96 算法流程



  概念上shuffle就是一个沟通数据连接的桥梁，实际上shuffle这一部分是如何实现的呢，下面就以Spark为例讲一下shuffle在Spark中的实现。

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

- **Spark Application的结构**

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

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

    构建SparkContext对象，如：

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

    参数介绍：

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

appName：构建的Application名称。

SparkHome：集群中安装Spark的目录。

jars：应用程序代码和依赖包。

- 主体程序：处理数据

  提交Application的描述请参见：**https://archive.apache.org/dist/spark/docs/3.1.1/submitting-applications.html**

- **Spark shell命令**

  Spark基本shell命令，支持提交Spark应用。命令为：

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

  参数解释：

  --class：Spark应用的类名。

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

  application-jar：Spark应用的jar包的路径。

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

- **Spark JobHistory Server**

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

1.4.24.2 Spark HA 方案介绍

Spark 多主实例 HA 原理与实现方案

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

多主实例模式相比主备模式的HA方案，优势主要体现在对以下两种场景的改进。

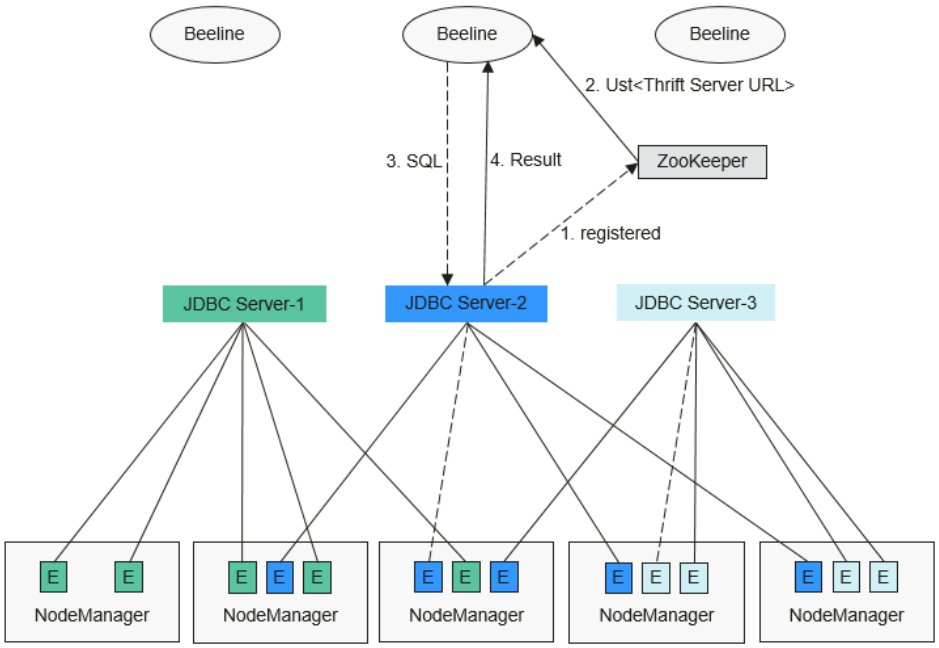
- 主备模式下，当发生主备切换时，会存在一段时间内服务不可用，该时间JDBCServer无法控制，取决于Yarn服务的资源情况。
- Spark中通过类似于HiveServer2的Thrift JDBC提供服务，用户通过Beeline以及JDBC接口访问。因此JDBCServer集群的处理能力取决于主Server的单点能力，可扩展性不够。

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

- **实现方案**

  多主实例模式的HA方案原理如下图所示。

图 1-97 Spark JDBCServer HA



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

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

多主实例方案遵循以下规则：

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

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

    - 安全模式下：

      Kinit认证方式下的JDBCURL如下所示：

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

说明

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

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

示例：安全模式下通过Beeline客户端连接时执行以下命令：

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

Keytab认证方式下的JDBCURL如下所示：

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

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

- 普通模式下：

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

  示例：普通模式下通过Beeline客户端连接时执行以下命令：

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

- 非多主实例模式

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

  示例：安全模式下通过Beeline客户端连接非多主实例模式时执行以下命令：

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

  说明

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

  多主实例模式与非多主实例模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。由于Spark JDBCServer是Hive中的HiveServer2的另外一个实现，其使用方法，请参见**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients**。

Spark 多租户 HA 方案实现

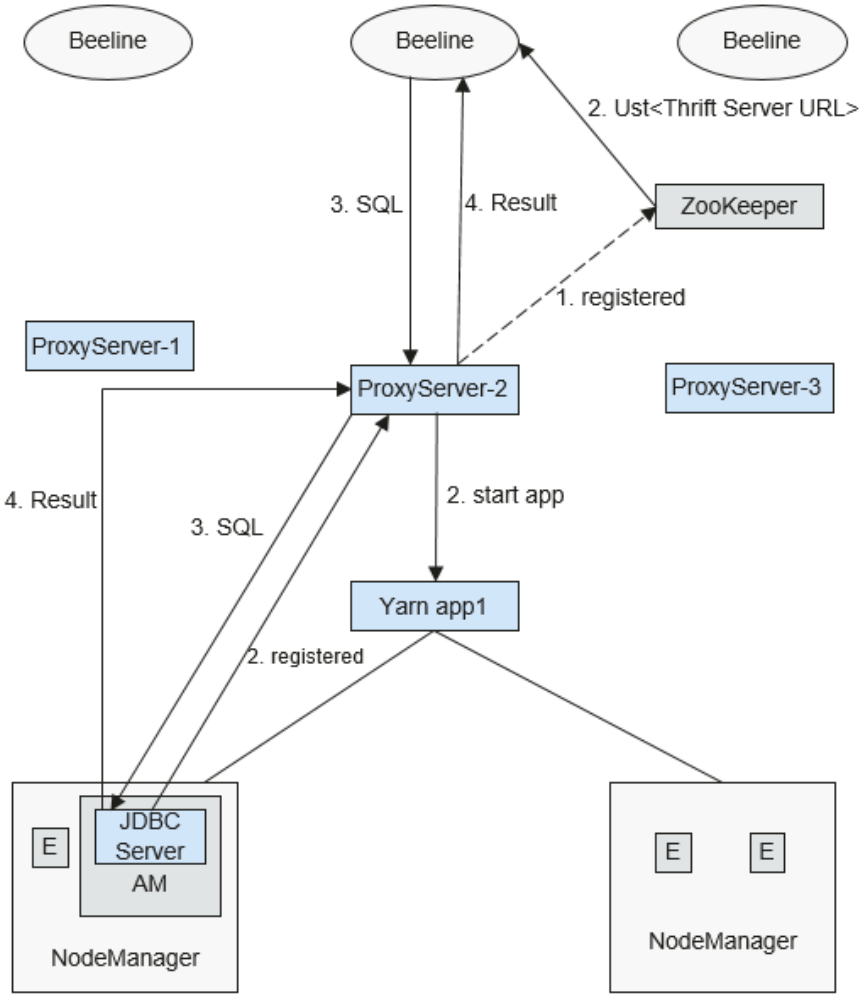
JDBCServer多主实例方案中，JDBCServer实现使用YARN-Client模式，但YARN资源队列只有一个，为了解决这种资源局限的问题，引入了多租户模式。

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

- **实现方案**

  多租户模式的HA方案原理如图1-98所示。

  图 1-98 Spark JDBCServer 多租户



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

  说明

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

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

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

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

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

- **URL连接介绍**
  - 多租户模式

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

    - 安全模式下：

      Kinit认证方式下的客户端URL如下所示：

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

      说明

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

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

      示例：安全模式下通过Beeline客户端连接时执行以下命令：

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

      Keytab认证方式下的URL如下所示：

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

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

    - 普通模式下：

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

      示例：普通模式下通过Beeline客户端连接时执行以下命令：

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

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

- 非多租户模式

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

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

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

  说明

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

  多租户模式与非多租户模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。由于Spark JDBCServer是Hive中的HiveServer2的另外一个实现，其使用方法，请参见Hive官网：**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients**。

  指定租户

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

  通过Beeline连接的命令示例如下（aaa为租户名称）：

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

1.4.24.3 Spark 与 HDFS 和 YARN 的关系

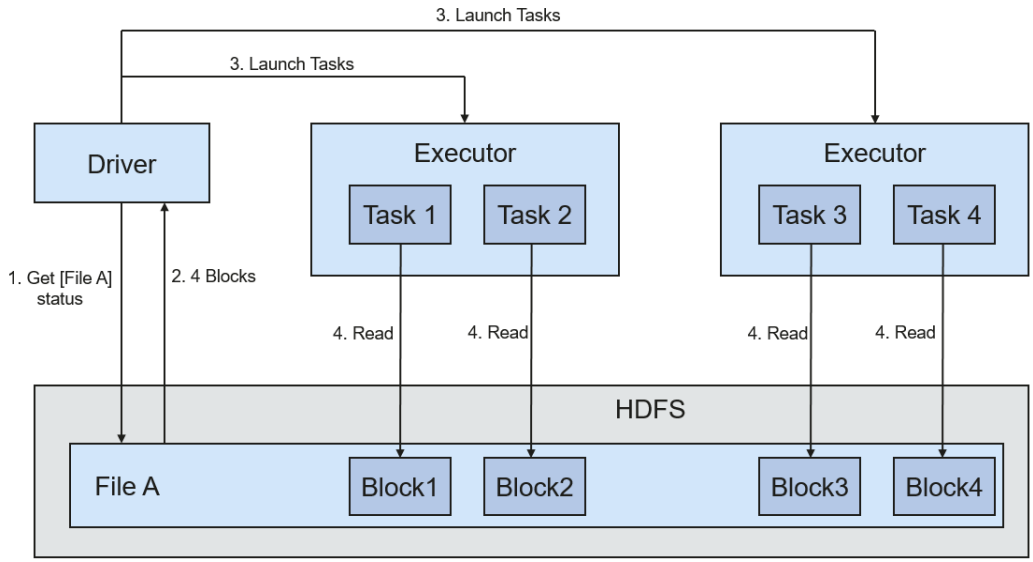
Spark 和 HDFS 的关系

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

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

读取文件的过程如**图 读取文件过程**所示。

图 1-99 读取文件过程

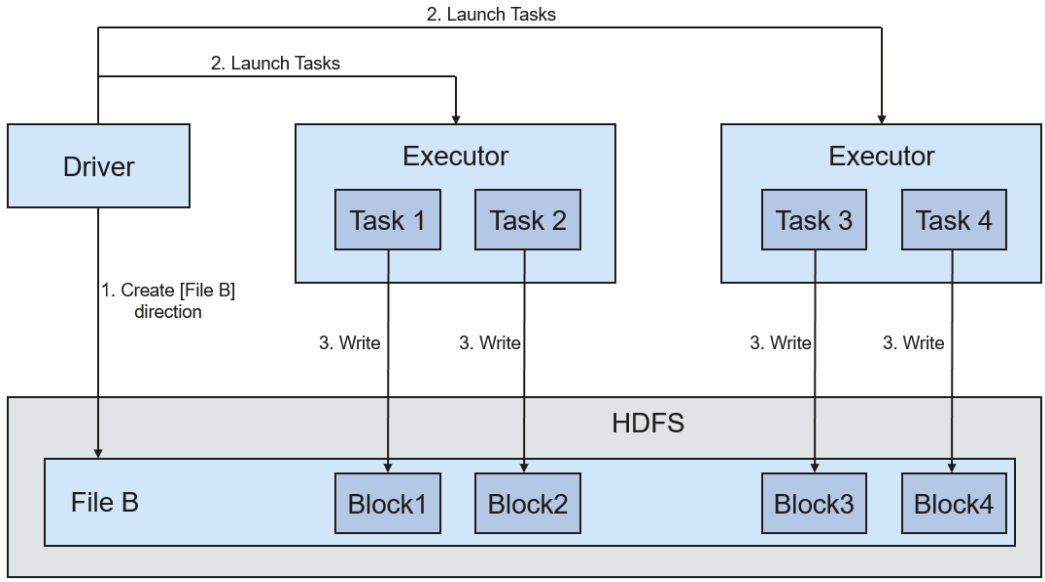


读取文件步骤的详细描述如下所示：

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

写入文件的过程如图 写入文件过程所示。

图 1-100 写入文件过程



HDFS文件写入的详细步骤如下所示：

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

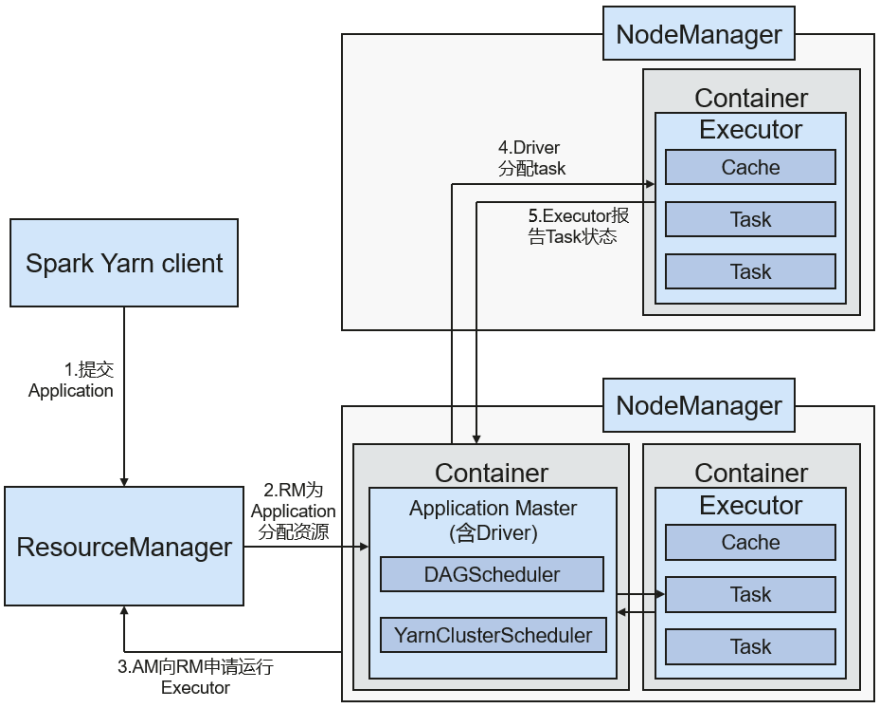
Spark 和 YARN 的关系

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

- YARN Cluster模式

  运行框架如图 Spark on yarn-cluster运行框架所示。

  图 1-101 Spark on yarn-cluster 运行框架



  Spark on yarn-cluster实现流程：

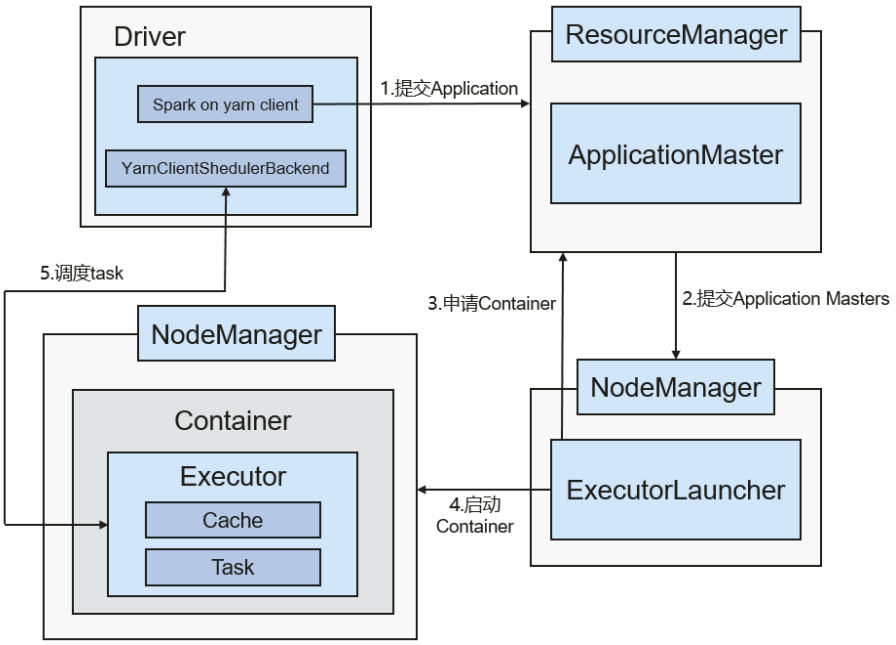
  a. 首先由客户端生成Application信息，提交给ResourceManager。
  b. ResourceManager为Spark Application分配第一个Container(ApplicationMaster)，并在该Container上启动Driver。
  c. ApplicationMaster向ResourceManager申请资源以运行Container。

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

- YARN Client模式

  运行框架如图 Spark on yarn-client运行框架所示。

图 1-102 Spark on yarn-client 运行框架



Spark on yarn-client实现流程：

说明

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

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

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

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

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

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

   说明

   正在运行的container不会被挂起释放资源。

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

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

场景描述

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

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

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

表 1-20 跨源查询增加特性对比

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

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

注意事项

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

1.4.25 Spark2x

1.4.25.1 Spark2x 基本原理

说明

Spark2x组件适用于MRS 3.x及后续版本。

简介

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

Spark的特点如下：

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

结构

Spark的架构如**图1-103**所示，各模块的说明如**表1-21**所示。

图 1-103 Spark 架构

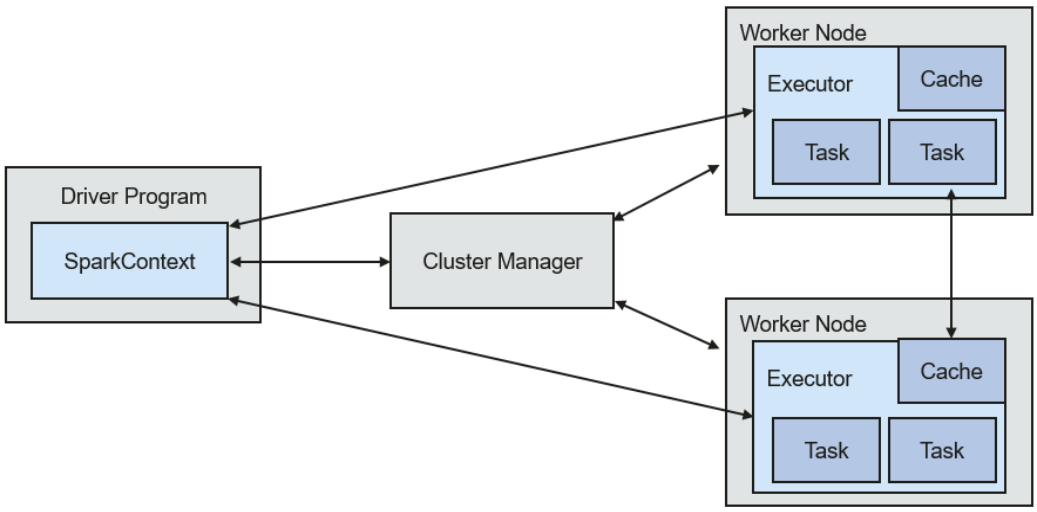


表 1-21 基本概念说明

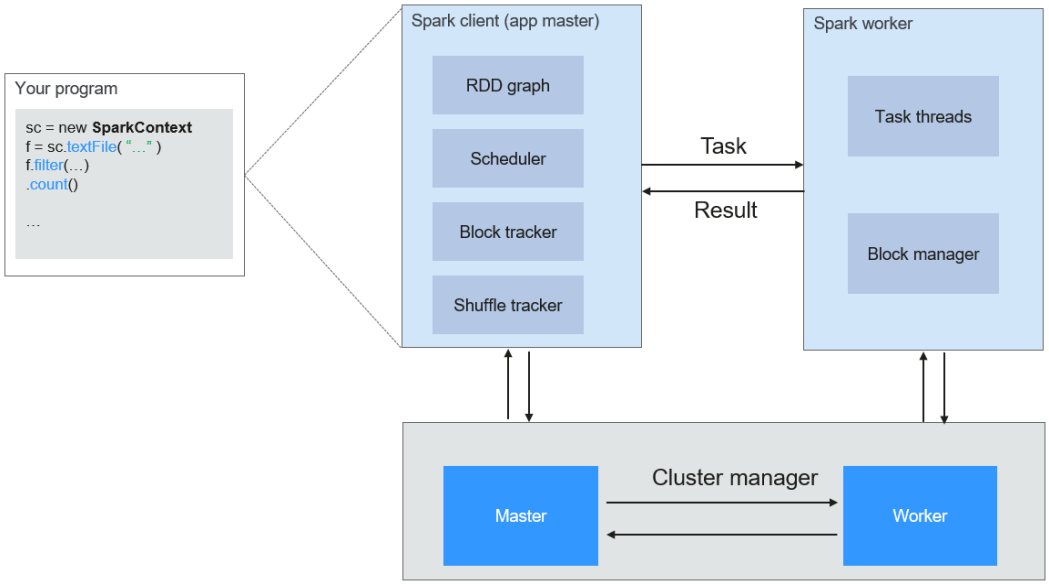
模块	说明
Cluster Manager	集群管理器，管理集群中的资源。Spark支持多种集群管理器，Spark自带的Standalone集群管理器、Mesos或YARN。Spark集群默认采用YARN模式。
Application	Spark应用，由一个Driver Program和多个Executor组成。
Deploy Mode	部署模式，分为cluster和client模式。cluster模式下，Driver会在集群内的节点运行；而在client模式下，Driver在客户端运行（集群外）。
Driver Program	是Spark应用程序的主进程，运行Application的main()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

Spark 原理

Spark的应用运行架构如图1-104所示，运行流程如下所示：

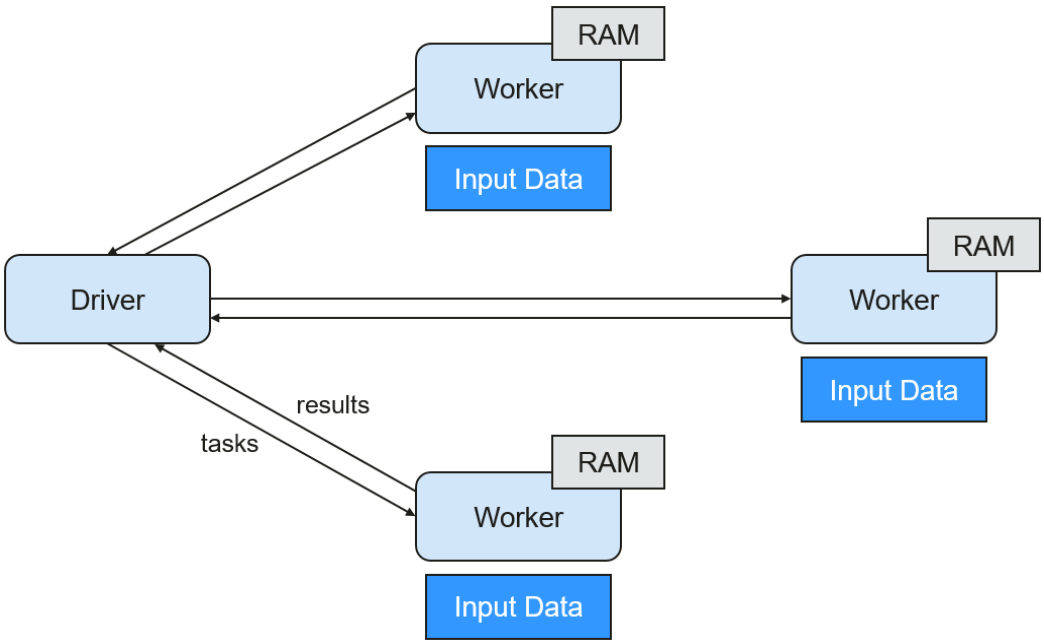
1. 应用程序（Application）是作为一个进程的集合运行在集群上的，由Driver进行协调。
2. 在运行一个应用时，Driver会去连接集群管理器（Standalone、Mesos、YARN）申请运行Executor资源，并启动ExecutorBackend。然后由集群管理器在不同的应用之间调度资源。Driver同时会启动应用程序DAG调度、Stage划分、Task生成。
3. 然后Spark会把应用的代码（传递给SparkContext的JAR或者Python定义的代码）发送到Executor上。
4. 所有的Task执行完成后，用户的应用程序运行结束。

图 1-104 Spark 应用运行架构



Spark采用Master和worker的模式，如图1-105所示。用户在Spark客户端提交应用程序，调度器将Job分解为多个Task发送到各个Worker中执行，各个Worker将计算的结果上报给Driver（即Master），Driver聚合结果返回给客户端。

图 1-105 Spark 的 Master 和 Worker



在此结构中，有几个说明点：

- 应用之间是独立的。

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

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

Spark on YARN有两种部署模式：

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

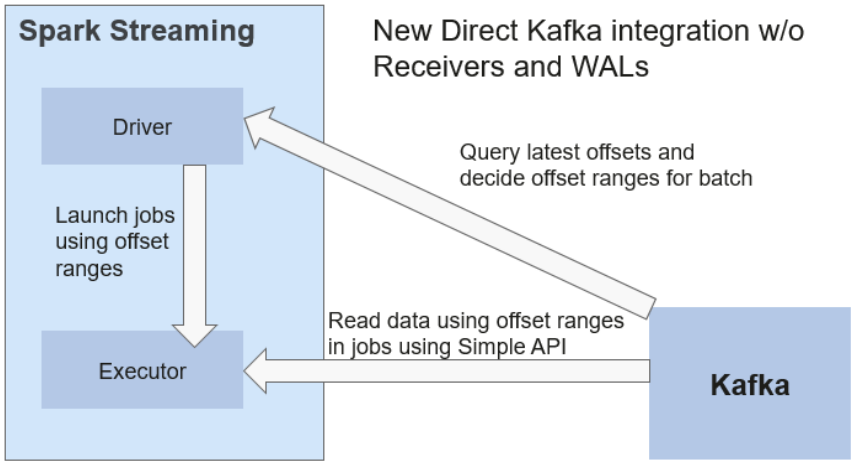
Spark Streaming 原理

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

Direct Streaming计算流程

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

图 1-106 Direct Kafka 接口数据传输



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

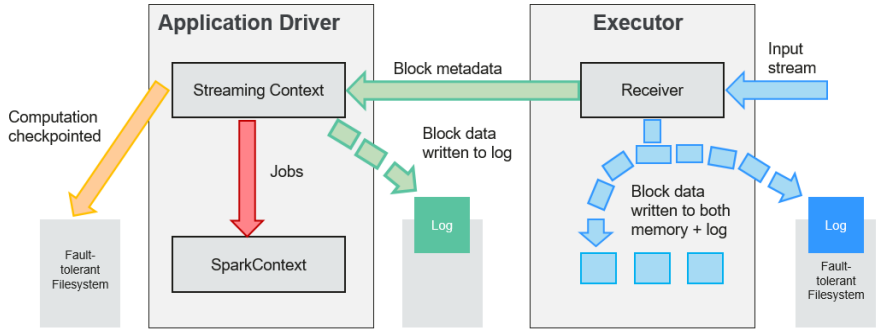
因此，Direct API消除了需要使用WAL和Receivers的情况，且确保每个Kafka记录仅被接收一次，这种接收更加高效。使得Spark Streaming和Kafka可以很好地整合在一起。总体来说，这些特性使得流处理管道拥有高容错性、高效性及易用性，因此推荐使用Direct Streaming方式处理数据。

Receiver计算流程

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

行任务。这些Receiver接收并保存流数据到Spark内存中以供处理。用户传送数据的生命周期如图1-107所示：

图 1-107 数据传输生命周期



1. 接收数据（蓝色箭头）

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

2. 通知Driver（绿色箭头）

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

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

3. 处理数据（红色箭头）

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

4. 周期性地设置检查点（橙色箭头）

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

容错性

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

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

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

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

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

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

如果需要启用预写日志功能，可以通过如下动作实现：

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

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

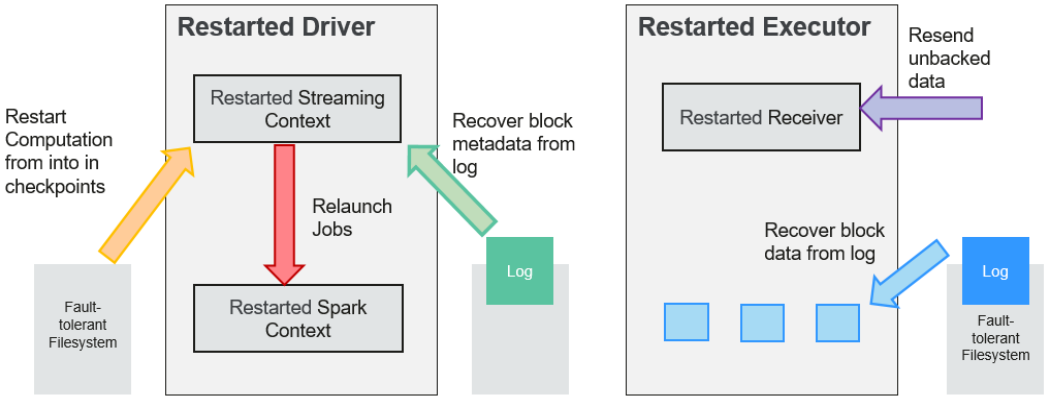
说明

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

恢复流程

当一个失败的Driver重启时，按如下流程启动：

图 1-108 计算恢复流程



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

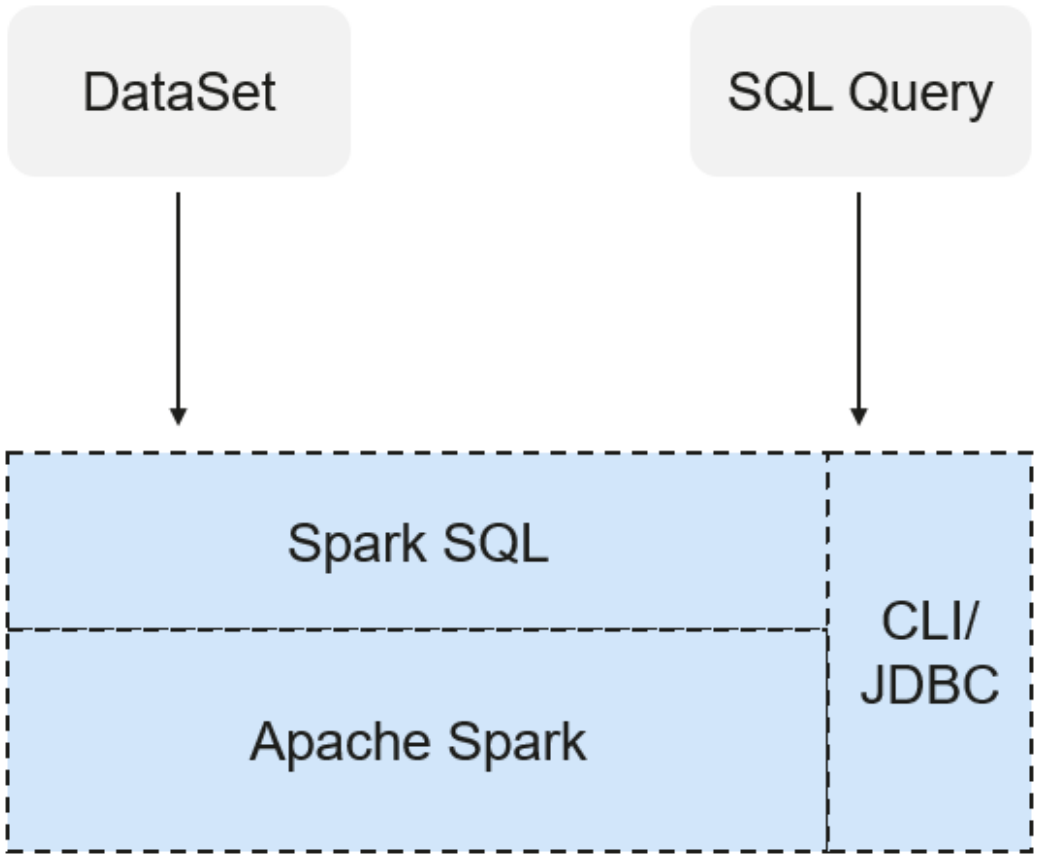
3. 未完成作业的重新形成（红色箭头）
   由于失败而没有处理完成的批处理，将使用恢复的元数据再次产生RDD和对应的作业。
4. 读取保存在日志中的块数据（蓝色箭头）
   在这些作业执行时，块数据直接从预写日志中读出。这将恢复在日志中可靠地保存的所有必要数据。
5. 重发尚未确认的数据（紫色箭头）
   失败时没有保存到日志中的缓存数据将由数据源再次发送。因为Receiver尚未对其确认。

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

SparkSQL 和 DataSet 原理

SparkSQL

图 1-109 SparkSQL 和 DataSet



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

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

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

Spark SQL Native DDL/DML

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

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

DataSet

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

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

可用于DataSet的操作分为Transformation和Action。

- Transformation操作可生成新的DataSet。

  如map、filter、select和aggregate (groupBy)。
- Action操作可触发计算及返回记结果。

  如count、show或向文件系统写数据。

通常使用两种方法创建一个DataSet：

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

CLI和JDBCServer

除了API编程接口之外，Spark SQL还对外提供CLI/JDBC接口：

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

SparkSession 原理

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

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

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

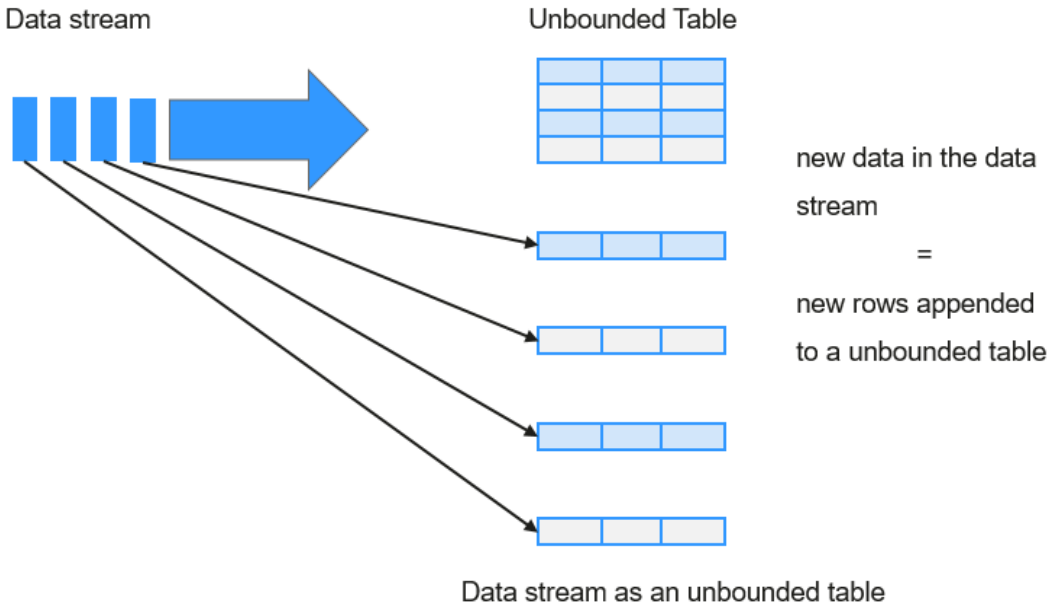
- SparkSession可以用于对数据执行SQL查询，将结果返回为DataFrame。
  ```
  sparkSession.sql("select * from person").show
  ```
- SparkSession可以用于设置运行时的配置项，这些配置项可以在SQL中使用变量替换。
  ```
  sparkSession.conf.set("spark.some.config", "abcd")
  sparkSession.conf.get("spark.some.config")
  sparkSession.sql("select ${spark.some.config}")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

Structured Streaming 原理

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

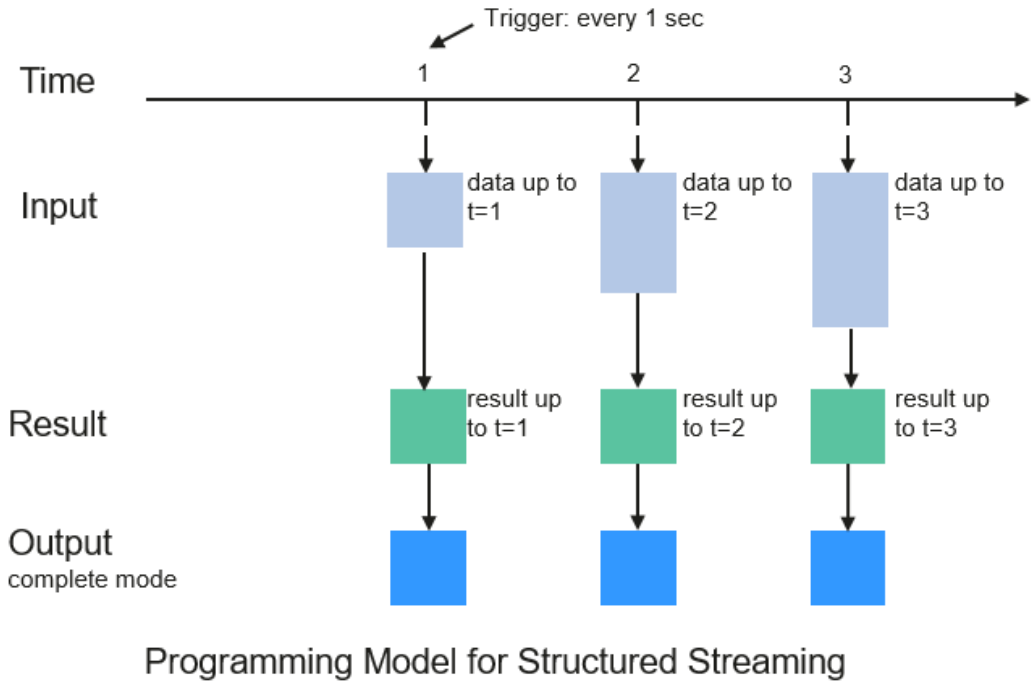
Structured Streaming的核心是将流式的数据看成一张不断增加的数据库表，这种流式的数据处理模型类似于数据块处理模型，可以把静态数据库表的一些查询操作应用在流式计算中，Spark执行标准的SQL查询，从不断增加的无边界表中获取数据。

图 1-110 Structured Streaming 无边界表



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

图 1-111 Structured Streaming 数据处理模型



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

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

基本概念

- **RDD**

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

  RDD的生成：

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

  RDD的存储：

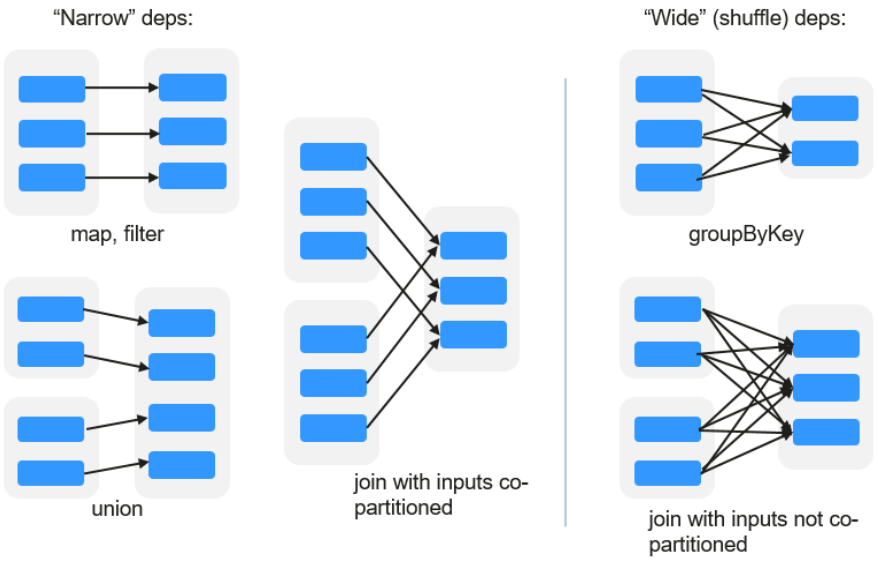
  - 用户可以选择不同的存储级别缓存RDD以便重用（RDD有11种存储级别）。

- 当前RDD默认是存储于内存，但当内存不足时，RDD会溢出到磁盘中。

- **Dependency（RDD的依赖）**

  RDD的依赖分别为：窄依赖和宽依赖。

  图 1-112 RDD 的依赖

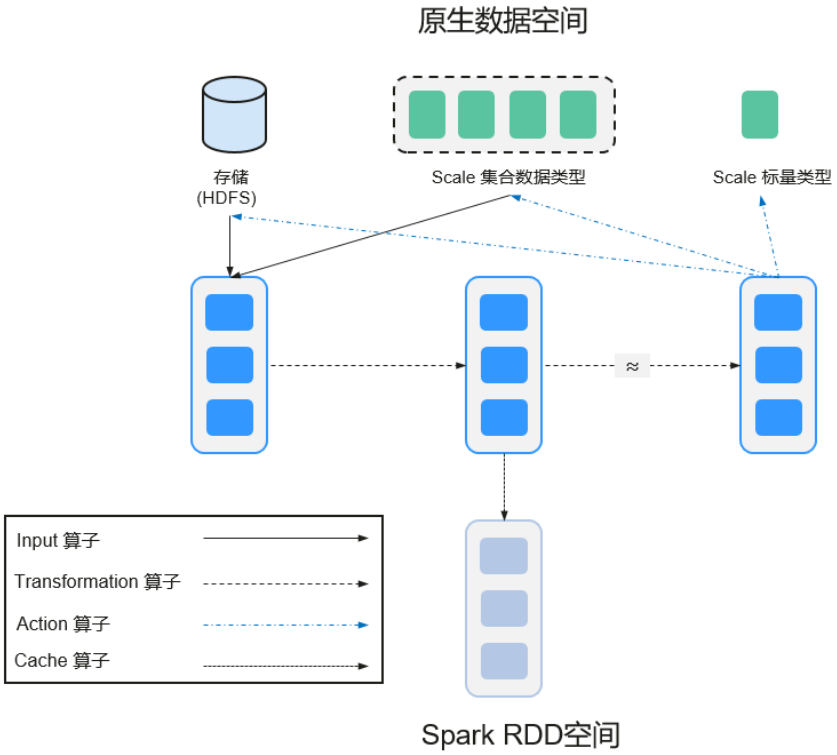
  - **窄依赖：**指父RDD的每一个分区最多被一个子RDD的分区所用。
  - **宽依赖：**指子RDD的分区依赖于父RDD的所有分区。

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

- **Transformation和Action（RDD的操作）**

  对RDD的操作包含Transformation（返回值还是一个RDD）和Action（返回值不是一个RDD）两种。RDD的操作流程如图1-113所示。其中Transformation操作是Lazy的，也就是说从一个RDD转换生成另一个RDD的操作不是马上执行，Spark在遇到Transformations操作时只会记录需要这样的操作，并不会去执行，需要等到有Actions操作的时候才会真正启动计算过程进行计算。Actions操作会返回结果或把RDD数据写到存储系统中。Actions是触发Spark启动计算的动因。

图 1-113 RDD 操作示例



RDD看起来与Scala集合类型没有太大差别，但数据和运行模型大相迥异。

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

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

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

c. cache算子缓存下来以备未来使用。

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

Transformation操作可以分为如下几种类型：

- 视RDD的元素为简单元素。

  输入输出一对一，且结果RDD的分区结构不变，主要是map。

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

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

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

- 视RDD的元素为Key-Value对。

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

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

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

对两个RDD基于key进行join和重组，如join、cogroup。

说明

后三种操作都涉及重排，称为shuffle类操作。

Action操作可以分为如下几种：

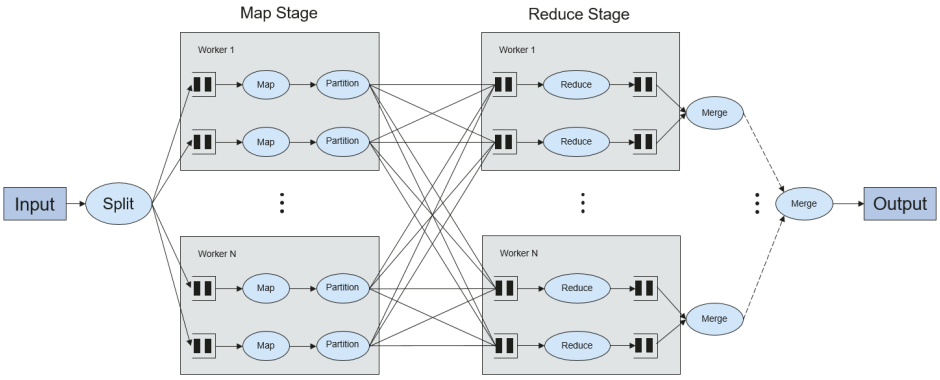
- 生成标量，如count（返回RDD中元素的个数）、reduce、fold/aggregate（返回几个标量）、take（返回前几个元素）。
- 生成Scala集合类型，如collect（把RDD中的所有元素倒入Scala集合类型）、lookup（查找对应key的所有值）。
- 写入存储，如与前文textFile对应的saveAsTextFile。
- 还有一个检查点算子checkpoint。当Lineage特别长时（这在图计算中时常发生），出错时重新执行整个序列要很长时间，可以主动调用checkpoint把当前数据写入稳定存储，作为检查点。

- **Shuffle**

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

  下图清晰地描述了MapReduce算法的整个流程。

  图 1-114 算法流程



  概念上shuffle就是一个沟通数据连接的桥梁，实际上shuffle这一部分是如何实现的呢，下面就以Spark为例讲一下shuffle在Spark中的实现。

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

- **Spark Application的结构**

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

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

    构建SparkContext对象，如：

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

    参数介绍：

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

appName：构建的Application名称。

SparkHome：集群中安装Spark的目录。

jars：应用程序代码和依赖包。

– 主体程序：处理数据

提交Application的描述请参见：**https://archive.apache.org/dist/spark/docs/3.1.1/submitting-applications.html**

- **Spark shell命令**

  Spark基本shell命令，支持提交Spark应用。命令为：

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

  参数解释：

  --class：Spark应用的类名。

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

  application-jar：Spark应用的jar包的路径。

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

- **Spark JobHistory Server**

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

1.4.25.2 Spark2x HA 方案介绍

1.4.25.2.1 Spark2x 多主实例

背景介绍

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

多主实例模式相比主备模式的HA方案，优势主要体现在对以下两种场景的改进。

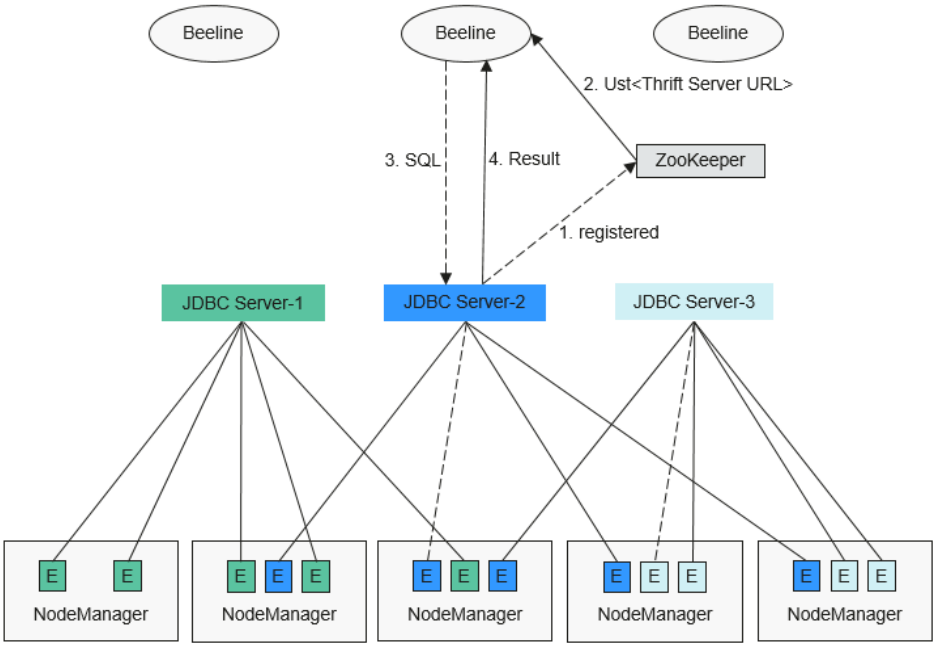
- 主备模式下，当发生主备切换时，会存在一段时间内服务不可用，该时间JDBCServer无法控制，取决于Yarn服务的资源情况。
- Spark中通过类似于HiveServer2的Thrift JDBC提供服务，用户通过Beeline以及JDBC接口访问。因此JDBCServer集群的处理能力取决于主Server的单点能力，可扩展性不够。

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

实现方案

多主实例模式的HA方案原理如下图所示。

图 1-115 Spark JDBCServer HA



1. JDBCServer在启动时，向ZooKeeper注册自身消息，在指定目录中写入节点，节点包含了该实例对应的IP，端口，版本号和序列号等信息（多节点信息之间以逗号隔开）。

   示例如下：

   ```
   [serverUri=192.168.169.84:22550
   ;version=8.1.0.1;sequence=0000001244,serverUri=192.168.195.232:22550 ;version=8.1.0.1;sequence=000
   0001242,serverUri=192.168.81.37:22550 ;version=8.1.0.1;sequence=0000001243]
   ```

2. 客户端连接JDBCServer时，需要指定Namespace，即访问ZooKeeper哪个目录下的JDBCServer实例。在连接的时候，会从Namespace下随机选择一个实例连接，详细URL参见**URL连接介绍**。
3. 客户端成功连接JDBCServer服务后，向JDBCServer服务发送SQL语句。
4. JDBCServer服务执行客户端发送的SQL语句后，将结果返回给客户端。

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

多主实例方案遵循以下规则：

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

URL 连接介绍

多主实例模式

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

- 安全模式下：
  - Kinit认证方式下的JDBCURL如下所示：
    jdbc:hive2://*<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=**sparkthriftserver2x**;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;

    说明

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

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

    示例：安全模式下通过Beeline客户端连接时执行以下命令：

    sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>***/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**
  - Keytab认证方式下的JDBCURL如下所示：
    jdbc:hive2://<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;user.principal=<principal_name>;user.keytab=<path_to_keytab>

    其中<principal_name>表示用户使用的Kerberos用户的principal，如“test@*<系统域名>*”。<path_to_keytab>表示<principal_name>对应的keytab文件路径，如“/opt/auth/test/user.keytab”。
- 普通模式下：
  jdbc:hive2://*<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=**sparkthriftserver2x**;

  示例：普通模式下通过Beeline客户端连接时执行以下命令：

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

非多主实例模式

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

示例：安全模式下通过Beeline客户端连接非多主实例模式时执行以下命令：

sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<server_IP>:<server_Port>***/;user.principal=spark2x/hadoop.<系统域名>@<系统域名>;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**

说明

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

多主实例模式与非多主实例模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。由于Spark JDBCServer是Hive中的HiveServer2的另外一个实现，其使用方法，请参见Hive官网：**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients**。

1.4.25.2.2 Spark2x 多租户

背景介绍

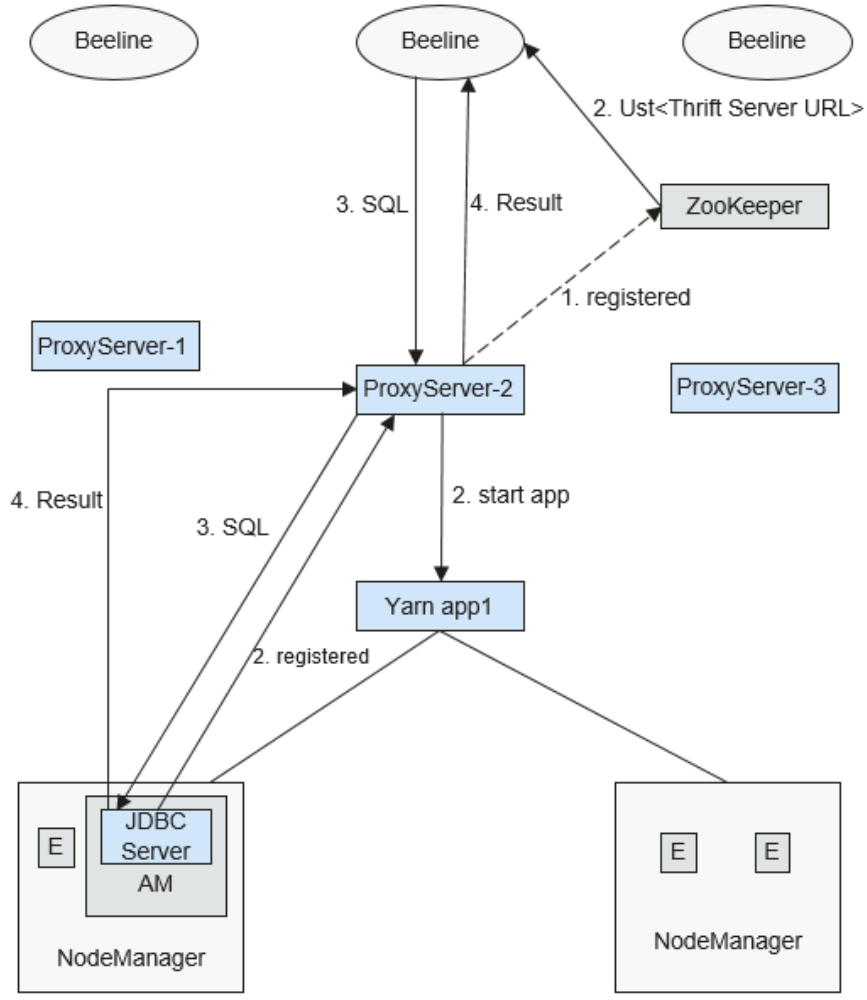
JDBCServer多主实例方案中，JDBCServer实现使用YARN-Client模式，但YARN资源队列只有一个，为了解决这种资源局限的问题，引入了多租户模式。

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

实现方案

多租户模式的HA方案原理如**图1-116**所示。

图 1-116 Spark JDBCServer 多租户



1. ProxyServer在启动时，向ZooKeeper注册自身消息，在指定目录中写入节点信息，节点信息包含了该实例对应的IP，端口，版本号和序列号等信息（多节点信息之间以逗号隔开）。

   说明

   多租户模式下，MRS页面上的JDBCServer实例是指ProxyServer（JDBCServer代理）。

   示例如下：

   ```
   serverUri=192.168.169.84:22550
   ;version=8.1.0.1;sequence=0000001244,serverUri=192.168.195.232:22550
   ;version=8.1.0.1;sequence=0000001242,serverUri=192.168.81.37:22550
   ;version=8.1.0.1;sequence=0000001243,
   ```

2. 客户端连接ProxyServer时，需要指定Namespace，即访问ZooKeeper哪个目录下的ProxyServer实例。在连接的时候，会从Namespace下随机选择一个实例连接，详细URL参见**URL连接介绍**。
3. 客户端成功连接ProxyServer服务，ProxyServer服务首先确认是否有该租户的JDBCServer存在，如果有，直接将Beeline连上真正的JDBCServer；如果没有，则以YARN-Cluster模式启动一个新的JDBCServer。JDBCServer启动成功后，ProxyServer会获取JDBCServer的地址，并将Beeline连上JDBCServer。

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

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

URL 连接介绍

多租户模式

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

- 安全模式下：
  - Kinit认证方式下的客户端URL如下所示：
    jdbc:hive2://*<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=**sparkthriftserver2x**;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;

    说明

    - 其中“<zkNode_IP>:<zkNode_Port>”是ZooKeeper的URL，多个URL以逗号隔开。
      例如：“192.168.81.37:2181,192.168.195.232:2181,192.168.169.84:2181”。
    - 其中sparkthriftserver2x是ZooKeeper上的目录，表示客户端从该目录下随机选择JDBCServer实例进行连接。

    示例：安全模式下通过Beeline客户端连接时执行以下命令：

    sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>***/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**
  - Keytab认证方式下的URL如下所示：
    jdbc:hive2://<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.*<系统域名>*@*<系统域名>*;user.principal=<principal_name>;user.keytab=<path_to_keytab>

    其中<principal_name>表示用户使用的Kerberos用户的principal，如“test@*<系统域名>*”。<path_to_keytab>表示<principal_name>对应的keytab文件路径，如“/opt/auth/test/user.keytab”。
- 普通模式下：
  jdbc:hive2://*<zkNode1_IP>:<zkNode1_Port>,<zkNode2_IP>:<zkNode2_Port>,<zkNode3_IP>:<zkNode3_Port>*/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=**sparkthriftserver2x**;

  示例：普通模式下通过Beeline客户端连接时执行以下命令：

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

非多租户模式

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

示例：安全模式下通过Beeline客户端连接非多租户模式时执行以下命令：

sh *CLIENT_HOME*/**spark/bin/beeline -u "jdbc:hive2://***<server_IP>:<server_Port>***/;user.principal=spark2x/hadoop.<系统域名>@<系统域名>;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;"**

说明

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

多租户模式与非多租户模式两种模式的JDBCServer接口相比，除连接方式不同外其他使用方法相同。由于Spark JDBCServer是Hive中的HiveServer2的另外一个实现，其使用方法，请参见Hive官网：**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients**。

指定租户

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

通过Beeline连接的命令示例如下（aaa为租户名称）：

beeline --hiveconf mapreduce.job.queuename=aaa -u 'jdbc:hive2://192.168.39.30:2181,192.168.40.210:2181,192.168.215.97:2181;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=sparkthriftserver2x;saslQop=auth-conf;auth=KERBEROS;principal=spark2x/hadoop.<系统域名>@<系统域名>;'

1.4.25.3 Spark2x 与组件的关系

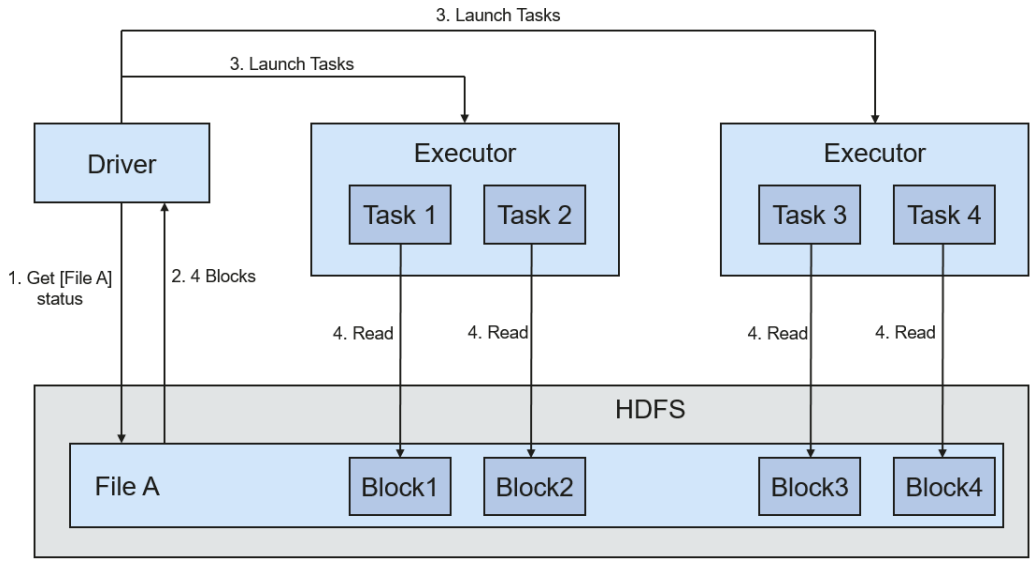
Spark 和 HDFS 的配合关系

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

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

读取文件的过程如**图1-117**所示。

图 1-117 读取文件过程

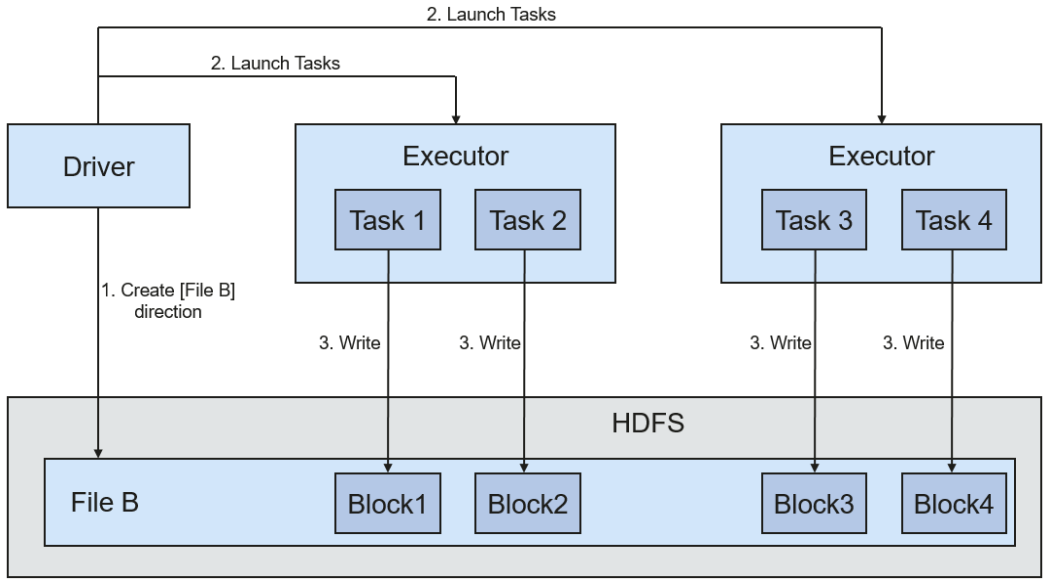


读取文件步骤的详细描述如下所示：

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

写入文件的过程如图1-118所示。

图 1-118 写入文件过程



HDFS文件写入的详细步骤如下所示：

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

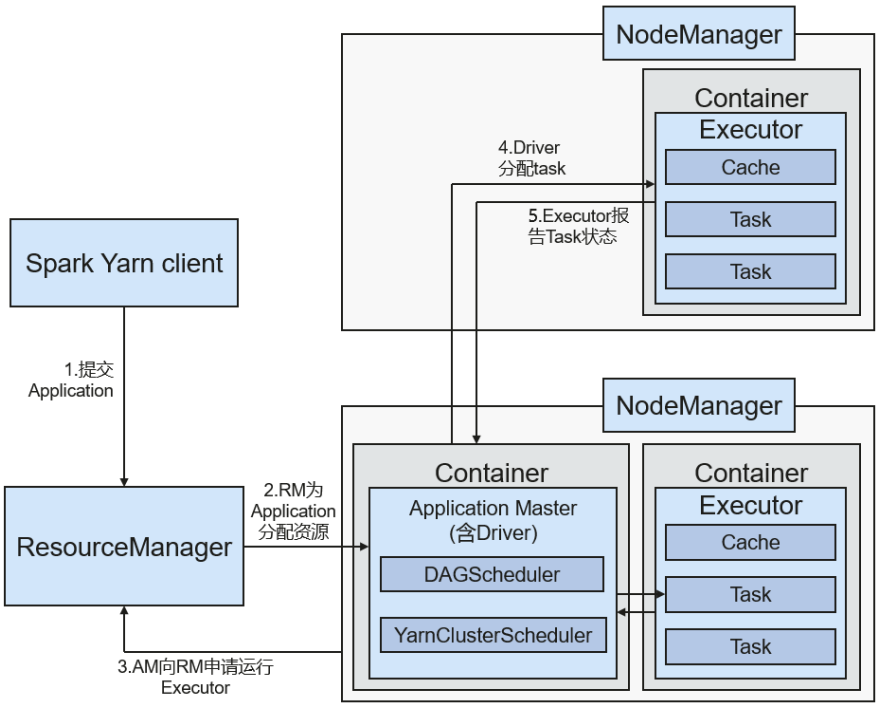
Spark 和 YARN 的配合关系

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

- YARN Cluster模式

  运行框架如图1-119所示。

  图 1-119 Spark on yarn-cluster 运行框架

  Spark on YARN-Cluster实现流程：

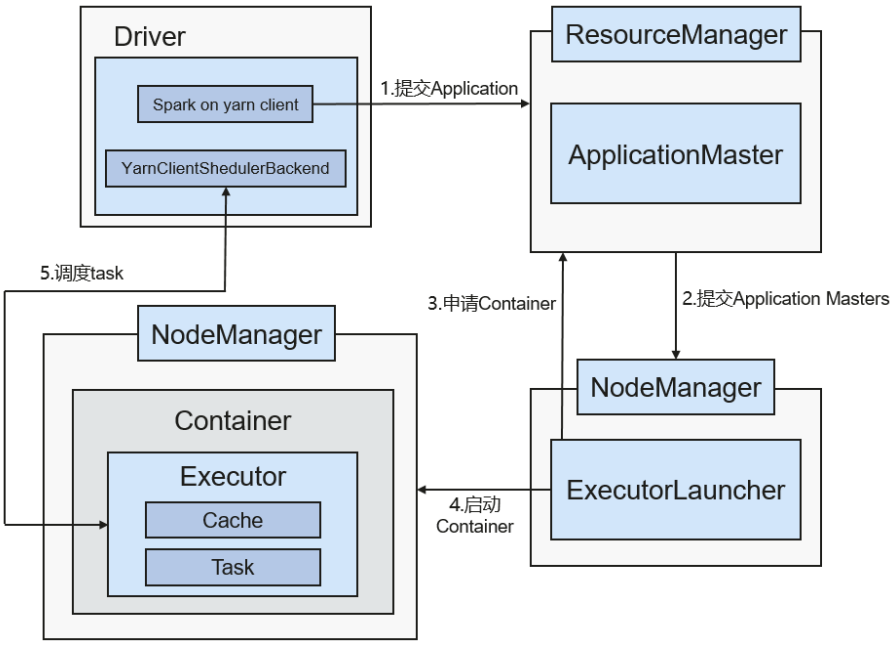
  a. 首先由客户端生成Application信息，提交给ResourceManager。
  b. ResourceManager为Spark Application分配第一个Container(ApplicationMaster)，并在该Container上启动Driver。
  c. ApplicationMaster向ResourceManager申请资源以运行Container。

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

- YARN Client模式

  运行框架如图1-120所示。

图 1-120 Spark on yarn-client 运行框架



Spark on YARN-Client实现流程：

说明

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

a. 客户端向ResourceManager发送Spark应用提交请求，Client端将启动ApplicationMaster所需的所有信息打包，提交给ResourceManager上，ResourceManager为其返回应答，该应答中包含多种信息(如ApplicationId、可用资源使用上限和下限等)。ResourceManager收到请求后，会为ApplicationMaster寻找合适的节点，并在该节点上启动它。ApplicationMaster是Yarn中的角色，在Spark中进程名字是ExecutorLauncher。

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

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

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

   说明

   正在运行的Container不会被挂起释放资源。

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

1.4.25.4 Spark2x 开源新特性

概述

Spark2x版本相对于Spark 1.5版本新增了一些开源特性。具体特性或相关概念如下：

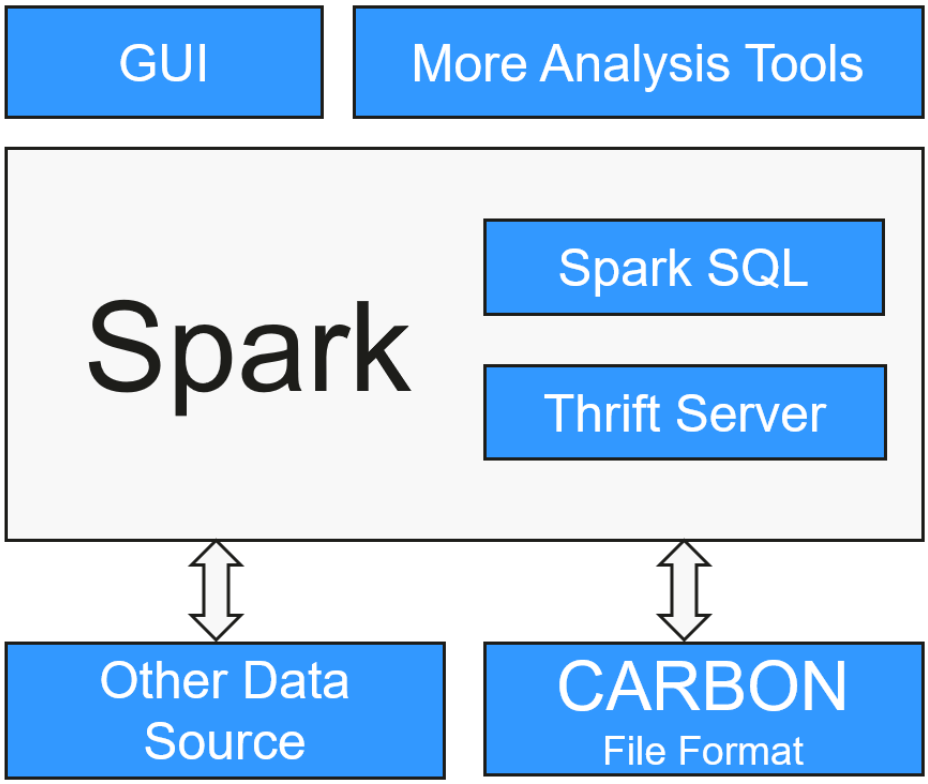
- DataSet，详见**SparkSQL和DataSet原理**。
- Spark SQL Native DDL/DML，详见**SparkSQL和DataSet原理**。
- SparkSession，详见**SparkSession原理**。
- Structured Streaming，详见**Structured Streaming原理**。
- 小文件优化。
- 聚合算法优化。
- Datasource表优化。
- 合并CBO优化。

1.4.25.5 Spark2x 开源增强特性

1.4.25.5.1 CarbonData 简介

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

图 1-121 CarbonData 基本架构



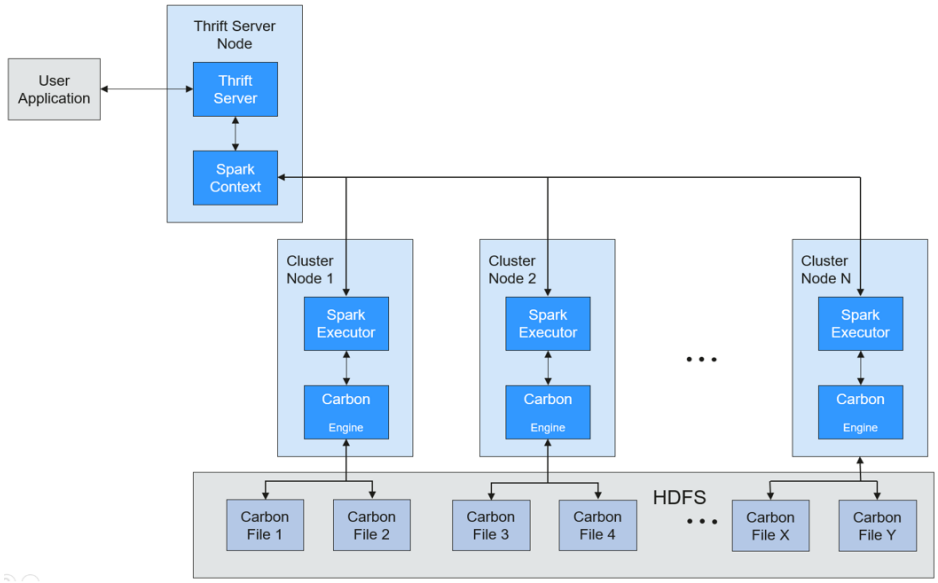
使用CarbonData的目的是对大数据即席查询提供超快速响应。从根本上说，CarbonData是一个OLAP引擎，采用类似于RDBMS中的表来存储数据。用户可将大量（10TB以上）的数据导入以CarbonData格式创建的表中，CarbonData将以压缩的多维索引列格式自动组织和存储数据。数据被加载到CarbonData后，就可以执行即席查询，CarbonData将对数据查询提供秒级响应。

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

CarbonData 结构

CarbonData作为Spark内部数据源运行，不需要额外启动集群节点中的其他进程，CarbonData Engine在Spark Executor进程之中运行。

图 1-122 CarbonData 结构



存储在CarbonData Table中的数据被分成若干个CarbonData数据文件，每一次数据查询时，CarbonData Engine模块负责执行数据集的读取、过滤等实际任务。CarbonData Engine作为Spark Executor进程的一部分运行，负责处理数据文件块的一个子集。

Table数据集数据存储在HDFS中。同一Spark集群内的节点可以作为HDFS的数据节点。

CarbonData 特性

- SQL功能：CarbonData与Spark SQL完全兼容，支持所有可以直接在Spark SQL上运行的SQL查询操作。
- 简单的Table数据集定义：CarbonData支持易于使用的DDL(数据定义语言)语句来定义和创建数据集。CarbonData DDL十分灵活、易于使用，并且足够强大，可以定义复杂类型的Table。
- 便捷的数据管理：CarbonData为数据加载和维护提供多种数据管理功能。CarbonData支持加载历史数据以及增量加载新数据。加载的数据可以基于加载时间进行删除，也可以撤销特定的数据加载操作。

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

CarbonData 关键技术和优势

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

CarbonData 索引缓存服务器

为了解决日益增长的数据量给driver带来的压力与出现的各种问题，现引入单独的索引缓存服务器，将索引从Carbon查询的Spark应用侧剥离。所有的索引内容全部由索引缓存服务器管理，Spark应用通过RPC方式获取需要的索引数据。这样，释放了大量的业务侧的内存，使得业务不会受集群规模影响而性能或者功能出现问题。

1.4.25.5.2 跨源复杂数据的 SQL 查询优化

场景描述

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

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

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

表 1-22 跨源查询增加特性对比

模块	增强前	增强后
aggregate	不支持aggregate下推	• 支持的聚合函数为：sum, avg, max, min, count 例如：select count(*) from table • 支持聚合函数内部表达式 例如：select sum(a+b) from table • 支持聚合函数运算，例如：select avg(a) + max(b) from table • 支持having下推 例如：select sum(a) from table where a>0 group by b having sum(a)>10 • 支持部分函数下推 支持对abs()、month()、length()等数学、时间、字符串函数进行下推。并且，除了以上内置函数，用户还可以通过SET命令新增数据源支持的函数。 例如：select sum(abs(a)) from table • 支持aggregate之后的limit、order by下推（由于Oracle不支持limit，所以Oracle中limit、order by不会下推） 例如：select sum(a) from table where a>0 group by b order by sum(a) limit 5
projection	仅支持简单projection下推，例如：select a, b from table	• 支持复杂表达式下推。 例如：select (a+b)*c from table • 支持部分函数下推，详细参见表下方的说明。 例如：select length(a)+abs(b) from table • 支持projection之后的limit、order by下推。 例如：select a, b+c from table order by a limit 3
predicate	仅支持运算符左边为列名右边为值的简单filter，例如 select * from table where a>0 or b in（“aaa”，“bbb”）	• 支持复杂表达数下推 例如：select * from table where a+b>c*d or a/c in (1, 2, 3) • 支持部分函数下推，详细参见表下方的说明。 例如：select * from table where length(a)>5

模块	增强前	增强后
aggregate over inner join	需要将两个表中相关的数据全部加载到Spark，先进行join操作，再进行aggregate操作	支持以下几种： ● 支持的聚合函数为：sum, avg, max, min,count ● 所有aggregate只能来自同一个表，group by可以来自一个表或者两个表，只支持inner join。 不支持的情形有： ● 不支持aggregate同时来自join左表和右表的下推。 ● 不支持aggregate内包含运算，如：sum(a+b)。 ● 不支持aggregate运算，如：sum(a)+min(b)。
aggregate over union all	需要将两个表中相关的数据全部加载到Spark，先进行union操作，再进行aggregate操作	支持情况： 支持的聚合函数为：sum, avg, max, min,count 不支持的情况： ● 不支持aggregate内包含运算，如：sum(a+b)。 ● 不支持aggregate运算，如：sum(a)+min(b)。

注意事项

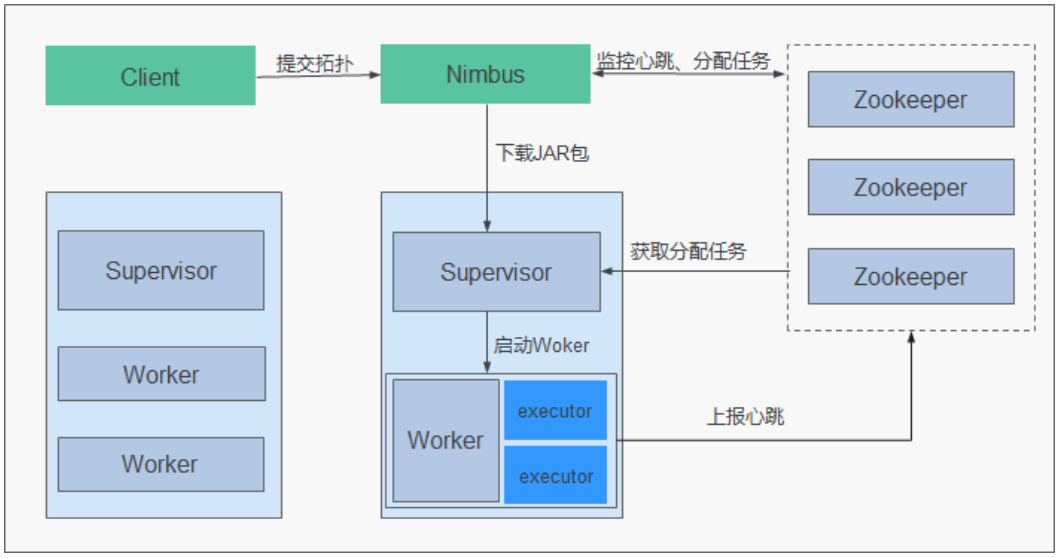
- 外部数据源是Hive的场景，通过Spark建的外表无法进行查询。
- 数据源只支持MySQL和Mppdb。

1.4.26 Storm

1.4.26.1 Storm 基本原理

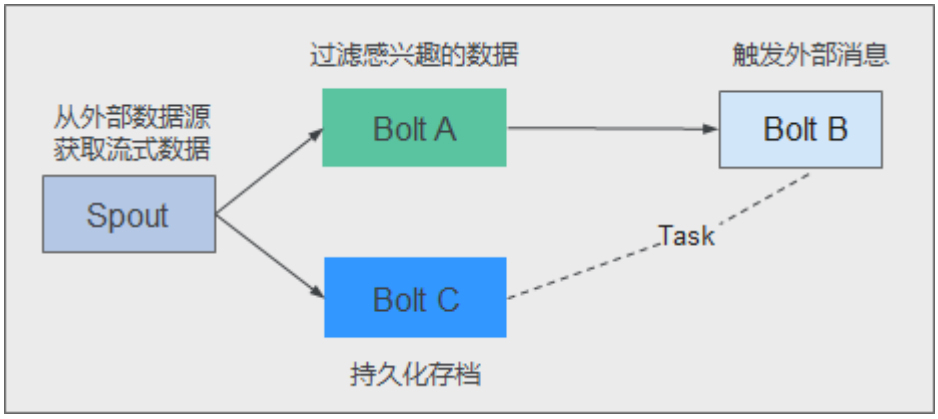
Apache Storm是一个分布式、可靠、容错的实时流式数据处理的系统。在Storm中，先要设计一个用于实时计算的图状结构，我们称之为拓扑（topology）。这个拓扑将会被提交给集群，由集群中的主控节点（master node）分发代码，将任务分配给工作节点（worker node）执行。一个拓扑中包括spout和bolt两种角色，其中spout发送消息，负责将数据流以tuple元组的形式发送出去；而bolt则负责转换这些数据流，在bolt中可以完成计算、过滤等操作，bolt自身也可以随机将数据发送给其他bolt。由spout发射出的tuple是不可变数组，对应着固定的键值对。

图 1-123 Storm 系统架构



业务处理逻辑被封装进Storm中的Topology中。一个Topology是由一组Spout组件（数据源）和Bolt组件（逻辑处理）通过Stream Groupings进行连接的有向无环图（DAG）。Topology里面的每一个Component（Spout/Bolt）节点都是并行运行的。在Topology里面， 可以指定每个节点的并行度，Storm则会在集群里面分配相应的Task来同时计算，以增强系统的处理能力。

图 1-124 Topology



Storm有众多适用场景：实时分析、持续计算、分布式ETL等。Storm有如下几个特点：

- 适用场景广泛
- 易扩展，可伸缩性高
- 保证无数据丢失
- 容错性好
- 易于构建和操控
- 多语言

Storm作为计算平台，在业务层为用户提供了更为易用的业务实现方式：CQL（Continuous Query Language—持续查询语言）。CQL具有以下几个特点：

- 使用简单：CQL语法和标准SQL语法类似，只要具备SQL基础，通过简单地学习，即可快速地进行业务开发。
- 功能丰富：CQL除了包含标准SQL的各类基本表达式等功能之外，还特别针对流处理场景增加了窗口、过滤、并发度设置等功能。
- 易于扩展：CQL提供了拓展接口，以支持日益复杂的业务场景，用户可以自定义输入、输出、序列化、反序列化等功能来满足特定的业务场景
- 易于调试：CQL提供了详细的异常码说明，降低了用户对各种错误的处理难度。

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

Storm 原理

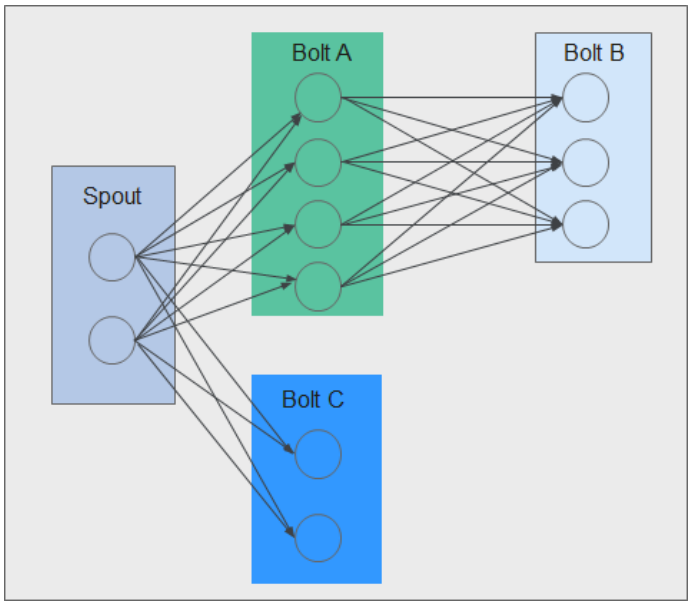
- **基本概念**

表 1-23 概念介绍

概念	说明
Tuple	Storm核心数据结构，是消息传递的基本单元，不可变Key-Value对，这些Tuple会以一种分布式的方式进行创建和处理。
Stream	Storm的关键抽象，是一个无边界的连续Tuple序列。
Topology	在Storm平台上运行的一个实时应用程序，由各个组件（Component）组成的一个DAG（Directed Acyclic Graph）。一个Topology可以并发地运行在多台机器上，每台机器上可以运行该DAG中的一部分。Topology与Hadoop中的MapReduce Job类似，不同的是，它是一个长驻程序，一旦开始就不会停止，除非人工中止。
Spout	Topology中产生源数据的组件，是Tuple的来源，通常可以从外部数据源（如消息队列、数据库、文件系统、TCP连接等）读取数据，然后转换为Topology内部的数据结构Tuple，由下一级组件处理。
Bolt	Topology中接受数据并执行具体处理逻辑（如过滤，统计、转换、合并、结果持久化等）的组件。
Worker	是Topology运行态的物理进程。每个Worker是一个JVM进程，每个Topology可以由多个Worker并行执行，每个Worker运行Topology中的一个逻辑子集。
Task	Worker中每一个Spout/Bolt的线程称为一个Task。
Stream groupings	Storm中的Tuple分发策略，即后一级Bolt以什么分发方式来接收数据。当前支持的策略有：Shuffle Grouping， Fields Grouping， All Grouping， Global Grouping， Non Grouping， Directed Grouping。

图1-125描述了一个由Spout、Bolt组成的DAG，即Topology。图中每个矩型框代表Spout或者Bolt，矩型框内的节点表示各个并发的Task，Task之间的“边”代表数据流——Stream。

图 1-125 Topology 示意图



- **可靠性**

  Storm提供三种级别的数据可靠性：

  - 至多一次：处理的数据可能会丢失，但不会被重复处理。此情况下，系统吞吐量最大。
  - 至少一次：保证数据传输可靠，但可能会被重复处理。此情况下，对在超时时间内没有获得成功处理响应的数据，会在Spout处进行重发，供后续Bolt再次处理，会对性能稍有影响。
  - 精确一次：数据成功传递，不丢失，不冗余处理。此情况下，性能最差。

  可靠性不同级别的选择，需要根据业务对可靠性的要求来选择、设计。例如对于一些对数据丢失不敏感的业务，可以在业务中不考虑数据丢失处理从而提高系统性能；而对于一些严格要求数据可靠性的业务，则需要使用精确一次的可靠性方案，以确保数据被处理且仅被处理一次。

- **容错**

  Storm是一个容错系统，提供较高可用性。**表1-24**从Storm的不同部件失效的情况角度解释其容错能力：

表 1-24 容错能力

失效场景	说明
Nimbus失效	Nimbus是无状态且快速失效的。当主Nimbus失效时，备Nimbus会接管，并对外提供服务。
Supervisor失效	Supervisor是工作节点的后台守护进程，是一种快速失效机制，且是无状态的，并不影响正在该节点上运行的Worker，但是会无法接收新的Worker分配。当Supervisor失效时，OMS会侦测到，并及时重启该进程。
Worker失效	该Worker所在节点上的Supervisor会在此节点上重新启动该Worker。如果多次重启失败，则Nimbus会将该任务重新分配到其它节点。

失效场景	说明
节点失效	该节点上的所有分配的任务会超时，而Nimbus会将这些Worker重新分配到其他节点。

Storm 开源特性

- 分布式实时计算框架

  开源Storm集群中的每台机器上都可以运行多个工作进程，每个工作进程又可创建多个线程，每个线程可以执行多个任务，任务是并发进行数据处理。

- 高容错

  如果在消息处理过程中有节点、进程等出现异常，提供重新部署该处理单元的能力。

- 可靠的消息保证

  支持At-Least Once、At-Most Once、Exactly Once的数据处理模式。

- 安全机制

  提供基于Kerberos的认证以及可插拔的授权机制，提供支持SSL的Storm UI以及Log Viewer界面，同时支持与大数据平台其他组件（如ZooKeeper，HDFS等）进行安全集成。

- 灵活的拓扑定义及部署

  使用Flux框架定义及部署业务拓扑，在业务DAG发生变化时，只需对YAML DSL（domain-specific language）定义进行修改，无需重新编译及打包业务代码。

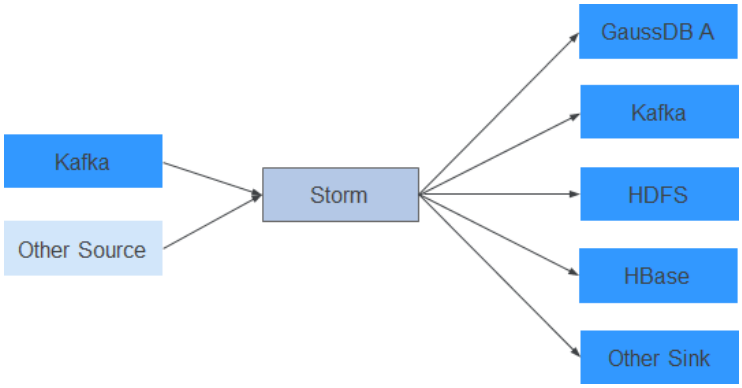
- 与外部组件集成

  支持与多种外部组件集成，包括：Kafka、HDFS、HBase、Redis或JDBC/RDBMS等服务，便于实现涉及多种数据源的业务。

1.4.26.2 Storm 与其他组件的关系

Storm，提供实时的分布式计算框架，它可以从数据源（如Kafka、TCP连接等）中获得实时消息数据，在实时平台上完成高吞吐、低延迟的实时计算，并将结果输出到消息队列或者进行持久化。Storm与其他组件的关系如图1-126所示：

图 1-126 组件关系图



Storm 和 Streaming 的关系

Storm和Streaming都使用的开源Apache Storm内核，不同的是，Storm使用的内核版本是1.2.1，Streaming使用的是0.10.0。Streaming组件一般用来在升级场景继承过度业务，比如之前版本已经部署Streaming并且有业务在运行的情况下，升级后仍然可以使用Streaming。如果是新搭建的集群，则建议使用Storm。

Storm 1.2.1新增特性说明：

- **分布式缓存**：提供命令行工具共享和更新拓扑的所需要的外部资源（配置），无需重新打包和部署拓扑。
- **Native Streaming Window API**：提供基于窗口的API。
- **资源调度器**：新增基于资源的调度器插件，可以在拓扑定义时指定可使用的最大资源，并且通过配置的方式指定用户的资源配额，从而管理该用户名下的拓扑资源。
- **State Management**：提供带检查点机制的Bolt接口，当事件失败时，Storm会自动管理bolt的状态并且执行恢复。
- **消息采样和调试：**在Storm UI界面可以开关拓扑或者组件级别的调试，将流消息按采样比率输出到指定日志中。
- **Worker动态分析：**在Storm UI界面可以收集Worker进程的Jstack、Heap日志，并且可以重启Worker进程。
- **拓扑日志级别动态调整：**提供命令行和Storm UI两种方式对运行中的拓扑日志进行动态修改。
- **性能提升：**与之前的版本相比，Storm的性能得到了显著提升。虽然，拓扑的性能和用例场景及外部服务的依赖有很大的关系，但是对于大多数场景来说，性能可以提升3倍。

1.4.26.3 Storm 开源增强特性

- CQL

  CQL（Continuous Query Language），持续查询语言，是一种用于实时数据流上的查询语言，它是一种SQL-like的语言，相对于SQL，CQL中增加了（时序）窗口的概念，将待处理的数据保存在内存中，进行快速的内存计算，CQL的输出结果为数据流在某一时刻的计算结果。使用CQL，可以快速进行业务开发，并方便地将业务提交到Storm平台开启实时数据的接收、处理及结果输出；并可以在合适的时候中止业务。

- 高可用性

  Nimbus HA机制，避免了开源Storm集群中Nimbus出现单点故障而导致集群无法提供Topology的新增及管理操作的问题，增强了集群可用性。

1.4.27 Tez

Tez是Apache最新的支持DAG（有向无环图）作业的开源计算框架，它可以将多个有依赖的作业转换为一个作业从而大幅提升DAG作业的性能。如果 Hive这样的项目使用Tez而不是MapReduce作为其数据处理的骨干，那么将会显著提升它们的响应时间，Tez构建在YARN之上，能够不需要做任何改动地运行MapReduce任务。

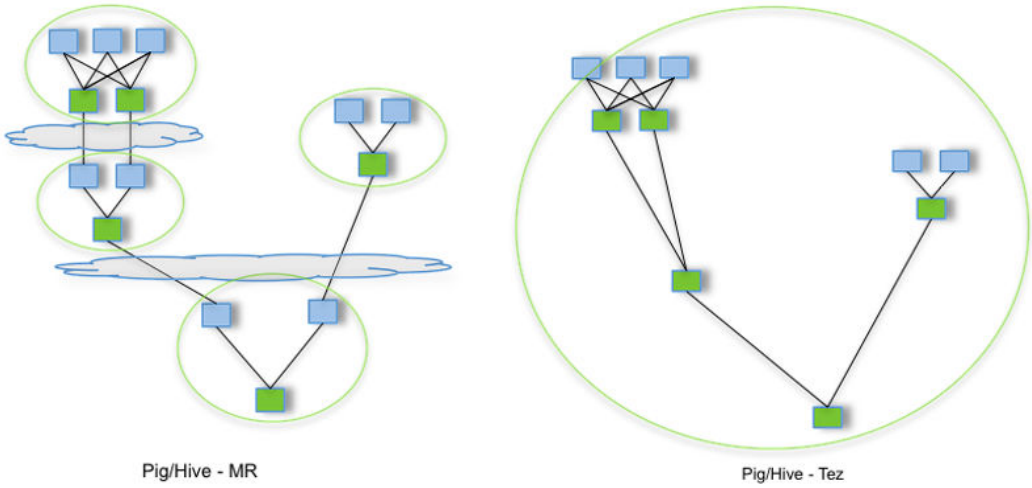
MRS将Tez作为Hive的默认执行引擎，执行效率远远超过原先的MapReduce的计算引擎。

有关Tez的详细说明，请参见：**https://tez.apache.org/**。

Tez 和 MapReduce 间的关系

Tez采用了DAG来组织MapReduce任务（DAG中一个节点就是一个RDD，边表示对RDD的操作）。它的核心思想是把将Map任务和Reduce任务进一步拆分，Map任务拆分为Input-Processor-Sort-Merge-Output，Reduce任务拆分为Input-Shuffer-Sort-Merge-Process-output，Tez将若干小任务灵活重组，形成一个大的DAG作业。

图 1-127 Hive 基于 MapReduce 提交任务和基于 Tez 提价任务流程图



Hive on MR任务中包含多个MapReduce任务，每个任务都会将中间结果存储到HDFS上——前一个步骤中的reducer为下一个步骤中的mapper提供数据。Hive on Tez任务仅在一个任务中就能完成同样的处理过程，任务之间不需要访问HDFS。

Tez 和 Yarn 间的关系

Tez是运行在Yarn之上的计算框架，运行时环境由Yarn的ResourceManager和ApplicationMaster组成。其中ResourceManager是一个全新的资源管理系统，而ApplicationMaster则负责MapReduce作业的数据切分、任务划分、资源申请和任务调度与容错等工作。此外，TezUI依赖Yarn提供的TimelineServer实现Tez任务运行过程呈现。

1.4.28 YARN

1.4.28.1 YARN 基本原理

为了实现一个Hadoop集群的集群共享、可伸缩性和可靠性，并消除早期MapReduce框架中的JobTracker性能瓶颈，开源社区引入了统一的资源管理框架**YARN**。

YARN是将JobTracker的两个主要功能（资源管理和作业调度/监控）分离，主要方法是创建一个全局的ResourceManager（RM）和若干个针对应用程序的ApplicationMaster（AM）。

说明

应用程序是指传统的MapReduce作业或作业的DAG（有向无环图）。

YARN 结构

YARN分层结构的本质是ResourceManager。这个实体控制整个集群并管理应用程序向基础计算资源的分配。ResourceManager将各个资源部分（计算、内存、带宽等）精心安排给基础NodeManager（YARN的每节点代理）。ResourceManager还与Application Master一起分配资源，与NodeManager一起启动和监视它们的基础应用程序。在此上下文中，Application Master承担了以前的TaskTracker的一些角色，ResourceManager 承担了JobTracker的角色。

Application Master管理一个在YARN内运行的应用程序的每个实例。Application Master负责协调来自ResourceManager的资源，并通过NodeManager监视容器的执行和资源使用（CPU、内存等的资源分配）。

NodeManager管理一个YARN集群中的每个节点。NodeManager提供针对集群中每个节点的服务，从监督对一个容器的终生管理到监视资源和跟踪节点健康。MRv1通过插槽管理Map和Reduce任务的执行，而NodeManager管理抽象容器，这些容器代表着可供一个特定应用程序使用的针对每个节点的资源。

图 1-128 YARN 结构

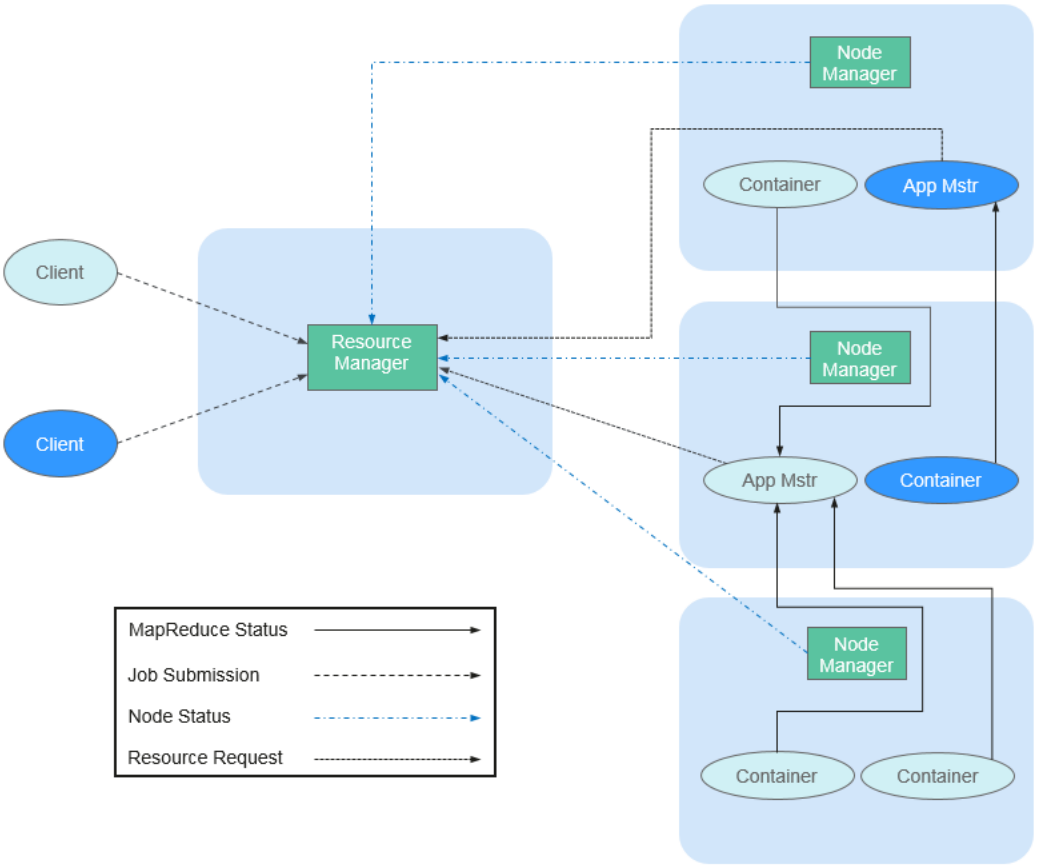


图1-128中各部分的功能如表1-25所示。

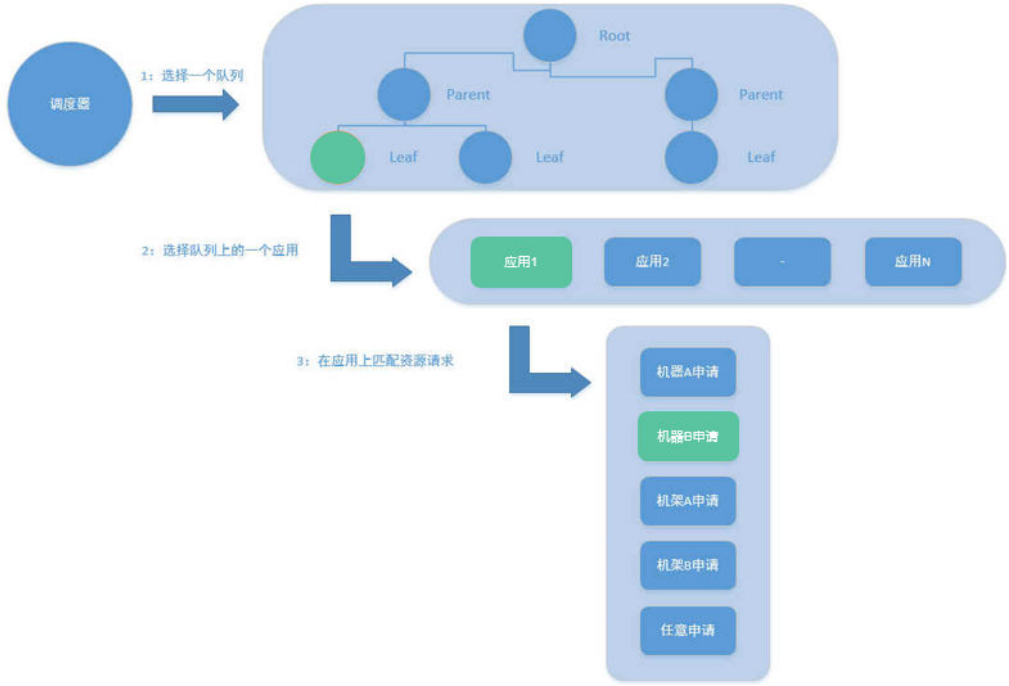
表 1-25 结构图说明

名称	描述
Client	YARN Application客户端，用户可以通过客户端向ResourceManager提交任务，查询Application运行状态等。
ResourceManager(RM)	负责集群中所有资源的统一管理和分配。接收来自各个节点（NodeManager）的资源汇报信息，并根据收集的资源按照一定的策略分配给各个应用程序。
NodeManager(NM)	NodeManager（NM）是YARN中每个节点上的代理，管理Hadoop集群中单个计算节点，包括与ResourceManger保持通信，监督Container的生命周期管理，监控每个Container的资源使用（内存、CPU等）情况，追踪节点健康状况，管理日志和不同应用程序用到的附属服务（auxiliary service）。
ApplicationMaster(AM)	即图中的App Mstr，负责一个Application生命周期内的所有工作。包括：与RM调度器协商以获取资源；将得到的资源进一步分配给内部任务（资源的二次分配）；与NM通信以启动/停止任务；监控所有任务运行状态，并在任务运行失败时重新为任务申请资源以重启任务。
Container	Container是YARN中的资源抽象，封装了某个节点上的多维度资源，如内存、CPU、磁盘、网络等（目前仅封装内存和CPU），当AM向RM申请资源时，RM为AM返回的资源便是用Container表示。YARN会为每个任务分配一个Container，且该任务只能使用该Container中描述的资源。

在YARN中，资源调度器是以层级队列方式组织资源的，这种组织方式有利于资源在不同队列间分配和共享，进而提高集群资源利用率。如下图所示，Superior Scheduler和Capacity Scheduler的核心资源分配模型相同。

调度器会维护队列的信息。用户可以向一个或者多个队列提交应用。每次NM心跳的时候，调度器会根据一定规则选择一个队列，再选择队列上的一个应用，并尝试在这个应用上分配资源。若因参数限制导致分配失败，将选择下一个应用。选择一个应用后，调度器会处理此应用的资源申请。其优先级从高到低依次为：本地资源的申请、同机架的申请，任意机器的申请。

图 1-129 资源分配模型



YARN 原理

新的Hadoop MapReduce框架被命名为MRv2或YARN。YARN主要包括ResourceManager、ApplicationMaster与NodeManager三个部分。

- ResourceManager：RM是一个全局的资源管理器，负责整个系统的资源管理和分配。主要由两个组件构成：调度器（Scheduler）和应用程序管理器（Applications Manager）。
  - 调度器根据容量、队列等限制条件（如每个队列分配一定的资源，最多执行一定数量的作业等），将系统中的资源分配给各个正在运行的应用程序。调度器仅根据各个应用程序的资源需求进行资源分配，而资源分配单位用一个抽象概念Container表示。Container是一个动态资源分配单位，将内存、CPU、磁盘、网络等资源封装在一起，从而限定每个任务使用的资源量。此外，该调度器是一个可插拔的组件，用户可根据自己的需要设计新的调度器，YARN提供了多种直接可用的调度器，比如Fair Scheduler和Capacity Scheduler等。
  - 应用程序管理器负责管理整个系统中所有应用程序，包括应用程序提交、与调度器协商资源以启动ApplicationMaster、监控ApplicationMaster运行状态并在失败时重新启动等。
- NodeManager：NM是每个节点上的资源和任务管理器，一方面，会定时向RM汇报本节点上的资源使用情况和各个Container的运行状态；另一方面，接收并处理来自AM的Container启动/停止等请求。
- ApplicationMaster：AM负责一个Application生命周期内的所有工作。包括：
  - 与RM调度器协商以获取资源。
  - 将得到的资源进一步分配给内部的任务(资源的二次分配)。
  - 与NM通信以启动/停止任务。
  - 监控所有任务运行状态，并在任务运行失败时重新为任务申请资源以重启任务。

开源容量调度器 Capacity Scheduler 原理

Capacity Scheduler是一种多用户调度器，它以队列为单位划分资源，为每个队列设定了资源最低保证和使用上限。同时，也为每个用户设定了资源使用上限以防止资源滥用。而当一个队列的资源有剩余时，可暂时将剩余资源共享给其他队列。

Capacity Scheduler支持多个队列，为每个队列配置一定的资源量，并采用FIFO调度策略。为防止同一用户的应用独占队列资源，Capacity Scheduler会对同一用户提交的作业所占资源量进行限定。调度时，首先计算每个队列使用的资源，选择使用资源最少的队列；然后按照作业优先级和提交时间顺序选择，同时考虑用户资源量的限制和内存限制。Capacity Scheduler主要有如下特性：

- 容量保证。MRS集群管理员可为每个队列设置资源最低保证和资源使用上限，而所有提交到队列的应用程序共享这些资源。
- 灵活性。如果一个队列中的资源有剩余，可以暂时共享给那些需要资源的队列，而一旦该队列有新的应用程序提交，则占用资源的队列将资源释放给该队列。这种资源灵活分配的方式可明显提高资源利用率。
- 多重租赁。支持多用户共享集群和多应用程序同时运行。为防止单个应用程序、用户或者队列独占集群中的资源，MRS集群管理员可为之增加多重约束（比如单个应用程序同时运行的任务数等）。
- 安全保证。每个队列有严格的ACL列表规定它的访问用户，每个用户可指定哪些用户允许查看自己应用程序的运行状态或者控制应用程序。此外，MRS集群管理员可指定队列管理员和集群系统管理员。
- 动态更新配置文件。MRS集群管理员可根据需要动态修改配置参数以实现在线集群管理

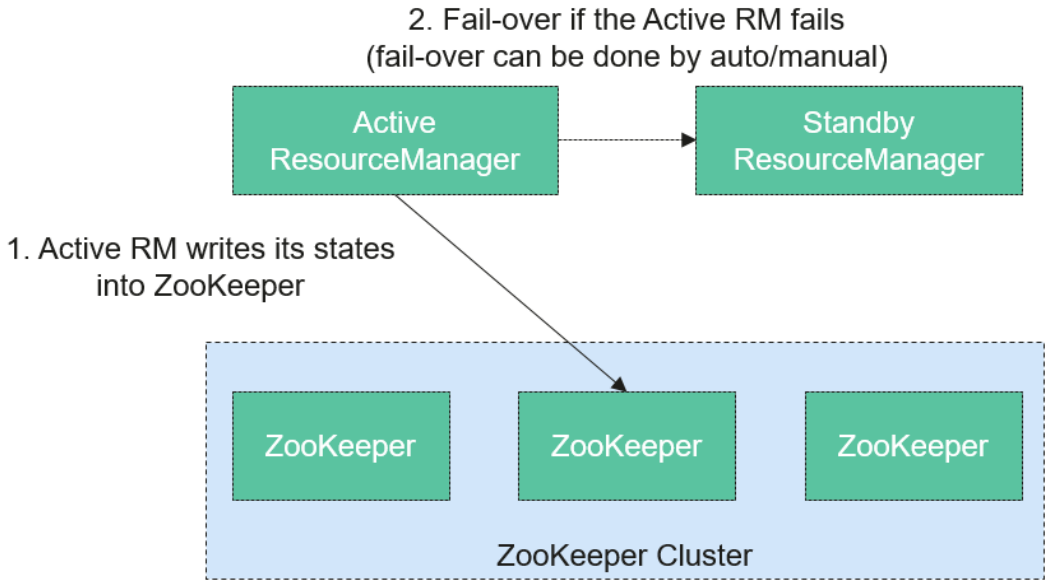
Capacity Scheduler中每个队列可以限制资源使用量。队列间的资源分配以使用量作为排列依据，使得容量小的队列有竞争优势。集群整体吞吐较大，延迟调度机制使得应用可以有机会放弃跨机器或者跨机架的调度，争取本地调度。

1.4.28.2 YARN HA 方案介绍

YARN HA 原理与实现方案

YARN中的ResourceManager负责整个集群的资源管理和任务调度，在Hadoop2.4版本之前，ResourceManager在YARN集群中存在单点故障的问题。YARN高可用性方案通过引入冗余的ResourceManager节点的方式，解决了这个基础服务的可靠性和容错性问题。

图 1-130 ResourceManager 高可用性实现架构



ResourceManager的高可用性方案是通过设置一组Active/Standby的ResourceManager节点来实现的（如图1-130）。与HDFS的高可用性方案类似，任何时间点上都只能有一个ResourceManager处于Active状态。当Active状态的ResourceManager发生故障时，可通过自动或手动的方式触发故障转移，进行Active/Standby状态切换。

在未开启自动故障转移时，YARN集群启动后，MRS集群管理员需要在命令行中使用***yarn rmadmin***命令手动将其中一个ResourceManager切换为Active状态。当需要执行计划性维护或故障发生时，则需要先手动将Active状态的ResourceManager切换为Standby状态，再将另一个ResourceManager切换为Active状态。

开启自动故障转移后，ResourceManager会通过内置的基于ZooKeeper实现的ActiveStandbyElector来决定哪一个ResourceManager应该成为Active节点。当Active状态的ResourceManager发生故障时，另一个ResourceManager将自动被选举为Active状态以接替故障节点。

当集群的ResourceManager以HA方式部署时，客户端使用的“yarn-site.xml”需要配置所有ResourceManager地址。客户端（包括ApplicationMaster和NodeManager）会以轮询的方式寻找Active状态的ResourceManager，也就是说客户端需要自己提供容错机制。如果当前Active状态的ResourceManager无法连接，那么会继续使用轮询的方式找到新的ResourceManager。

备RM升主后，能够恢复故障发生时上层应用运行的状态（详见**ResourceManger Restart**）。当启用ResourceManager Restart时，重启后的ResourceManager就可以通过加载之前Active的ResourceManager的状态信息，并通过接收所有NodeManager上container的状态信息重构运行状态继续执行。这样应用程序通过定期执行检查点操作保存当前状态信息，就可以避免工作内容的丢失。状态信息需要让Active/Standby的ResourceManager都能访问。当前系统提供了三种共享状态信息的方法：通过文件系统共享（FileSystemRMStateStore）、通过LevelDB数据库共享（LeveldbRMStateStore）或通过ZooKeeper共享（ZKRMStateStore）。这三种方式中只有ZooKeeper共享支持Fencing机制。Hadoop默认使用ZooKeeper共享。

关于YARN高可用性方案的更多信息，可参考如下链接：

http://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/ResourceManagerHA.html

1.4.28.3 Yarn 与其他组件的关系

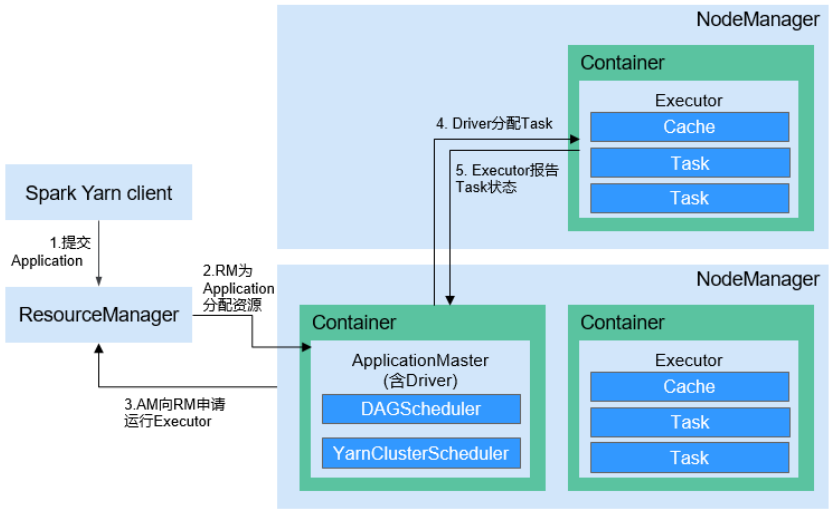
Yarn 和 Spark 组件的关系

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

- Yarn Cluster模式

  运行框架如图1-131所示。

  图 1-131 Spark on yarn-cluster 运行框架



  Spark on yarn-cluster实现流程：

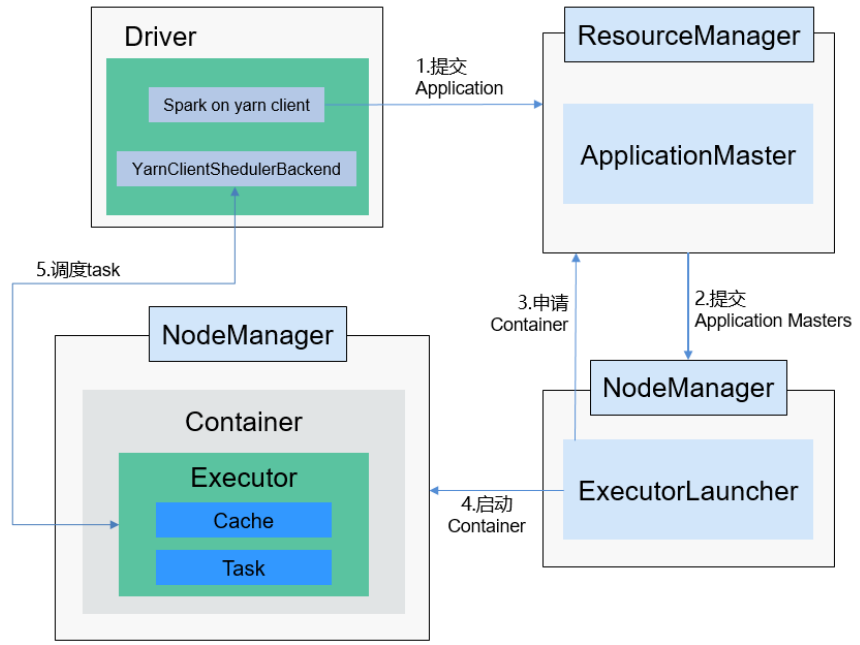
  a. 首先由客户端生成Application信息，提交给ResourceManager。
  b. ResourceManager为Spark Application分配第一个Container(ApplicationMaster)，并在该Container上启动Driver。
  c. ApplicationMaster向ResourceManager申请资源以运行Container。

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

- Yarn Client模式

  运行框架如图1-132所示。

图 1-132 Spark on yarn-client 运行框架



Spark on yarn-client实现流程：

说明

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

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

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

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

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

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

   说明

   正在运行的container不会被挂起释放资源。

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

Yarn 和 MapReduce 的关系

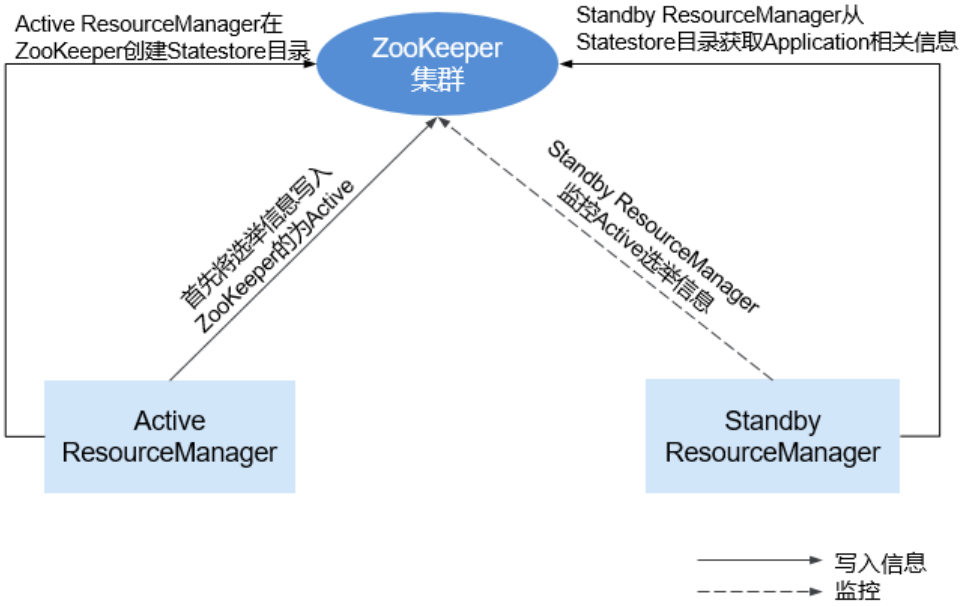
MapReduce是运行在Yarn之上的一个批处理的计算框架。MRv1是Hadoop 1.0中的MapReduce实现，它由编程模型（新旧编程接口）、运行时环境（由JobTracker和

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

Yarn 和 ZooKeeper 的关系

ZooKeeper与Yarn的关系如图1-133所示。

图 1-133 ZooKeeper 与 Yarn 的关系



1. 在系统启动时，ResourceManager会尝试把选举信息写入ZooKeeper，第一个成功写入ZooKeeper的ResourceManager被选举为Active ResourceManager，另一个为Standby ResourceManager。Standby ResourceManager定时去ZooKeeper监控Active ResourceManager选举信息。
2. Active ResourceManager还会在ZooKeeper中创建Statestore目录，存储Application相关信息。当Active ResourceManager产生故障时，Standby ResourceManager会从Statestore目录获取Application相关信息，恢复数据。

Yarn 和 Tez 的关系

Hive on Tez作业信息需要Yarn提供TimeLine Server能力，以支持Hive任务展示应用程序的当前和历史状态，便于存储和检索。

1.4.28.4 YARN 开源增强特性

任务优先级调度

在原生的YARN资源调度机制中，如果先提交的MapReduce Job长时间地占据整个Hadoop集群的资源，会使得后提交的Job一直处于等待状态，直到Running中的Job执行完并释放资源。

MRS集群提供了任务优先级调度机制。此机制允许用户定义不同优先级的Job，后启动的高优先级Job能够获取运行中的低优先级Job释放的资源；低优先级Job未启动的计算容器被挂起，直到高优先级Job完成并释放资源后，才被继续启动。

该特性使得业务能够更加灵活地控制自己的计算任务，从而达到更佳的集群资源利用率。

说明

容器可重用与任务优先级调度有冲突，若启用容器重用，资源会被持续占用，优先级调度将不起作用。

YARN 的权限控制

Hadoop YARN的权限机制是通过访问控制列表（ACL）实现的。按照不同用户授予不同权限控制，主要介绍下面两个部分：

- 集群运维管理员控制列表（Admin Acl）

  该功能主要用于指定YARN集群的运维管理员，其中，MRS集群管理员列表由参数“yarn.admin.acl”指定。集群运维管理员可以访问ResourceManager WebUI，还能操作NodeManager节点、队列、NodeLabel等，**但不能提交任务**。

- 队列访问控制列表（Queue Acl）

  为了方便管理集群中的用户，YARN将用户/用户组分成若干队列，并指定每个用户/用户组所属的队列。每个队列包含两种权限：提交应用程序权限和管理应用程序权限（比如终止任意应用程序）。

开源功能：

虽然目前YARN服务的用户层面上支持如下三种角色：

- 集群运维管理员
- 队列管理员
- 普通用户

但是当前开源YARN提供的WebUI/RestAPI/JavaAPI等接口上不会根据用户角色进行权限控制，任何用户都有权限访问应用和集群的信息，无法满足多租户场景下的隔离要求。

增强：

安全模式下，对开源YARN提供的WebUI/RestAPI/JavaAPI等接口上进行了权限管理上的增强，支持根据不同的用户角色，进行相应的权限控制。

各个角色对应的权限如下：

- 集群运维管理员：拥有在YARN集群上执行管理操作（如访问ResourceManager WebUI、刷新队列、设置NodeLabel、主备倒换等）的权限。

- 队列管理员：拥有在YARN集群上所管理队列的修改和查看权限。
- 普通用户：拥有在YARN集群上对自己提交应用的修改和查看权限。

自研超级调度器 Superior Scheduler 原理

Superior Scheduler是一个专门为Hadoop YARN分布式资源管理系统设计的调度引擎，是针对企业客户融合资源池，多租户的业务诉求而设计的高性能企业级调度器。

Superior Scheduler可实现开源调度器、Fair Scheduler以及Capacity Scheduler的所有功能。另外，相较于开源调度器，Superior Scheduler在企业级多租户调度策略、租户内多用户资源隔离和共享、调度性能、系统资源利用率和支持大集群扩展性方面都做了针对性的增强。设计的目标是让Superior Scheduler直接替代开源调度器。

类似于开源Fair Scheduler和Capacity Scheduler，Superior Scheduler通过YARN调度器插件接口与YARN Resource Manager组件进行交互，以提供资源调度功能。图1-134为其整体系统图。

图 1-134 Superior Scheduler 内部架构

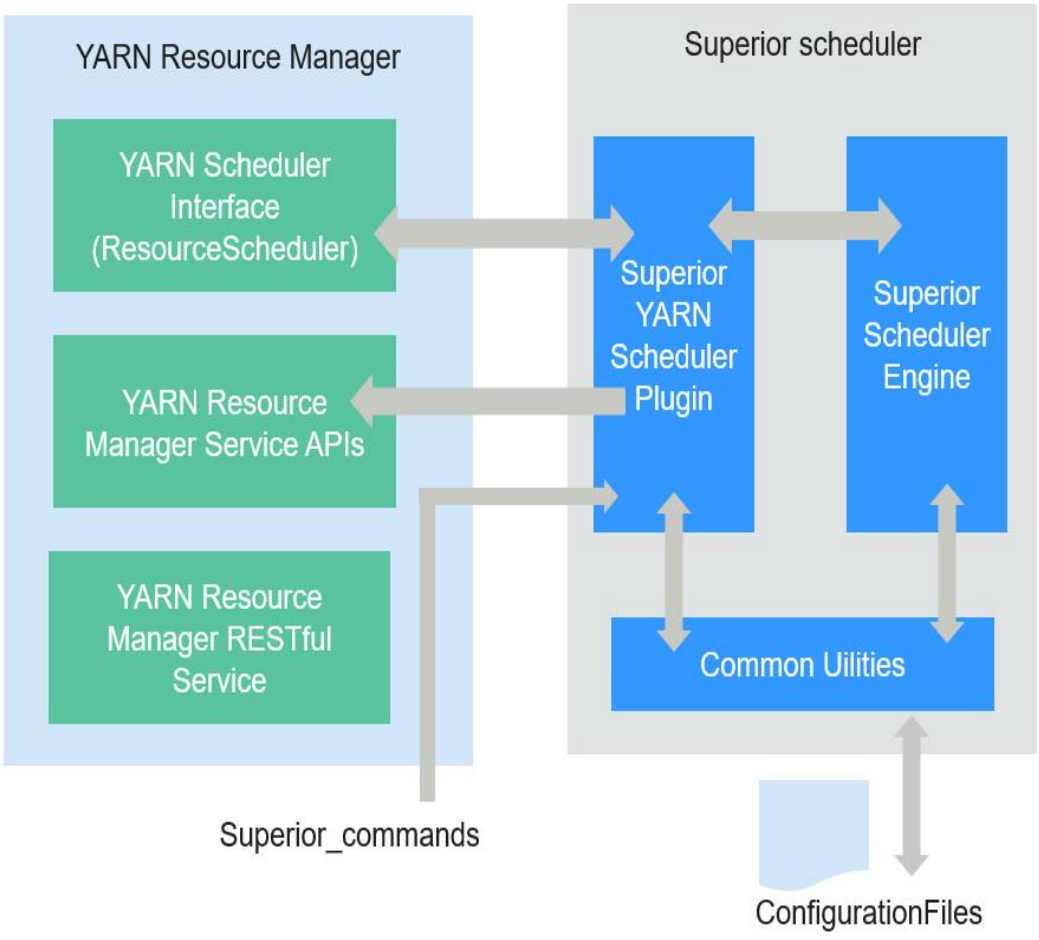


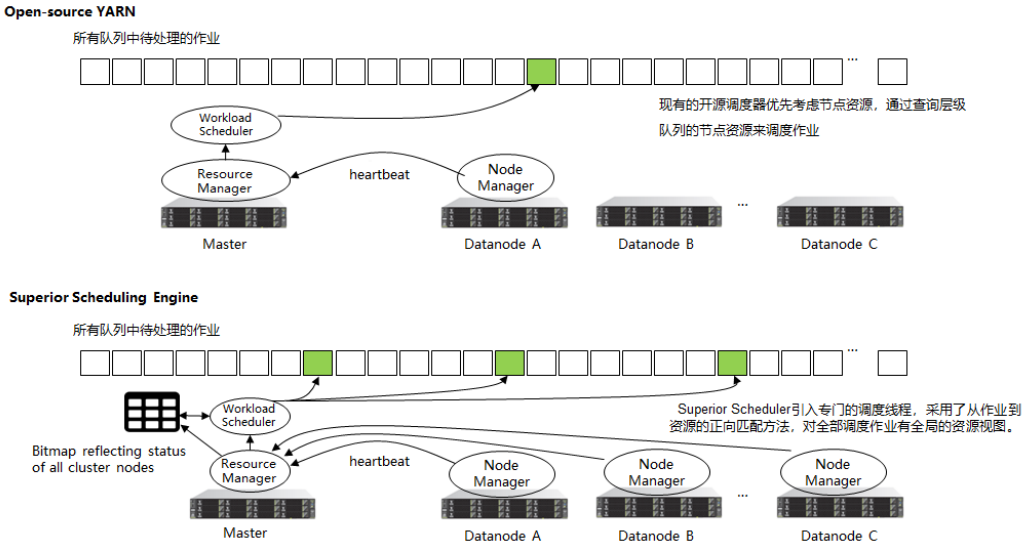
图1-134中，Superior Scheduler的主要模块如下：

- Superior Scheduler Engine：具有丰富调度策略的高性能调度器引擎。
- Superior YARN Scheduler Plugin：YARN Resource Manager和Superior Scheduler Engine之间的桥梁，负责同YARN Resource Manager交互。

在调度原理上，开源的调度器都是基于计算节点心跳驱动的资源反向匹配作业的调度机制。具体来讲，每个计算节点定期发送心跳到YARN的Resource Manager通知该节点状态并同时启动调度器为这个节点分配作业。这种调度机制把调度的周期同心跳结合在一起，当集群规模增大时，会遇到系统扩展性以及调度性能瓶颈。另外，因为采用了资源反向匹配作业的调度机制，开源调度器在调度精度上也有局限性，例如数据亲和性偏于随机，另外系统也无法支持基于负载的调度策略等。主要原因是调度器在选择作业时，缺乏全局的资源视图，很难做到最优选择。

Superior Scheduler内部采用了不同的调度机制。Superior Scheduler的调度器引入了专门的调度线程，把调度同心跳剥离开，避免了系统心跳风暴问题。另外，Superior Scheduler调度流程采用了从作业到资源的正向匹配方法，这样每个调度的作业都有全局的资源视图，可以很大的提到调度的精度。相比开源调度器，Superior Scheduler在系统吞吐量、利用率、数据亲和性等方面都有很大提升。

图 1-135 Superior Scheduler 性能对比



Superior Scheduler除了提高系统吞吐量和利用率，还提供了以下主要调度功能：

- 多资源池

  多资源池有助于在逻辑上划分集群资源并在多个租户/队列之间共享它们。资源池的划分可以基于异构的资源或完全按照应用资源隔离的诉求来划分。对于一个资源池，不同队列可配置进一步的策略。

- 每个资源池多租户调度（reserve、min、share、max）

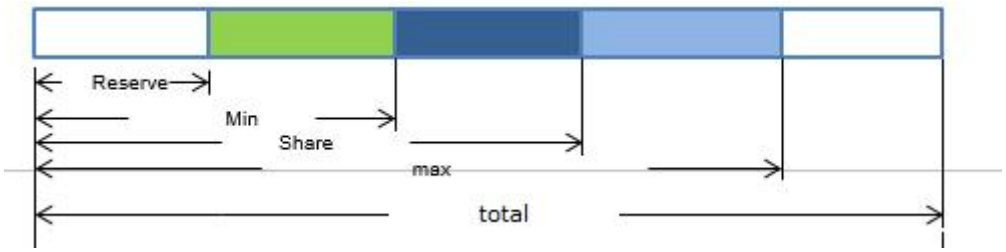
  Superior Scheduler提供了灵活的层级多租户调度策略。并允许针对不同的资源池可以访问的租户/队列，配置不同策略，如下所示。

表 1-26 策略描述

策略名称	描述
reserve	预留租户资源。即使租户没有作业，其他租户也不能使用该预留的资源。其值可以是百分比或绝对值。如果两者都配置，调度系统动态计算转换为资源绝对值，并取两者的最大值。缺省的reserve值为0。相对于定义一个专用资源池并指定具体机器的方式，reserve的策略可以认为提供了一种灵活的浮动预留功能，由于并不限定具体的机器，可以提高计算的数据亲和性，也不会受具体机器故障的影响。
min	具有抢占支持的最低保证资源。其他租户可以使用这部分资源，但是本租户享有优先使用权。其值可以是百分比或绝对值。如果两者都配置，调度系统动态计算转换为资源绝对值，并取两者的最大值。缺省值是0。
share	不支持抢占的共享资源。本租户要使用这部分资源时，需要等待其他租户完成作业并释放资源。其值是百分比或绝对值。
max	允许的最大资源数量。租户无法获得比允许的最大资源多的资源。其值是百分比或绝对值。如果两者都配置，调度系统动态计算转换为资源绝对值，并取两者最大值。缺省值不受限制。

租户资源分配策略示意图，如图1-136所示。

图 1-136 策略示意图



说明

其中“total”表示总资源，不是调度策略。

同开源的调度器相比，Superior Scheduler同时提供了租户级百分比和绝对值的混配策略，可以很好的适应各种灵活的企业级租户资源调度诉求。例如，用户可以在一级租户提供最大绝对值的资源保障，这样租户的资源不会因为集群的规模改变而受影响。但在下层的子租户之间，可以提供百分比的分配策略，这样可以尽可能提升一级租户内的资源利用率。

- 异构和多维资源调度

  Superior Scheduler支持CPU和内存资源的调度外，还支持扩展支持以下功能：

  - 节点标签可用于识别像GPU_ENABLED，SSD_ENBALED等节点的多维属性，可以根据这些标签进行调度。
  - 资源池可用于对同一类别的资源进行分组并分配给特定的租户/队列。
- 租户内多用户公平调度

在叶子租户里，多个用户可以使用相同的队列来提交作业。相比开源调度器，Superior Scheduler可以支持在同一租户内灵活配置不同用户的资源共享策略。例如可以为VIP用户配置更多的资源访问权重。

- 数据位置感知调度

  Superior Scheduler采用“从作业到节点的调度策略”，即尝试在可用节点之间调度给定的作业，使得所选节点适合于给定作业。通过这样做，调度器将具有集群和数据的整体视图。如果有机会使任务更接近数据，则保证了本地化。而开源调度器采用“从节点到作业的调度策略”，在给定节点中尝试匹配适当的作业。

- Container调度时动态资源预留

  在异构和多样化的计算环境中，一些container需要更多的资源或多种资源，例如Spark作业可能需要更大的内存。当这些container与其他需要较小资源的container竞争时，可能没有机会在合理的时间内获得所需的资源而处于饥饿状态。由于开源的调度器是基于资源反向匹配作业的调度方式，会为这些作业盲目的进行资源预留以防进入饥饿状态。这就导致了系统资源的整体浪费。Superior Scheduler与开源特性的不同之处在于：

  - 基于需求的匹配：由于Superior Scheduler采用“从作业到节点的调度”，能够选择合适的节点来预留资源提升这些特殊container的启动时间，并避免浪费。
  - 租户重新平衡：启用预留逻辑时，开源调度器并不遵循配置的共享策略。Superior Scheduler采取不同的方法。在每个调度周期中，Superior Scheduler将遍历租户，并尝试基于多租户策略重新达到平衡，且尝试满足所有策略（reserve，min，share等），以便可以释放预留的资源，将可用资源流向不同租户下的其他本应得到资源的container。

- 动态队列状态控制（Open/Closed/Active/Inactive）

  支持多个队列状态，有助于MRS集群管理员操作和维护多个租户。

  - Open状态（Open/Closed）：如果是Open（默认）状态，将接受提交到此队列的应用程序，如果是Closed状态，则不接受任何应用程序。
  - Active状态（Active/Inactive）：如果处于Active（默认）状态，租户内的应用程序是可以被调度和分配资源。如果处于Inactive状态则不会进行调度。

- 应用等待原因

  如果应用程序尚未启动，则提供作业等待原因信息。

Superior Scheduler和YARN开源调度器作了对比分析，如表1-27所示：

表 1-27 对比分析

领域	YARN开源调度器	Superior Scheduler
多租户调度	在同构集群上，只能选择容量调度器（Capacity Scheduler）或公平调度器（Fair Scheduler）两者之一，且集群当前不支持公平调度器（Fair Scheduler）。容量调度器只支持百分比方式配置，而公平调度器只支持绝对值方式。	• 支持**异构集群**和**多资源池**。 • 支持**预留**，以保证直接访问资源。

领域	YARN开源调度器	Superior Scheduler
数据位置感知调度	从节点到作业的调度策略导致降低数据本地话命中率，潜在影响应用的执行性能。	**从作业到节点的调度策略。**可具有更精确的数据位置感知，数据本地化调度的作业命中率比较高。
基于机器负载的均衡调度	不支持	**Superior Scheduler在调度时考虑机器的负载和资源分配情况，做到均衡调度。**
租户内多用户公平调度	不支持	租户内用户的公平调度，支持关键字default、others。
作业等待原因	不支持	作业等待原因信息可显示为什么作业需等待。

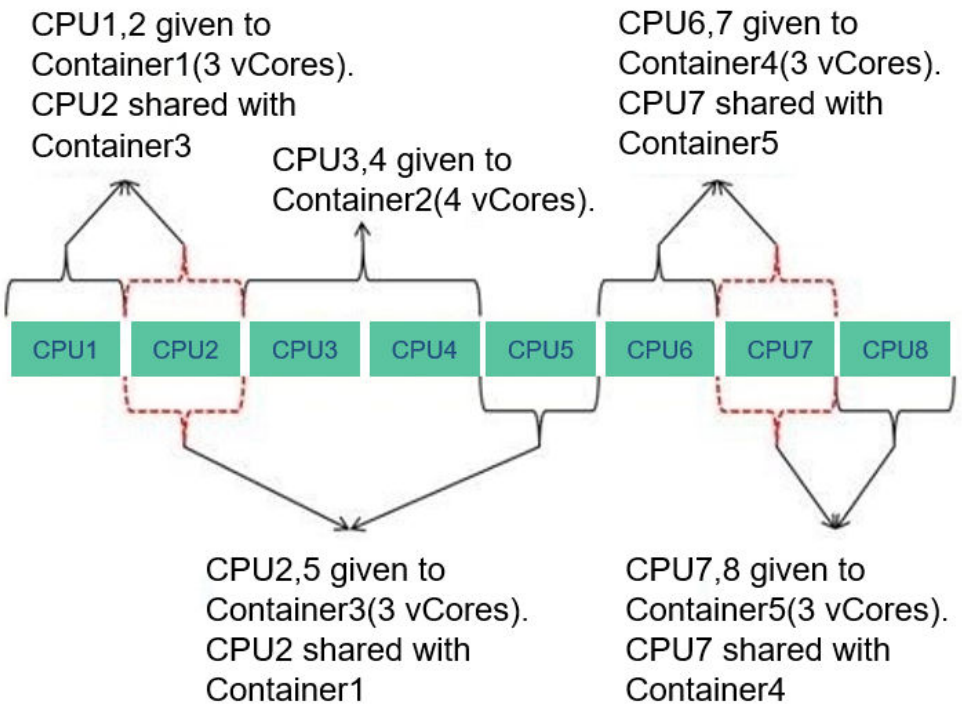
综上所述，Superior Scheduler是一个高性能调度器，拥有丰富的调度策略，在功能、性能、资源利用率和扩展性方面都优于Capacity Scheduler。

支持 CPU 硬隔离

YARN无法严格控制每个container使用的CPU资源。在使用CPU子系统时，container可能会超额占用资源。所以，我们使用CPUset控制资源分配。

为了解决这个问题，CPU将会被严格按照虚拟核和物理核的比例分配至各个container。如果container需要一整个物理核，则分配给它一整个物理核。若container只需要部分物理核，则可能发生几个container共享同一个物理核的情况。下图为CPU配额示例，假定虚拟核和物理核的比例为2:1。

图 1-137 CPU 配额

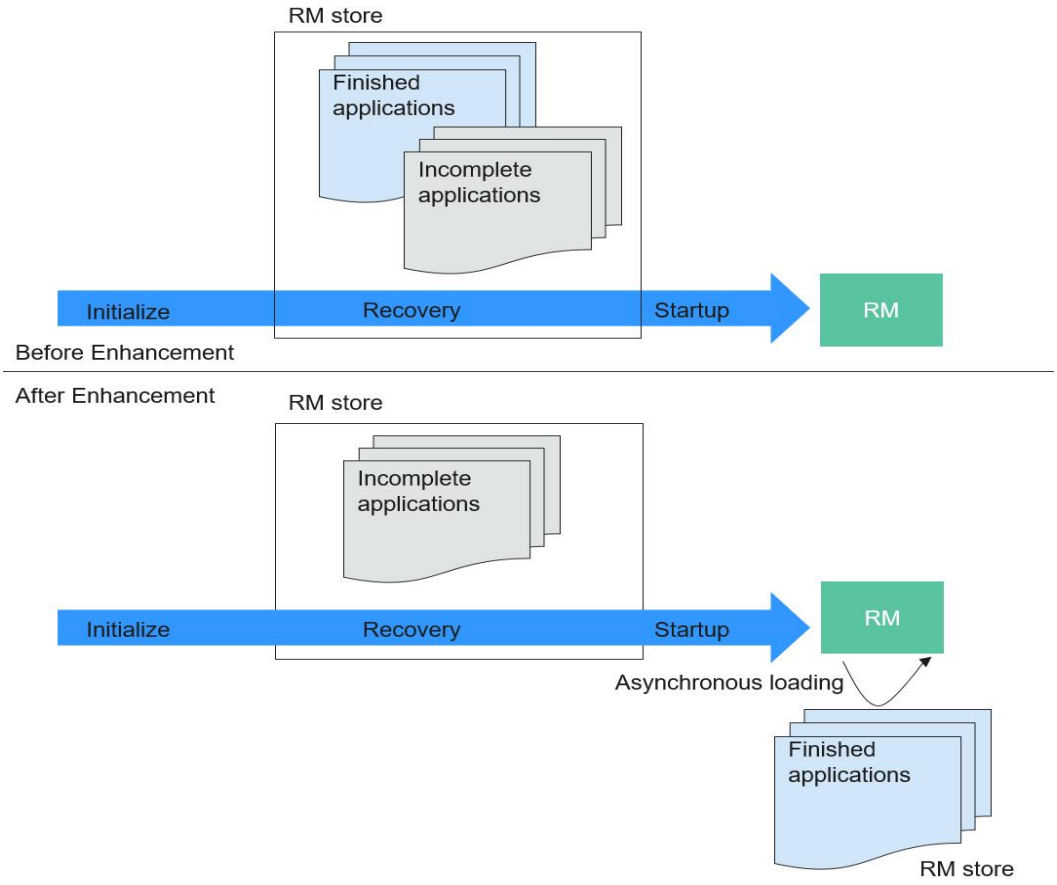


YARN 开源增强特性：重启性能优化

一般情况下，RM恢复会获取正在运行和已完成的应用。而大量的已完成的应用可能导致RM启动过慢、HA切换/重启耗时过长等问题。

为了加速RM的启动，现在优先获取未完成的应用列表，再启动RM。此时，已完成的应用会在一个后台异步线程中继续恢复。下图展示了RM的启动恢复流程。

图 1-138 RM 启动恢复流程



1.4.29 ZooKeeper

1.4.29.1 ZooKeeper 基本原理

ZooKeeper 简介

ZooKeeper是一个分布式、高可用性的协调服务。在大数据产品中主要提供两个功能：

- 帮助系统避免单点故障，建立可靠的应用程序。
- 提供分布式协作服务和维护配置信息。

ZooKeeper 结构

ZooKeeper集群中的节点分为三种角色：Leader、Follower和Observer，其结构和相互关系如图1-139所示。通常来说，需要在集群中配置奇数个（2N+1）ZooKeeper服务，至少（N+1）个投票才能成功的执行写操作。

图 1-139 ZooKeeper 结构

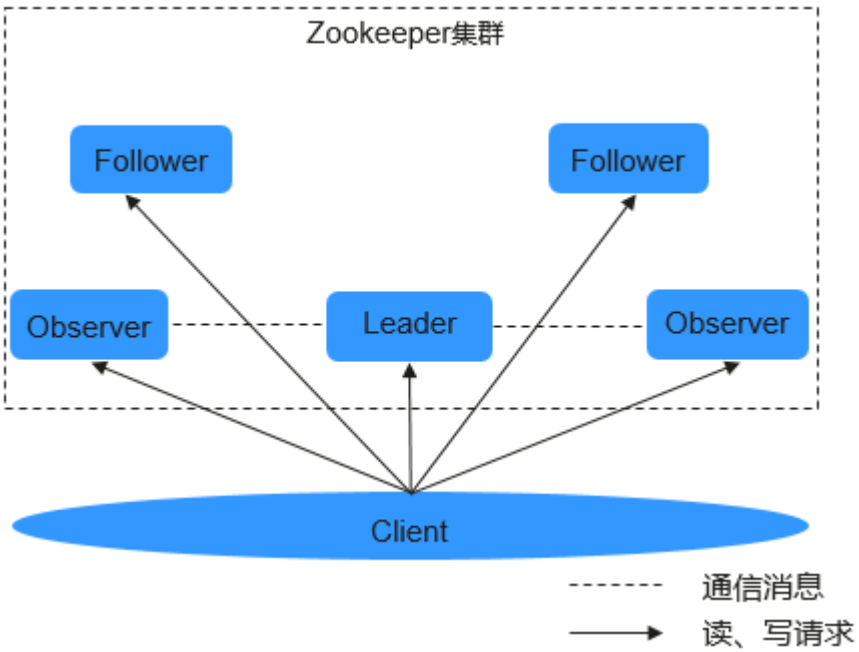


图1-139中各部分的功能说明如表1-28所示。

表 1-28 结构图说明

名称	描述
Leader	在ZooKeeper集群中只有一个节点作为集群的领导者，由各Follower通过ZooKeeper Atomic Broadcast(ZAB)协议选举产生，主要负责接收和协调所有写请求，并把写入的信息同步到Follower和Observer。
Follower	Follower的功能有两个： • 每个Follower都作为Leader的储备，当Leader故障时重新选举Leader，避免单点故障。 • 处理读请求，并配合Leader一起进行写请求处理。
Observer	Observer不参与选举和写请求的投票，只负责处理读请求、并向Leader转发写请求，避免系统处理能力浪费。
Client	ZooKeeper集群的客户端，对ZooKeeper集群进行读写操作。例如HBase可以作为ZooKeeper集群的客户端，利用ZooKeeper集群的仲裁功能，控制其HMaster的“Active”和“Standby”状态。

如果集群启用了安全服务，在连接ZooKeeper时需要进行身份认证，认证方式有以下两种：

- keytab方式：需要从MRS集群管理员处获取一个“人机”用户，用于登录MRS平台并通过认证，并且获取到该用户的keytab文件。
- 票据方式：从MRS集群管理员处获取一个“人机”用户，用于后续的安全登录，开启Kerberos服务的renewable和forwardable开关并且设置票据刷新周期，开启成功后重启kerberos及相关组件。

  说明

  - 默认情况下，用户的密码有效期是90天，所以获取的keytab文件的有效期是90天。
  - Kerberos服务的renewable、forwardable开关和票据刷新周期的设置在Kerberos服务的配置页面的“系统”标签下，票据刷新周期的修改可以根据实际情况修改“kdc_renew_lifetime”和“kdc_max_renewable_life”的值。

ZooKeeper 原理

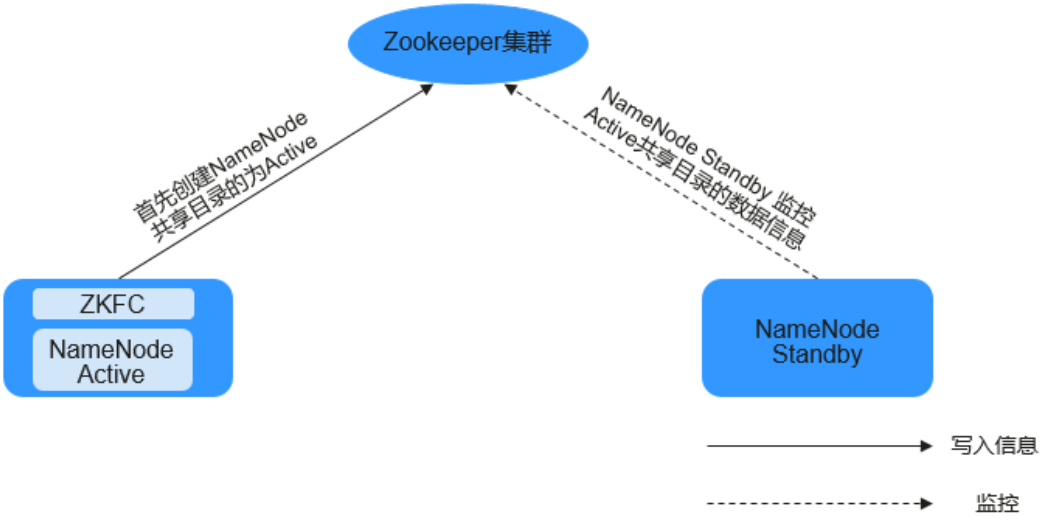
- **写请求**
  a. Follower或Observer接收到写请求后，转发给Leader。
  b. Leader协调各Follower，通过投票机制决定是否接受该写请求。
  c. 如果超过半数以上的Leader、Follower节点返回写入成功，那么Leader提交该请求并返回成功，否则返回失败。
  d. Follower或Observer返回写请求处理结果。
- **只读请求**
  客户端直接向Leader、Follower或Observer读取数据。

1.4.29.2 ZooKeeper 与其他组件的关系

ZooKeeper 和 HDFS 的关系

ZooKeeper与HDFS的关系如图1-140所示。

图 1-140 ZooKeeper 和 HDFS 的关系



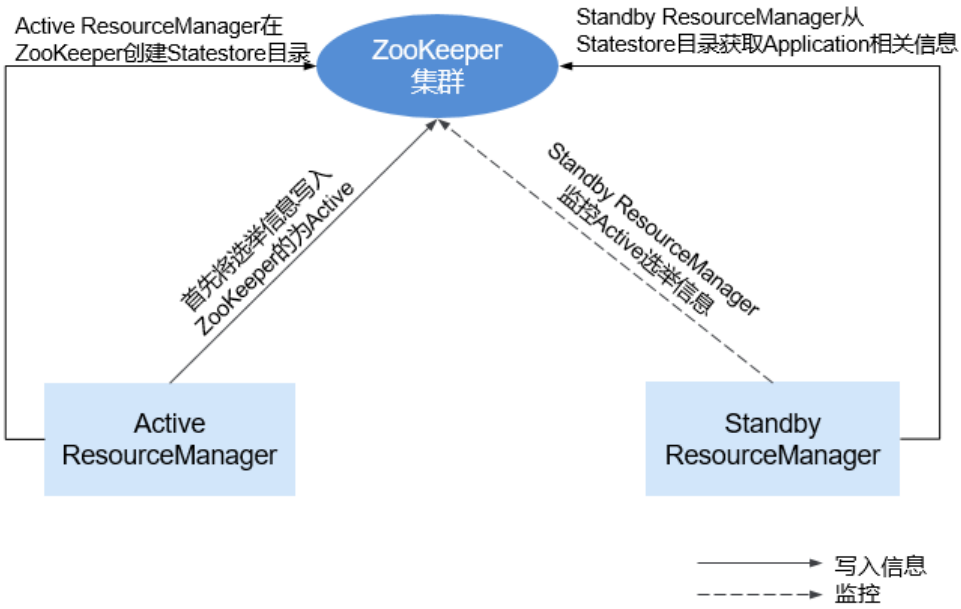
ZKFC（ZKFailoverController）作为一个ZooKeeper集群的客户端，用来监控NameNode的状态信息。ZKFC进程仅在部署了NameNode的节点中存在。HDFS NameNode的Active和Standby节点均部署有zkfc进程。

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

ZooKeeper 和 YARN 的关系

ZooKeeper与YARN的关系如图1-141所示。

图 1-141 ZooKeeper 与 YARN 的关系

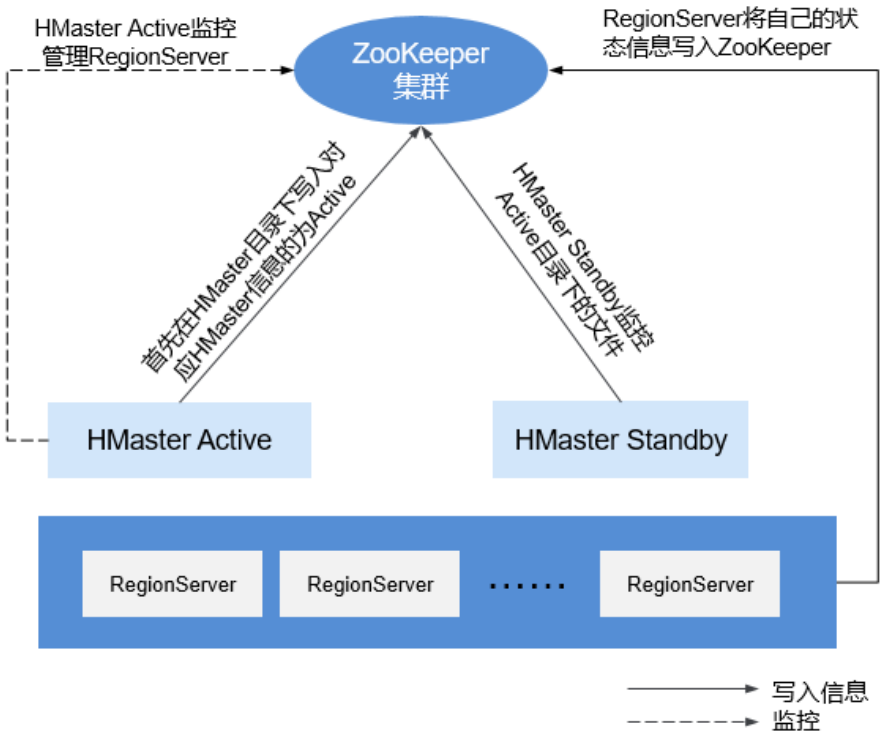


1. 在系统启动时，ResourceManager会尝试把选举信息写入ZooKeeper，第一个成功写入ZooKeeper的ResourceManager被选举为Active ResourceManager，另一个为Standby ResourceManager。Standby ResourceManager定时去ZooKeeper监控Active ResourceManager选举信息。
2. Active ResourceManager还会在ZooKeeper中创建Statestore目录，存储Application相关信息。当Active ResourceManager产生故障时，Standby ResourceManager会从Statestore目录获取Application相关信息，恢复数据。

ZooKeeper 和 HBase 的关系

ZooKeeper与HBase的关系如图1-142所示。

图 1-142 ZooKeeper 和 HBase 的关系

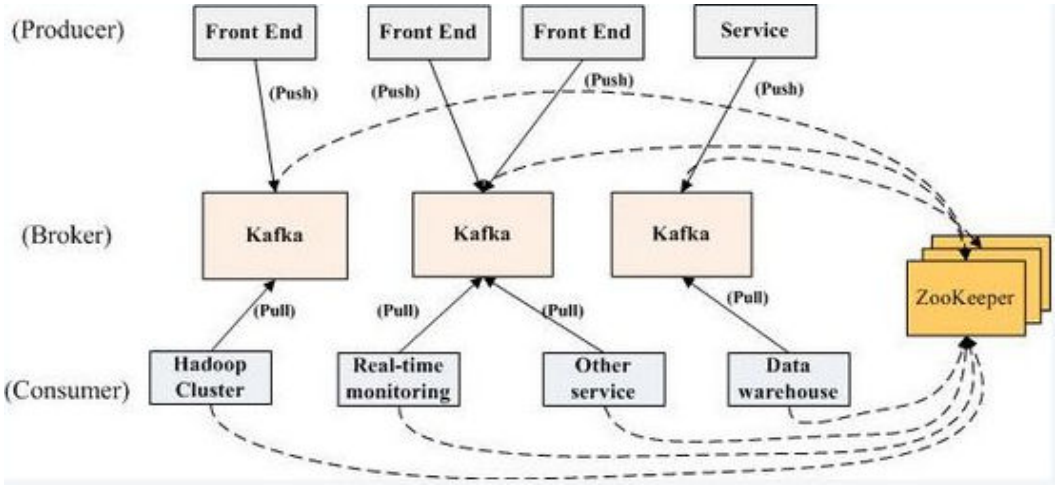


1. HRegionServer以Ephemeral node的方式注册到ZooKeeper中。其中ZooKeeper存储HBase的如下信息：HBase元数据、HMaster地址。
2. HMaster通过ZooKeeper随时感知各个HRegionServer的健康状况，以便进行控制管理。
3. HBase也可以部署多个HMaster，类似HDFS NameNode，当HMaster主节点出现故障时，HMaster备用节点会通过ZooKeeper获取主HMaster存储的整个HBase集群状态信息。即通过ZooKeeper实现避免HBase单点故障问题的问题。

ZooKeeper 和 Kafka 的配合关系

ZooKeeper与Kafka的关系如图 ZooKeeper和Kafka的关系所示。

图 1-143 ZooKeeper 和 Kafka 的关系



1. Broker端使用ZooKeeper用来注册broker信息，并进行partition leader选举。
2. Consumer端使用ZooKeeper用来注册consumer信息，其中包括consumer·消费的partition列表等，同时也用来发现broker列表，并和partition leader建立socket连接，并获取消息。

1.4.29.3 ZooKeeper 开源增强特性

日志增强

安全模式下，Ephemeral node（临时节点）在session过期之后就会被系统删除，在审计日志中添加Ephemeral node被删除的审计日志，以便了解当时Ephemeral node的状态信息。

所有ZooKeeper客户端的操作都要在审计日志中添加Username。

从ZooKeeper客户端创建znode，其kerberos principal是“zkcli/hadoop.*<系统域名>*@*<系统域名>*”。

例如打开日志<ZOO_LOG_DIR>/zookeeper_audit.log，内容如下：

```
2016-12-28 14:17:10,505 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test1?result=success
2016-12-28 14:17:10,530 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test2?result=success
2016-12-28 14:17:10,550 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test3?result=success
2016-12-28 14:17:10,570 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test4?result=success
2016-12-28 14:17:10,592 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test5?result=success
2016-12-28 14:17:10,613 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test6?result=success
2016-12-28 14:17:10,633 | INFO  | CommitProcWorkThread-4 | session=0x12000007553b4903?
user=10.177.223.78,zkcli/hadoop.hadoop.com@HADOOP.COM?ip=10.177.223.78?operation=create znode?
target=ZooKeeperServer?znode=/test7?result=success
```

输出显示了在审计日志中添加了ZooKeeper客户端用户“zkcli/hadoop.hadoop.com@HADOOP.COM”的日志。

ZooKeeper中的用户详情：

在ZooKeeper中，不同的认证方案使用不同的凭证作为用户。 基于认证供应商的要求，任何参数都可以被认为是用户。

示例：

- SAMLAuthenticationProvider使用客户端主体作为用户。
- X509AuthenticationProvider使用户客户端证书作为用户。
- IAuthenticationProvider使用客户端IP作为用户。
- 自定义认证提供程序实现org.apache.zookeeper.server.auth.ExtAuthenticationProvider.getUserName（String）方法以获取用户名。 如果没有实现，从认证提供程序实例获取用户名将被跳过。

ZooKeeper 开源增强特性：ZooKeeper SSL 通信（Netty 连接）

ZooKeeper设计最初含有Nio包，且不能较好的支持3.5版本后的SSL。为了解决这个问题，Netty被加入到ZooKeeper中。所以如果用户需要使用SSL，启用Netty并设置Server端和Client端的以下参数。

开源的服务端只支持简单的文本密码，这可能导致相关安全问题。为此我们在服务端将不再使用此类文本密码。

- Client端
  a. 将“zkCli.sh/zkEnv.sh”文件中的参数“-Dzookeeper.client.secure”设置为“true”以在Client端使用安全通信。之后客户端可以连接服务端的secureClientPort。
  b. 通过设置“zkCli.sh/zkEnv.sh”文件中的以下参数配置客户端环境。

参数	描述
-Dzookeeper.clientCnxnSocket	用于客户端的Netty通信。 默认值： "org.apache.zookeeper.ClientCnxnSocketNetty"
-Dzookeeper.ssl.keyStore.location	keystore文件路径。
-Dzookeeper.ssl.keyStore.password	加密密码。
-Dzookeeper.ssl.trustStore.location	truststore文件路径。
-Dzookeeper.ssl.trustStore.password	加密密码。
-Dzookeeper.config.crypt.class	用于加密密码的解密。
-Dzookeeper.ssl.password.encrypted	默认值：false 当keystore和truststore的密码为加密密码时设置为true。

参数	描述
-Dzookeeper.ssl.enabled.protocols	通过配置此参数定义SSL协议以适用于SSL上下文。
-Dzookeeper.ssl.exclude.cipher.ext	通过配置此参数定义SSL上下文中应排除的密码列表，之间以逗号间隔。

说明

以上参数须在“zkCli.sh/zk.Env.sh”文件内设置。

- Server端
  a. 在文件“zoo.cfg”中将监听SSL端口参数“secureClientPort”设置为“3381”。
  b. 在server端将文件“zoo.cfg”中的参数“zookeeper.serverCnxnFactory”设置为“org.apache.zookeeper.server.NettyServerCnxnFactory”。
  c. 设置文件zoo.cfg(路径：“zookeeper/conf/zoo.cfg”)中的以下参数来配置服务端环境。

参数	描述
ssl.keyStore.location	keystore.jks文件路径。
ssl.keyStore.password	加密密码。
ssl.trustStore.location	truststore文件路径。
ssl.trustStore.password	加密密码。
config.crypt.class	用于加密密码的解密。
ssl.keyStore.password.encrypted	默认值：false 设置为true时可使用加密密码。
ssl.trustStore.password.encrypted	默认值：false 设置为true时可使用加密密码。
ssl.enabled.protocols	通过配置此参数定义SSL协议以适用于SSL上下文。
ssl.exclude.cipher.ext	通过配置此参数定义SSL上下文中应排除的密码列表，之间以逗号间隔。

  d. 启动ZKserver，然后将安全客户端连接到安全端口。
- 凭证
  ZooKeeper上Client和Server之间的凭证由X509AuthenticationProvider执行。根据以下参数指定服务端证书及信任客户端证书，并通过这些证书初始化X509AuthenticationProvider。

- zookeeper.ssl.keyStore.location
- zookeeper.ssl.keyStore.password
- zookeeper.ssl.trustStore.location
- zookeeper.ssl.trustStore.password

说明

若用户不想使用ZooKeeper的默认机制，可根据所需配置不同的ZooKeeper信任机制。

1.5 产品功能

1.5.1 多租户

特性简介

现代企业的数据集群在向集中化和云化方向发展，企业级大数据集群需要满足：

- 不同用户在集群上运行不同类型的应用和作业（分析、查询、流处理等），同时存放不同类型和格式的数据。
- 某些类型的用户（例如银行、政府单位等）对数据安全非常关注，很难容忍将自己的数据与其他用户放在一起。

这给大数据集群带来了以下挑战：

- 合理地分配和调度资源，以支持多种应用和作业在集群上平稳运行。
- 对不同的用户进行严格的访问控制，以保证数据和业务的安全。

多租户将大数据集群的资源隔离成一个个资源集合，彼此互不干扰，用户通过“租用”需要的资源集合，来运行应用和作业，并存放数据。在大数据集群上可以存在多个资源集合来支持多个用户的不同需求。

因此，MRS大数据集群提供了完整的企业级大数据多租户解决方案。多租户是MRS大数据集群中的多个资源集合（每个资源集合是一个租户），具有分配和调度资源（资源包括计算资源和存储资源）的能力。

特性优势

- 合理配置和隔离资源

  租户之间的资源是隔离的，一个租户对资源的使用不影响其它租户，保证了每个租户根据业务需求去配置相关的资源，可提高资源利用效率。
- 测量和统计资源消费

  系统资源以租户为单位进行计划和分配，租户是系统资源的申请者和消费者，其资源消费能够被测量和统计。
- 保证数据安全和访问安全

  多租户场景下，分开存放不同租户的数据，以保证数据安全；控制用户对租户资源的访问权限，以保证访问安全。

调度器增强

多租户根据调度器类型分为开源的Capacity调度器和自主研发的增强型Superior调度器。

为满足企业需求，克服Yarn社区在调度上遇到的挑战与困难，自主研发的Superior调度器，不仅集合了当前Capacity调度器与Fair调度器的优点，还做了以下增强：

- 增强资源共享策略

  Superior调度器支持队列层级，在同集群集成开源调度器的特性，并基于可配置策略进一步共享资源。针对实例，MRS集群管理员可通过Superior调度器为队列同时配置绝对值或百分比的资源策略计划。Superior调度器的资源共享策略将YARN的标签调度增强为资源池特性，YARN集群中的节点可根据容量或业务类型不同，进行分组以使队列更有效地利用资源。

- 基于租户的资源预留策略

  部分租户可能在某些时间中运行关键任务，租户所需的资源应保证可用。Superior调度器构建了支持资源预留策略的机制，在这些租户队列运行的任务可立即获取到预留资源，以保证计划的关键任务可正常执行。

- 租户和资源池的用户公平共享

  Superior调度器提供了队列内用户间共享资源的配置能力。每个租户中可能存在不同权重的用户，高权重用户可能需要更多共享资源。

- 大集群环境下的调度性能优势

  Superior调度器接收到各个NodeManager上报的心跳信息，并将资源信息保存在内存中，使得调度器能够全局掌控集群的资源使用情况。Superior调度器采用了push调度模型，令调度更加精确、高效，大大提高了大集群下的资源使用率。另外，Superior调度器在NodeManager心跳间隔较大的情况下，调度性能依然优异，不牺牲调度性能，也能避免大集群环境下的“心跳风暴”。

- 优先策略

  当某个服务在获取所有可用资源后还无法满足最小资源的要求，则会发生优先抢占。抢占功能默认关闭。

1.5.2 安全增强

MRS作为一个海量数据管理和分析的平台，具备高安全性。MRS主要从以下几个方面保障用户的数据和业务运行安全。

- 网络隔离

  整个系统部署在云上的虚拟私有云中，提供隔离的网络环境，保证集群的业务、管理的安全性。结合虚拟私有云的子网划分、路由控制、安全组等功能，为用户提供高安全、高可靠的网络隔离环境。

- 资源隔离

  MRS服务支持资源专属区内部署，专属区内物理资源隔离，用户可以在专属区内灵活地组合计算存储资源，包括专属计算资源+共享存储资源、共享计算资源+专属存储资源、专属计算资源+专属存储资源。

- 主机安全

  MRS支持与云安全服务集成，支持漏洞扫描、安全防护、应用防火墙、堡垒机、网页防篡改等。针对操作系统和端口部分，提供如下安全措施：

  - 操作系统内核安全加固
  - 更新操作系统最新补丁
  - 操作系统权限控制
  - 操作系统端口管理
  - 操作系统协议与端口防攻击

- 应用安全

  通过如下措施保证大数据业务正常运行：

  - 身份鉴别和认证
  - Web应用安全
  - 访问控制
  - 审计安全
  - 密码安全

- 数据安全

  针对海量用户数据，提供如下措施保障客户数据的机密性、完整性和可用性。

  - 容灾：MRS支持将数据备份到OBS（对象存储服务）中，支持跨区域的高可靠性。
  - 备份：MRS支持针对DBService、NameNode、LDAP的元数据备份和对HDFS、HBase的业务数据备份。

- 数据完整性

  通过数据校验，保证数据在存储、传输过程中的数据完整性。

  - 用户数据保存在HDFS上，HDFS默认采用CRC32C校验数据的正确性。
  - HDFS的DataNode节点负责存储校验数据，如果发现客户端传递过来的数据有异常（不完整）就上报异常给客户端，让客户端重新写入数据。
  - 客户端从DataNode读数据的时候会同步检查数据是否完整，如果发现数据不完整，尝试从其它的DataNode节点上读取数据。

- 数据保密性

  MRS分布式文件系统在Apache Hadoop版本基础上，提供对文件内容的加密存储功能，避免敏感数据明文存储，提升数据安全性。业务应用只需对指定的敏感数据进行加密，加解密过程业务完全不感知。在文件系统数据加密基础上，Hive实现表级加密，HBase实现列族级加密，在创建表时指定采用的加密算法，即可实现对敏感数据的加密存储。

  从数据的存储加密、访问控制来保障用户数据的保密性。

  - HBase支持将业务数据存储到HDFS前进行压缩处理，且用户可以配置AES和SMS4算法加密存储。
  - 各组件支持本地数据目录访问权限设置，无权限用户禁止访问数据。
  - 所有集群内部用户信息提供密文存储。

- 安全认证
  - 基于用户和角色的认证统一体系，遵从账户/角色RBAC（Role-Based Access Control）模型，实现通过角色进行权限管理，对用户进行批量授权管理。
  - 支持安全协议Kerberos，MRS使用LDAP作为账户管理系统，并通过Kerberos对账户信息进行安全认证。
  - 提供单点登录，统一了MRS系统用户和组件用户的管理及认证。
  - 对登录Manager的用户进行审计。

1.5.3 组件 WebUI 便捷访问

大数据组件都有自己的WebUI页面管理自身系统，但是由于网络隔离的原因，用户并不能很简便地访问到该页面。比如访问HDFS的WebUI页面，传统的操作方法是需要用户创建ECS，使用ECS远程登录组件的UI，这使得组件的页面UI访问很是繁琐，对于很多初次接触大数据的用户很不友好。

MRS提供了基于弹性公网IP来便捷访问组件WebUI的安全通道，并且比用户自己绑定弹性公网IP更便捷，只需界面鼠标操作，即可简化原先用户需要自己登录虚拟私有云添加安全组规则，获取公网IP等步骤，减少了用户操作步骤。分析集群Hadoop、Spark、HBase、Hue及流式集群Storm，都可以在Manager上找到组件页面入口，快速访问。

1.5.4 可靠性增强

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

系统可靠性

- 管理节点均实现HA

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

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

- 异常场景下的可靠性保证

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

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

- 数据备份与恢复

  为应对数据丢失或损坏对用户业务造成不利影响，在异常情况下快速恢复系统，MRS根据用户业务的需要提供全量备份、增量备份和恢复功能。

  - 自动备份

    MRS对集群管理系统Manager上的数据提供自动备份功能，根据制定的备份策略可自动备份集群上的数据，包括LdapServer、DBService的数据。

  - 手动备份

    在系统进行扩容、打补丁等重大操作前，需要通过手动备份集群管理系统的数据，以便在系统故障时，恢复集群管理系统功能。

    为进一步提供系统的可靠性，在将Manager、HBase上的数据备份到第三方服务器时，也需要通过手动备份。

节点可靠性

- 操作系统健康状态监控

  周期采集操作系统硬件资源使用率数据，包括CPU、内存、硬盘、网络等资源的使用率状态。

- 进程健康状态监控

  MRS提供业务实例的状态以及业务实例进程的健康指标的检查，能够让用户第一时间感知进程健康状态。

- 硬盘故障的自动处理

  MRS对开源版本进行了增强，可以监控各节点上的硬盘以及文件系统状态。如果出现异常，立即将相关分区移出存储池；如果硬盘恢复正常（通常是因为用户更换了新硬盘），也会将新硬盘重新加入业务运作。这样极大简化了维护人员的工作，更换故障硬盘可以在线完成；同时用户可以设置热备盘，从而极大缩减了故障硬盘的修复时间，有利于提高系统的可靠性。

- 节点磁盘LVM配置

  MRS支持将多个磁盘配置成LVM（Logic Volume Management），多个磁盘规划成一个逻辑卷组。配置成LVM可以避免各磁盘间使用不均的问题，保持各个磁盘间均匀使用在HDFS和Kafka等能够利用多磁盘能力的组件上尤其重要。并且LVM可以支持磁盘扩容时不需要重新挂载，避免了业务中断。

数据可靠性

MRS可以利用弹性云服务器ECS提供的反亲和节点组以及放置组的能力，结合Hadoop的机架感知能力，将数据冗余到多个物理宿主机上，避免物理硬件的失效造成数据的失效。

1.5.5 作业管理

作业管理为用户提供向集群提交作业的入口，支持包括MapReduce、Spark、HiveQL和SparkSQL等类型的作业。结合数据治理中心DataArts Studio，提供一站式的大数据协同开发环境、全托管的大数据调度能力，帮助用户快速构建大数据处理中心。

通过数据治理中心DataArts Studio，用户可以先在线开发调试MRS HiveQL/SparkSQL脚本、拖拽式地开发MRS作业，完成MRS与其他20多种异构数据源之间的数据迁移和数据集成；通过强大的作业调度与灵活的监控告警，轻松管理数据作业运维。

1.5.6 自定义引导操作

特性简介

MRS提供标准的云上弹性大数据集群，目前可安装部署包括hadoop、spark等9种大数据组件。当前标准的云上大数据集群不能满足所有用户需求，例如如下几种场景：

- 通用的操作系统配置不能满足实际数据处理需求，例如需调大系统最大连接数。
- 需要安装自身业务所需的软件工具或运行环境，例如须安装gradle、业务需要依赖R语言包。
- 根据自身业务对大数据组件包做修改，例如对hadoop或spark安装包做修改。
- 需要安装其他MRS还未支持的大数据组件。

对于上述定制化的场景，可以选择登录到每个节点上手动操作，之后每扩容一个新节点，再执行一次同样的操作，操作相对繁琐，也容易出错。同时手动执行记录不便追溯，不能实现“按需创建、创建成功后即处理数据”的目标。

因此，MRS提供了自定义引导操作，在启动集群组件前（或后）可以在指定的节点上执行脚本。用户可以通过引导操作来完成安装MRS还没支持的第三方软件，修改集群运行环境等自定义操作。如果集群扩容，选择执行引导操作，则引导操作也会以相同

方式在新增节点上执行。MRS会使用root用户执行您指定的脚本，脚本内部您可以通过**su - *xxx***命令切换用户。

客户价值

MRS提供了自定义引导操作，用户可以以此为入口，灵活、便捷地配置自己的专属集群，自定义安装软件。

1.5.7 元数据

当创建MRS集群时选择部署Hive和Ranger组件时，MRS提供多种元数据存储方式，您可以根据自身需要进行选择：

- 本地元数据：元数据存储于集群内的本地GaussDB中，当集群删除时元数据同时被删除，如需保存元数据，需提前前往数据库手动保存元数据。
- 数据连接：可选择关联与当前集群同一虚拟私有云和子网的RDS服务中的PostgresDB或MySQL数据库，元数据将存储于关联的数据库中，不会随当前集群的删除而删除，多个MRS集群可共享同一份元数据。

说明

Hive组件可选元数据存储方式功能在MRS 1.9.x及之后版本支持。

Ranger组件可选元数据存储方式功能目前仅在MRS 1.9.x版本支持关联RDS服务中的MySQL数据库。

1.5.8 集群管理

1.5.8.1 集群生命周期管理

MRS支持集群的生命周期管理包括创建集群和删除集群。

- 创建集群：支持用户定制集群的类型，组件范围，各类型的节点数、虚拟机规格、可用区、VPC网络、认证信息，MRS将为用户自动创建一个符合配置的集群，全程无需用户参与；同时支持用户在集群中运行自定义内容；支持快速创建多应用场景集群，比如创建Hadoop分析集群、HBase集群、Kafka集群。大数据平台同时支持部署异构集群，在集群中存在不同规格的虚机，允许在CPU类型，硬盘容量，硬盘类型，内存大小灵活组合。在集群中支持多种虚机规格混合使用。
- 删除集群：当集群不再需要时（包括集群中的数据和配置），用户可以选择删除集群，MRS会将集群相关的资源全部删除。

创建集群

通过在MRS服务管理面，客户可以创建MRS集群，通过选择集群所建的区域及使用的云资源规格，一键式创建适合企业业务的MRS集群。MRS服务会根据用户选择的集群类型、版本和节点规格，帮助客户自动完成企业级大数据平台的安装部署和参数调优。

MRS服务为客户提供完全可控的大数据集群，客户在创建时可设置虚拟机的登录方式（密码或者密钥对），所创建的MRS集群资源完全归客户所用。同时MRS支持在最小可在两节点4U8G的ECS上部署大数据集群，为客户测试开发提供更多的灵活选择。

MRS集群类型包括分析集群、流式集群和混合集群。

- 分析集群：用来做离线数据分析，提供的是Hadoop体系的组件。
- 流式集群：用来做流处理任务，提供的是流式处理组件。
- 混合集群：既可以用来做离线数据分析，又可以用来做流处理任务，提供的是Hadoop体系的组件和流式处理组件。
- 自定义：根据业务需求，可以灵活搭配所需组件（MRS 3.x及后续版本）。

MRS集群节点类型包括Master节点、Core节点和Task节点。

- Master节点：集群中的管理节点，分布式系统的Master进程和Manager以及数据库均部署在该节点；该类型节点不可扩容。该类型节点的处理能力决定了整个集群的管理上限，MRS服务支持将Master节点规格提高，以支持更大集群的管理。
- Core节点：支持存储和计算两种目标的节点，可扩容、缩容。因承载的数据存储，因此在缩容时，为保证数据不丢失，有较多限制，无法进行弹性伸缩。
- Task节点：仅用于计算的节点，可扩容、缩容。因只承载计算任务，因此可以进行弹性伸缩。

MRS创建集群方式支持自定义创建集群和快速创建集群两种。快速创建集群请参考**快速集群**，自定义购买集群请参考**创建自定义集群**。

- 自定义创建集群：自定义创建可以灵活地选择配置项，针对不同的应用场景，可以选择不同规格的弹性云服务器，全方位贴合您的业务诉求。
- 快速创建集群：用户可以根据应用场景，快速创建对应配置的集群，提高了配置效率，更加方便快捷。当前支持快速创建Hadoop分析集群、HBase集群、Kafka集群。
  - Hadoop分析集群：Hadoop分析集群完全使用开源Hadoop生态，采用YARN管理集群资源，提供Hive、Spark离线大规模分布式数据存储和计算，SparkStreaming、Flink流式数据计算，Presto交互式查询，Tez有向无环图的分布式计算框等Hadoop生态圈的组件，进行海量数据分析与查询。
  - HBase集群：HBase集群使用Hadoop和HBase组件提供一个稳定可靠，性能卓越、可伸缩、面向列的分布式云存储系统，适用于海量数据存储以及分布式计算的场景，用户可以利用HBase搭建起TB至PB级数据规模的存储系统，对数据轻松进行过滤分析，毫秒级得到响应，快速发现数据价值。
  - Kafka集群：Kafka集群使用Kafka和Storm组件提供一个开源高吞吐量，可扩展性的消息系统。广泛用于日志收集、监控数据聚合等场景，实现高效的流式数据采集，实时数据处理存储等。

删除集群

MRS服务支持用户在不需要大数据集群时执行删除集群操作，集群删除后，所有大数据使用的相关云资源都会同时被释放。删除集群前，建议完成数据搬迁或者备份，确认集群无任何业务运行或者集群异常且经运维分析无法继续提供服务时再执行集群删除操作。对于数据存放在云硬盘EVS或直通盘的大数据集群，集群删除后，数据也随之删除，强烈建议您慎重选择删除集群。

1.5.8.2 集群扩缩容

大数据集群的处理能力通常可以通过增加集群的节点数来横向扩展，当集群规模不符合业务要求时，用户可以通过该功能进行集群节点规模的调整，进行扩容或者缩容；在缩容节点时，MRS会智能地选择负载最少或者迁移数据量最小节点，并且在缩容过程中，缩容节点不再接收新的任务，正在执行的任务继续执行，同时将该节点数据拷贝至其他节点，该节点进入退服状态，当该节点任务长时间运行无法结束时，会迁移至其他节点运行，最大限度地减少对集群业务的影响。

扩容集群

目前支持扩容集群Core节点或Task节点，用户可通过增加节点数量处理业务峰值负载。MRS集群节点扩容中和扩容后对现有集群的业务没有影响，扩容后引起的数据倾斜问题可进行规避。

缩容集群

用户可以根据业务需求量，通过简单的缩减Core节点或者Task节点，对集群进行缩容，以使MRS拥有更优的存储、计算能力，降低运维成本。用户执行MRS集群缩容后，MRS服务将根据节点已安装的服务类型自动选择可以缩容的节点。

Core节点在缩容的时候，会对原节点上的数据进行迁移。业务上如果对数据位置做了缓存，客户端自动刷新位置信息可能会影响时延。缩容节点可能会影响部分HBase on HDFS数据的第一次访问响应时长，可以重启HBase或者对相关的表Disable/Enable来避免。

Task节点本身不存储集群数据，属于计算节点，不存在节点数据迁移的问题。

1.5.8.3 自动弹性伸缩

特性简介

随着企业的数据越来越多，越来越多的企业选择使用Spark/Hive等技术来进行分析，由于数据量大，处理任务繁重，资源的消耗比较高，因此使用成本也是比较高。当前并不是每个企业在每时每刻在进行分析，而一般是在一天的一个时间段内进行分析汇总，因此MRS提供了弹性伸缩能力，可以自动在业务在繁忙时申请额外资源，业务不繁忙时释放闲置资源，让用户按需使用，尽可能的帮助客户降低使用成本，聚焦核心业务。

在大数据应用，尤其是周期性的数据分析处理场景中，需要根据业务数据的周期变化，动态调整集群计算资源以满足业务需要。MRS的弹性伸缩规则功能支持根据集群负载对集群进行弹性伸缩。此外，如果数据量为周期有规律的变化，并且希望在数据量变化前提前完成集群的扩缩容，可以使用MRS的资源计划特性。

MRS服务支持规则和时间计划两种弹性伸缩的策略：

- 弹性伸缩规则：根据集群实时负载对Task节点数量进行调整，数据量变化后触发扩缩容，有一定的延后性。
- 资源计划：若数据量变化存在周期性规律，则可通过资源计划在数据量变化前提前完成集群的扩缩容，避免出现增加或减少资源的延后。

弹性伸缩规则与资源计划均可触发弹性伸缩，两者即可同时配置也可单独配置。资源计划与基于负载的弹性伸缩规则叠加使用可以使得集群节点的弹性更好，足以应对偶尔超出预期的数据峰值出现。

当某些业务场景要求在集群扩缩容之后，根据节点数量的变化对资源分配或业务逻辑进行更改时，手动扩缩容的场景客户可以登录集群节点进行操作。对于弹性伸缩场景，MRS支持通过自定义弹性伸缩自动化脚本来解决。自动化脚本可以在弹性伸缩前后执行相应操作，自动适应业务负载的变化，免去了人工操作。同时，自动化脚本给用户实现个性需求提供了途径，完全自定义的脚本与多个可选的执行时机基本可以满足用户的各项需求，使弹性伸缩更具灵活性。

客户价值

MRS的自动弹性伸缩可以帮助用户实现以下价值。

- 降低使用成本

  部分企业在进行批量分析时，并不是时时刻刻都在进行分析，例如一般都存在数据持续接入，而到了特定时间段（例如凌晨3点）进行批量分析，可能仅需要消耗2小时。

  MRS提供的弹性伸缩能力，可以帮助客户，在晚上的时候，将分析节点扩容到指定规模，而计算完毕后，则自动释放计算节点，尽可能的降低使用成本。

- 平衡突发查询

  大数据集群上，由于有大量的数据，企业会经常面临临时的分析任务，例如支撑企业决策的临时数据报表等，都会导致对于资源的消耗在极短时间内剧增。MRS提供的弹性伸缩能力，可以让突发大数据分析时，可以及时的补充计算节点，避免因为计算能力不足，导致业务宕机，使用户无需创建额外资源，当突发事件结束后，MRS会自动判断缩容时机，自动完成缩容。

- 聚焦核心业务

  大数据作为二次开发平台，开发人员非常难判断具体的资源消耗，因为查询分析的条件复杂性（例如全局排序，过滤，合并等）以及数据的复杂性，例如增量数据的不确定性等，都会导致预估多少计算量是非常困难的行为，而使用弹性伸缩能力，可以让业务人员专注于业务开发，无需分心再做各种资源评估。

1.5.8.4 创建 Task 节点

特性简介

支持创建Task节点，只作为计算节点，不存放持久化的数据，是实现弹性伸缩的基础。

客户价值

在MRS服务只作为计算资源的场景下，使用Task节点可以节省成本，并可以更加方便快捷地对集群节点进行扩缩容，满足用户对集群计算能力随时增减的需求。

用户场景

当集群数据量变化不大而集群业务处理能力需求变化比较大，大的业务处理能力只是临时需要，此时选择添加Task节点。

- 临时业务量增大，如年底报表处理。
- 需要在短时间内处理完原来需要处理很久的任务，如一些紧急分析任务。

1.5.8.5 升级 Master 节点规格

MRS大数据集群采用Manager实现集群的管理，而管理集群的相关服务，如HDFS存储系统的NameNode，Yarn资源管理的ResourceManager，以及MRS的Manager管理服务都部署在集群的Master节点。

随着新业务的上线，集群规模不断扩大，Master节点承担的管理负荷也越来越高，企业用户面临CPU负载过高，内存使用率超过阈值的问题。通常自建大数据集群需要完成数据搬迁，采购升级节点硬件配置实现Master规格提升，而MRS服务借助云服务的

优势，实现一键式Master节点升级，并在升级过程中通过Master节点的主备HA保证已有业务的不间断，方便快捷帮助用户解决主节点规格升级问题。

1.5.8.6 隔离主机

用户发现某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。在为集群安装补丁的场景中，也支持排除指定节点不安装补丁。隔离主机仅支持隔离非管理节点。

主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。另外，主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。 隔离主机具体请参考**隔离主机**。

1.5.8.7 标签管理

标签是集群的标识，为集群添加标签，可以方便用户识别和管理拥有的集群资源。MRS服务通过与标签管理服务（TMS）关联，可以让拥有大量云资源的用户，通过给云资源打标签，快速查找具有同一标签属性的云资源，进行统一检视、修改、删除等管理操作，方便用户对大数据集群及其他相关云资源的统一管理。

您可以在创建集群时添加标签，也可以在集群创建完成后，在集群的详情页添加标签，您最多可以给集群添加10个标签。

1.5.9 集群运维

告警管理

MRS可以实时监控大数据集群，通过告警和事件可以识别系统健康状态。同时MRS也支持用户自定义配置监控与告警阈值用于关注各指标的健康情况，当监控数据达到告警阈值，系统将会触发一条告警信息。

MRS还可以与消息通知服务(SMN)的消息服务系统对接，将告警信息通过短信或者邮件等形式推送给用户。具体介绍请参见**消息通知**。

补丁管理

MRS集群支持补丁操作，会及时发布开源大数据组件的补丁。用户能够在MRS集群管理页面上查看到运行集群相关的补丁发布信息，包括其修复问题的详细说明及影响场景，客户可以根据业务运行情况自行选择是否安装补丁。补丁安装过程是一键式操作，无需人工干预，通过滚动安装，补丁升级不会停止业务，保障用户集群长期可用。

MRS服务可以展示详细的补丁安装过程，补丁管理也支持补丁的卸载和失败回滚。

说明

MRS 3.x及之后版本暂不支持在管理控制台执行补丁管理操作。

运维支撑

MRS提供的集群的资源是完全属于用户的，通常情况下，当集群出现问题，需要运维人员支撑时，运维人员是无法直接访问的。为了更好的服务客户，MRS提供两种方式来减少定位问题时的信息传递：

- 日志共享：用户可以在MRS 页面发起日志共享，选择日志范围共享给运维人员，以便运维人员在不接触集群的情况下帮助定位问题。
- 运维授权：MRS服务提供运维授权功能，用户在使用MRS集群过程中，发生问题可以在MRS页面发起运维授权，由运维人员帮助客户快速定位问题，用户可以随时收回该授权。

健康检查

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

1.5.10 消息通知

特性简介

大数据集群运行过程中经常会进行如下操作：

- 大数据集群经常会发生变更，比如扩容、缩容集群。
- 业务数据量突然变化，集群触发弹性伸缩。
- 相关业务结束，需要终止大数据集群等。

用户想要及时得知这些操作是否成功了，以及当集群出现大数据服务不可用，或节点故障时，用户希望不用隔段时间就登录集群查看，而是可以及时地收到告警通知。MRS联合消息通知服务(SMN)，可以将以上信息主动地通知到用户的手机及邮箱，让维护更加省心省力。

客户价值

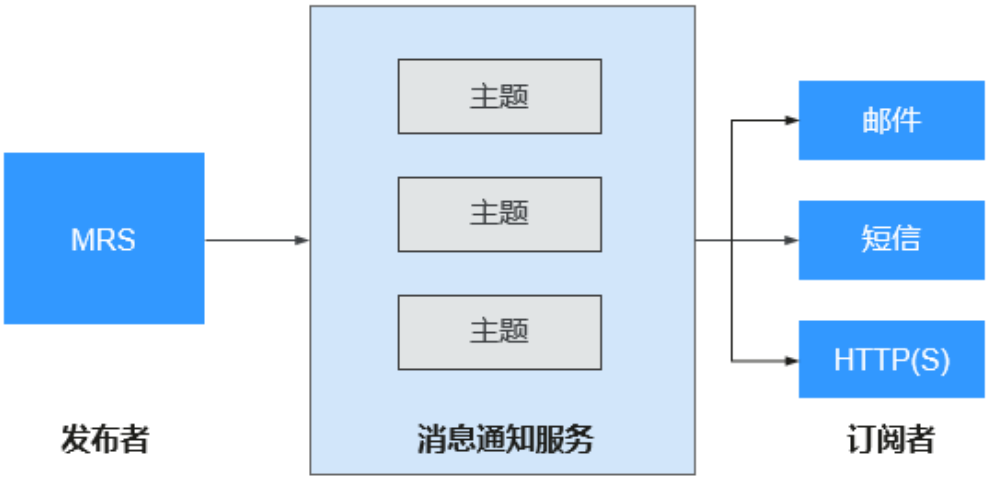
配置消息通知后，可以实时给用户发送MRS集群健康状态，用户可以通过手机短信或邮箱实时接收到MRS集群变更及组件告警信息。MRS可以帮助用户轻松运维，实时监控，实时发送告警，操作灵活，大数据业务部署更加省心省力。

特性描述

MRS联合消息通知服务(SMN)，采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式。

首先，作为主题拥有者，我们可以先创建一个主题，并对主题设置访问控制权限来决定哪些发布者和订阅者可以通过该主题进行交流。MRS将集群消息发送至您有权限发布消息的主题，然后所有订阅了该主题的订阅者（可以是手机短信、邮箱等）都将收到集群变更以及组件告警的消息。

图 1-144 实现过程



1.6 约束与限制

使用MRS前，您需要认真阅读并了解以下使用限制。

- MRS集群必须创建在VPC子网内。
- 建议使用支持的浏览器登录MRS。
  - Google Chrome：36.0及更高版本
  - Internet Explorer：9.0及更高版本
- 创建MRS集群时，支持自动创建安全组，也可从下拉框中选择已有的安全组。集群创建完成后，请勿随意删除或更改已使用的安全组。否则可以能导致集群异常，影响MRS集群的使用。
- MRS集群使用的安全组请勿随意放开权限，避免被恶意访问。
- 请勿随意执行如下操作，避免集群进入异常状态，影响MRS集群的使用。
  - 在ECS中对MRS集群的节点进行关机、重启、删除、变更OS、重装OS和修改规格等操作。
  - 删除集群节点上已有的进程、安装的应用程序和文件。
- 集群处于非人为异常状态时，可以联系技术支持人员，技术支持人员征得您同意后会请您提供密钥，登录MRS集群进行问题排查。
- 请根据业务需要规划集群节点的磁盘，如果需要存储大量业务数据，请增加云硬盘数量或存储空间。以防止存储空间不足影响节点正常运行。
- 集群节点仅用于存储用户业务数据，非业务数据建议保存在对象存储服务或其他弹性云服务器中。
- 集群节点仅用于运行MRS集群，其他客户端应用程序、用户业务程序建议申请独立弹性云服务器部署。
- MRS集群的节点（包含Master，Core和Task节点）扩充存储容量时，不建议通过扩容原有磁盘的方式实现。建议通过创建磁盘再挂载的方式实现。
- MRS集群扩容（包含存储能力和计算能力）时，可以通过增加Core节点或者Task节点的方式实现。
- 当关闭集群中的某一个Master节点后仍然使用集群执行任务或修改配置，且任务执行或配置修改后未启动被关闭的Master节点就关闭集群中其他Master节点时，

存在由于主备倒换导致的数据丢失风险。在该场景下，请在任务执行或配置修改后先启动被关闭的Master节点，再关闭全部节点。若集群中节点已经被全部关闭，请按照节点关机顺序的倒序启动集群节点。

- 当使用MRS集群过程中进行Capacity和Superior调度器切换时只完成调度器的切换，不保证配置同步。如果您需要配置同步，请基于新的调度器重新配置。

1.7 技术支持

云原生数据湖MRS 服务（MapReduce Service）是租户完全可控的半托管云服务，为用户提供一站式企业级大数据平台，用户可以在MRS集群上轻松运行Hadoop、Hive、Spark、HBase、Kafka、Flink等大数据组件，帮助企业快速构建海量数据信息处理系统，并通过对海量信息数据实时与非实时的分析挖掘，发现全新价值点和企业商机。

维护策略声明

MRS租户集群资源归属于用户，MRS提供基于该资源的半托管云服务能力，用户拥有对集群使用的完全控制权，默认情况下，云服务无权限对客户集群进行操作，集群日常运维管理由用户负责，如果在大数据集群运维过程中遇到了相关技术问题，可以联系华为云技术支持团队获得帮助，该技术支持仅协助分析处理MRS云服务相关求助，不包含云服务以外的求助，例如用户基于大数据平台构建的应用系统等。

技术支持范围

- MRS云服务管理控制台提供的相关功能：
  - 集群的创删扩缩
  - 集群作业管理
  - 集群告警管理
  - 集群补丁管理
  - IAM用户委托管理
  - 对外API接口管理
- MRS服务提供的开源大数据组件，其中开源组件请参考对应MRS版本组件列表。
- 支持客户进行MRS服务相关开源组件漏洞分析，如影响分析、修复建议，由用户负责评估对应的业务影响和负责最终实施。

1.8 权限管理

如果您需要对云上创建的MapReduce服务资源，给企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制云资源的访问。

通过IAM，您可以在云账号中给员工创建IAM用户，并授权控制他们对资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有MapReduce服务的使用权限，但是不希望他们拥有删除MRS集群等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用MRS，但是不允许删除MRS集群的权限策略，控制他们对MRS集群资源的使用范围。

如果云账号已经能满足您的要求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用MRS服务的其它功能。

MRS 权限说明

默认情况下，IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

MRS部署时通过物理区域划分，为项目级服务。授权时，“作用范围”需要选择“区域级项目”，然后在指定区域对应的项目中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问MRS时，需要先切换至授权区域。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对MRS服务，能够控制IAM用户仅能对集群进行指定的管理操作。如不允许某用户组删除集群，仅允许操作MRS集群基本操作，如创建集群、查询集群列表等。多数细粒度策略以API接口为粒度进行权限拆分。

如表1-29所示，包括了MRS的所有系统策略。

表 1-29 MRS 系统策略

策略名称	描述	策略类别
MRS FullAccess	MRS管理员权限，拥有该权限的用户可以拥有MRS所有权限。	细粒度策略
MRS CommonOperations	MRS服务普通用户权限，拥有该权限的用户可以拥有MRS服务使用权限，无新增、删除资源权限。	细粒度策略
MRS ReadOnlyAccess	MRS服务只读权限，拥有该权限的用户仅能查看MRS的资源。	细粒度策略
MRS Administrator	操作权限： • 对MRS服务的所有执行权限。 • 拥有该权限的用户必须同时拥有Tenant Guest和Server Administrator权限。	RBAC策略

表1-30列出了MRS常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 1-30 常用操作与系统策略的授权关系

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
创建集群	√	x	x	√
调整集群	√	x	x	√
升级节点规格	√	x	x	√
删除集群	√	x	x	√
查询集群详情	√	√	√	√
查询集群列表	√	√	√	√
设置弹性伸缩策略	√	x	x	√
查询主机列表	√	√	√	√
查询操作日志	√	√	√	√
创建并执行作业	√	√	x	√
停止作业	√	√	x	√
删除单个作业	√	√	x	√
批量删除作业	√	√	x	√
查询作业详情	√	√	√	√
查询作业列表	√	√	√	√
新建文件夹	√	√	x	√
删除文件	√	√	x	√
查询文件列表	√	√	√	√
批量操作集群标签	√	√	x	√
创建单个集群标签	√	√	x	√

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
删除单个集群标签	√	√	x	√
按照标签查询资源列表	√	√	√	√
查询集群标签	√	√	√	√
访问Manager页面	√	√	x	√
查询补丁列表	√	√	√	√
安装补丁	√	√	x	√
卸载补丁	√	√	x	√
运维通道授权	√	√	x	√
运维通道日志共享	√	√	x	√
查询告警列表	√	√	√	√
订阅告警消息提醒	√	√	x	√
提交SQL语句	√	√	x	√
查询SQL结果	√	√	x	√
取消SQL执行任务	√	√	x	√

MRS FullAccess

```
{
   "Version": "1.1",
   "Statement": [
      {
         "Action": [
            "mrs:*:*",
            "ecs:*:*",
            "bms:*:*",
            "evs:*:*",
            "vpc:*:*",
```

```
            "bss:*:*",
            "kms:*:*",
            "rds:*:*"
        ],
        "Effect": "Allow"
    }
  ]
}
```

MRS CommonOperations

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Action": [
        "mrs:*:get*",
        "mrs:*:list*",
        "ecs:*:get*",
        "ecs:*:list*",
        "bms:*:get*",
        "bms:*:list*",
        "evs:*:get*",
        "evs:*:list*",
        "vpc:*:get*",
        "vpc:*:list*",
        "mrs:job:submit",
        "mrs:job:stop",
        "mrs:job:delete",
        "mrs:job:checkSql",
        "mrs:job:batchDelete",
        "mrs:file:create",
        "mrs:file:delete",
        "mrs:tag:batchOperate",
        "mrs:tag:create",
        "mrs:tag:delete",
        "mrs:manager:access",
        "mrs:patch:install",
        "mrs:patch:uninstall",
        "mrs:ops:grant",
        "mrs:ops:shareLog",
        "mrs:alarm:subscribe",
        "mrs:alarm:delete",
        "bss:*:get*",
        "bss:*:list*",
        "kms:*:get*",
        "kms:*:list*",
        "rds:*:get*",
        "rds:*:list*",
        "mrs:bootstrap:*"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "mrs:cluster:create",
        "mrs:cluster:resize",
        "mrs:cluster:scaleUp",
        "mrs:cluster:delete",
        "mrs:cluster:policy"
      ],
      "Effect": "Deny"
```

```
        }
    ]
}
```

MRS ReadOnlyAccess

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "mrs:*:get*",
                "mrs:*:list*",
                "mrs:tag:count",
                "ecs:*:get*",
                "ecs:*:list*",
                "bms:*:get*",
                "bms:*:list*",
                "evs:*:get*",
                "evs:*:list*",
                "vpc:*:get*",
                "vpc:*:list*",
                "bss:*:get*",
                "bss:*:list*",
                "kms:*:get*",
                "kms:*:list*",
                "rds:*:get*",
                "rds:*:list*"
            ],
            "Effect": "Allow"
        },
        {
            "Action": [
                "mrs:cluster:create",
                "mrs:cluster:resize",
                "mrs:cluster:scaleUp",
                "mrs:cluster:delete",
                "mrs:cluster:policy",
                "mrs:job:submit",
                "mrs:job:stop",
                "mrs:job:delete",
                "mrs:job:batchDelete",
                "mrs:file:create",
                "mrs:file:delete",
                "mrs:tag:batchOperate",
                "mrs:tag:create",
                "mrs:tag:delete",
                "mrs:manager:access",
                "mrs:patch:install",
                "mrs:patch:uninstall",
                "mrs:ops:grant",
                "mrs:ops:shareLog",
                "mrs:alarm:subscribe"
            ],
            "Effect": "Deny"
        }
    ]
}
```

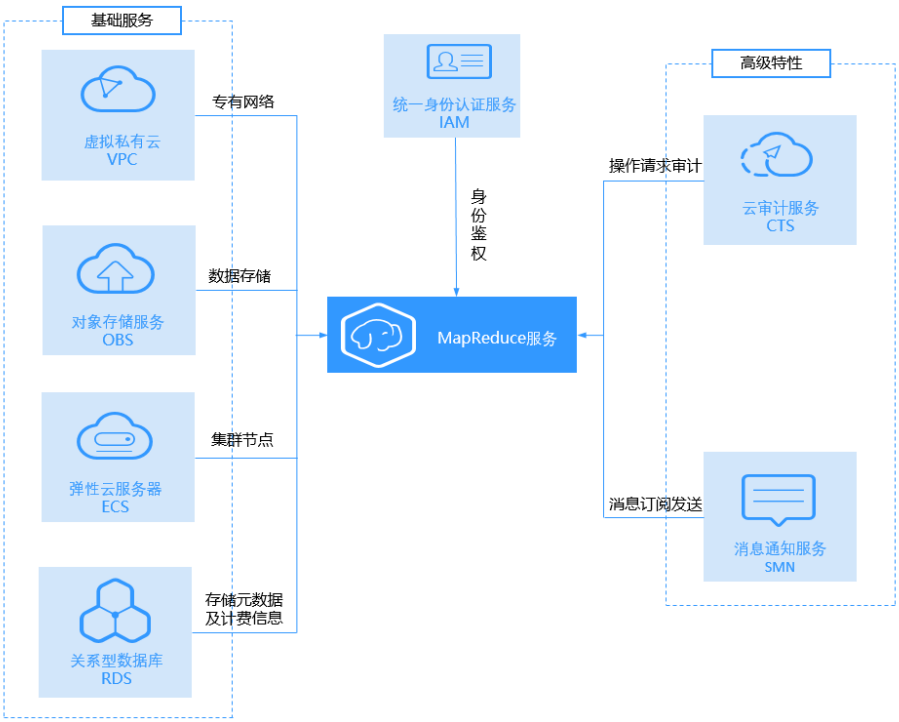
MRS Administrator

```
{
   "Version": "1.0",
   "Statement": [
      {
         "Action": [
            "MRS:MRS:*"
         ],
         "Effect": "Allow"
      }
   ],
   "Depends": [
      {
         "catalog": "BASE",
         "display_name": "Server Administrator"
      },
      {
         "catalog": "BASE",
         "display_name": "Tenant Guest"
      }
   ]
}
```

1.9 与其他云服务的关系

MRS服务与周边其他服务的关系如图图1-145所示。

图 1-145 MRS 与其他服务的关系图



MRS 服务与其他服务的关系

表 1-31 MRS 服务与其他服务的关系

服务名称	MRS服务与其他服务的关系
虚拟私有云（Virtual Private Cloud）	MRS集群创建在虚拟私有云（VPC）的子网内，VPC通过逻辑方式进行网络隔离，为用户的MRS集群提供安全、隔离的网络环境。
对象存储服务（Object Storage Service）	对象存储服务（OBS）用于存储用户数据，包括MRS作业输入数据和作业输出数据： ● MRS作业输入数据：用户程序和数据文件 ● MRS作业输出数据：作业输出的结果文件和日志文件 MRS中HDFS、Hive、MapReduce、YARN、Spark、Flume和Loader支持从OBS导入、导出数据。 MRS使用OBS的并行文件系统提供服务。
弹性云服务器（Elastic Cloud Server）	MRS服务使用弹性云服务器（Elastic Cloud Server，简称ECS）作为集群的节点，每个弹性云服务器是集群中的一个节点。
关系型数据库（Relational Database Service）	关系型数据库（RDS）用于存储MRS系统运行数据，包括MRS集群元数据等。
统一身份认证服务（Identity and Access Management）	统一身份认证服务（IAM）为MRS提供了鉴权功能。
消息通知服务（SMN）	MRS联合消息通知服务（SMN），采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式。
云审计服务（Cloud Trace Service）	云审计服务（CTS）为用户提供MRS资源操作请求及请求结果的操作记录，供用户查询、审计和回溯使用。

表 1-32 云审计支持的 MRS 操作列表

操作名称	资源类型	事件名称
创建集群	cluster_mrs	createCluster
删除集群	cluster_mrs	deleteCluster
集群扩容	cluster_mrs	scaleOutCluster
集群缩容	cluster_mrs	scaleInCluster

在您开启了云审计服务后，系统开始记录云服务资源的操作。云审计服务管理控制台保存最近7天的操作记录。详细操作步骤请参考“云审计服务（CTS）> 快速入门 > 查看追踪事件”。

1.10 配额说明

配额是用户账号在对应环境配置的可用资源额度，限定配额仅仅是为了防止资源滥用。

MapReduce服务应用的基础设施如下，配额由各个基础服务管理，如需扩大配额，请联系对应服务的技术支持进行扩容：

- 弹性云服务器
- 裸金属服务器
- 虚拟私有云
- 云硬盘
- 镜像服务
- 对象存储服务
- 弹性公网IP
- 消息通知服务
- 统一身份认证服务

1.11 常见概念

HBase 表

HBase的表是三个维度排序的映射。从行主键、列主键和时间戳映射为单元格的值。所有的数据存储在HBase的表单元格中。

列

HBase表的一个维度。列名称的格式为“<family>:<label>”，<family>和<label>为任意字符组合。表由<family>的集合组成（<family>又称为列族）。HBase表中的每个列都归属于某个列族。

列族

列族是预定义的列集合，存储在HBase Schema中。如果需要在列族下创建一些列，首先需创建列族。列族将HBase中具有相同性质的数据进行重组，且没有类型的限制。同一列族的每行数据存储在同一个服务器中。每个列族像一个属性，如压缩包、时间戳、数据块缓存等等。

MemStore

MemStore是HBase存储的核心，当WAL中数据存储达到一定量时，加载到MemStore进行排序存储。

RegionServer

RegionServer是HBase集群运行在每个工作节点上的服务。一方面维护Region的状态，提供对于Region的管理和服务；另一方面，上传Region的负载信息，参与Master的分布式协调管理。

时间戳

用于索引同一份数据的不同版本，时间戳的类型是64位整型。时间戳可以由HBase在数据写入时自动赋值或者由客户显式赋值。

Store

HBase存储的核心，一个Store拥有一个MemStore和多个StoreFile，一个Store对应一个分区中表的列族。

索引

一种数据结构，提高了对数据库表中的数据检索效率。可以使用一个数据库表中的一列或多列，提供了快速随机查找和有效访问有序记录的基础。

协处理器

HBase提供的在RegionServer执行的计算逻辑的接口。协处理器分两种类型，系统协处理器可以全局导入RegionServer上的所有数据表，表协处理器即是用户可以指定一张表使用协处理器。

Block Pool

Block Pool是隶属于单个Namespace的块的集合。DataNode存储来自集群中所有块池的块。每个块池都是独立管理的。这就允许一个Namespace为新块生成块ID，而不需要和其他Namespace合作。一个NameNode失效，不会影响DataNode为集群中其他NameNode提供服务。

DataNode

HDFS集群的工作节点。根据客户端或者是元数据节点的调度存储和检索数据，定期向元数据及客户端发送他们所存储的文件块的列表。

文件块

HDFS中存储的最小逻辑单元。每个HDFS文件由一个或多个文件块存储。所有的文件块存储在DataNode中。

文件块副本

一个副本是存储在HDFS中的一些文件块拷贝件。同一个文件块存储多个拷贝件主要用于系统的可用性和容错。

Namespace Volume

一种独立的（自给自足的）管理单元。一个Namespace和它的Block Pool，合称为“Namespace Volume”。当一个NameNode/Namespace被删除后，DataNode上的相关块池也会被删除。在集群升级时，每个命名空间卷将作为一个整体被升级。

NodeManager

负责执行应用程序的容器，同时监控应用程序的资源使用情况（CPU、内存、硬盘、网络）并且向ResourceManager汇报。

ResourceManager

集群的资源管理器，基于应用程序对资源的需求进行调度。资源管理器提供一个调度策略的插件，它负责将集群资源分配给多个队列和应用程序。调度插件可以基于现有的能力调度和公平调度模型。

分区

每一个Topic可以被分为多个Partition，每个Partition对应一个可持续追加的有序且不可变的log文件。

跟随者

跟随者（Follower）负责处理读请求的模块，配合Leader一起进行写请求处理。也可作为Leader的储备，当Leader故障时从Follower当中选举出Leader，避免出现单点故障。

观察者

观察者（Observer）不参与选举和写请求的投票，只负责处理读请求、并向Leader转发写请求，避免系统处理能力浪费。

领导者

作为ZooKeeper集群的领导者，由各Follower通过Zab协议选举产生。主要负责接受和协调所有写请求，并把写入的信息同步到Follower和Observer。

CarbonData

华为基于Spark SQL开放架构，将自研的MOLAP引擎和Spark深度集成，快速构建基于Spark的分布式多维分析引擎，使Spark分析性能从分钟级提升到秒级，增强了Spark的多维分析能力。

离散流

Spark Streaming提供的抽象概念。表示一个连续的数据流，是从数据源获取或者通过输入流转换生成的数据流。从本质上说，一个DStream表示一系列连续的RDD。

堆内存（Heap memory）

堆是JVM运行时数据区域，所有类实例和数组的内存均从此处分配。初始堆内存根据JVM启动参数-Xms控制，最大堆内存通过JVM启动参数的-Xmx进行控制。

- 最大堆内存（Maximum Heap memory）：系统可以分配给程序的最大堆内存，JVM启动参数的-xmx指定。
- 分配的堆内存（Committed Heap memory）：为保证程序运行的系统堆分配的堆内存总量，Committed heap memory在程序运行期间根据使用情况，会在初始堆内存和最大堆内存之间波动变化。
- 使用的堆内存（Used heap memory）：当前程序运行时已经使用的堆内存，这个内存小于Committed heap memory。
- 非堆内存：在JVM中堆之外的内存称为非堆内存（Non-heap memory），JVM自身运行时所需要的内存区域，非堆内存有多个内存池，通常包括以下3个部分：

  - 代码缓存区（Code Cache）主要用于存放JIT所编译的代码。默认限制240M，可以通过JVM启动参数-XX:InitialCodeCacheSize -XX:ReservedCodeCacheSize进行设置。
  - 类指针压缩空间（Compressed Class Space）存储类指针的元数据，默认限制1024M，通过JVM启动参数-XX:CompressedClassSpaceSize进行设置。
  - 元空间（Metaspace）用于存放元数据，通过JVM启动参数-XX:MetaspaceSize -XX:MaxMetaspaceSize进行设置。
- 最大非堆内存（Maximum Non Heap Memory）：系统可以分配给程序的最大非堆内存。其值为代码缓存区（Code Cache），类指针压缩空间（Compressed Class Space），元空间（Metaspace）最大值之和。
- 分配的非堆内存（Committed Non Heap Memory）：为保证程序运行的系统非堆内存总量，Committed Non Heap Memory在程序运行期间根据使用的非堆内存情况，会在初始非堆内存和最大非堆内存之间波动变化。
- 使用非堆内存（Used Non Heap memory）：当前程序运行时已经使用的非堆内存，这个值小于分配的非堆内存（Committed Non Heap Memory）。

Hadoop

一个分布式系统框架。用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用了集群的高速运算和存储。 Hadoop能够对大量数据以可靠的、高效的、可伸缩的方式进行分布式处理。Hadoop是可靠的，因为它假设计算单元和存储会失败，因此维护多个工作数据副本，确保对失败节点重新分布处理；Hadoop是高效的，因为它以并行的方式工作，从而加快处理速度；Hadoop是可伸缩的，能够处理PB级数据。 Hadoop主要由HDFS、MapReduce、HBase和Hive组成。

角色

角色是服务的组成要素，每个服务由一个或多个角色组成。服务通过角色安装到主机（即服务器）上，保证服务正常运行。

集群

将多个服务器集中起来使它们能够像一台服务器一样提供服务的计算机技术。采用集群通常是为了提高系统的稳定性、可靠性、数据处理能力或服务能力。例如，可以减少单点故障、共享存储资源、负荷分担或提高系统性能等。

实例

当一个服务的角色安装到主机上，即形成一个实例。每个服务有各自对应的角色实例。

元数据

元数据（Metadata），又称中介数据、中继数据，为描述数据的数据（data about data），主要是描述数据属性（property）的信息，用来支持如指示存储位置、历史数据、资源查找、文件记录等功能。

2 入门

2.1 如何使用 MRS

MRS是一个在云上部署和管理Hadoop系统的服务，一键即可部署Hadoop集群。MRS提供租户完全可控的企业级大数据集群云服务，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件。

MRS使用简单，通过使用在集群中连接在一起的多台计算机，您可以运行各种任务，处理或者存储（PB级）巨量数据。MRS的基本使用流程如下：

1. 上传程序和数据文件到对象存储服务（OBS）中，用户需要先将本地的程序和数据文件上传至OBS中。
2. **创建自定义集群**，用户可以指定集群类型用于离线数据分析和流处理任务，指定集群中预置的弹性云服务器实例规格、实例数量、数据盘类型（普通IO、高 IO、超高 IO）、要安装的组件（Hadoop、Spark、HBase、Hive、Kafka、Storm等）。用户可以使用**引导操作**在集群启动前（或后）在指定的节点上执行脚本，安装其他第三方软件或修改集群运行环境等自定义操作。
3. **管理作业**，MRS为用户提供程序执行平台，程序由用户自身开发，MRS负责程序的提交、执行和监控。
4. **管理集群**，MRS为用户提供企业级的大数据集群的统一管理平台，帮助用户快速掌握服务及主机的健康状态，通过图形化的指标监控及定制及时的获取系统的关键信息，根据实际业务的性能需求修改服务属性的配置，对集群、服务、角色实例等实现一键启停等操作。
5. **删除集群**，如果作业执行结束后不需要集群，可以删除MRS集群。

2.2 创建集群

使用MRS的首要操作就是创建集群，本章节为您介绍如何在MRS管理控制台创建一个新的集群。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

> **说明**
>
> 创建集群时需要注意配额提醒。当资源配额不足时，建议按照提示申请足够的资源，再创建集群。

步骤3 在创建集群页面，选择“自定义创建”页签。

步骤4 配置集群软件信息。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：默认最新版本即可。
- 集群类型：默认选择“分析集群”即可。
- 组件选择：分析集群勾选Spark2x、HBase和Hive等组件。流式集群勾选Kafka和Storm等组件。混合集群可同时勾选分析集群流式集群的组件。
- 元数据：默认即可。

> **说明**
>
> 针对MRS 3.x之前版本，分析集群勾选Spark、HBase和Hive等组件。

步骤5 单击“下一步”。

- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 安全组：选择“自动创建”。
- 弹性公网IP：选择“暂不绑定”。
- 实例规格：Master和Core节点都选择“通用计算型S3->8核 16GB(s3.2xlarge.2)”。
- 系统盘：存储类型选择“普通IO”，存储空间默认即可。
- 数据盘：存储类型选择“普通IO”，存储空间默认即可，数据盘数量默认即可。
- 实例数量：Master节点数量默认为2，Core节点数量配置为3。

步骤6 单击“下一步”进入高级配置页签，配置参数，其他参数保持默认。

- Kerberos认证：
  - Kerberos认证：关闭Kerberos认证。
  - 用户名：Manager管理员用户，目前默认为**admin**用户。
  - 密码：Manager管理员用户的密码。
- 密钥对：从下拉框中选择密钥对，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。
- 通信安全授权：勾选确认授权。

步骤7 单击“立即申请”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤8 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

----结束

2.3 上传示例数据和程序

用户通过“文件管理”页面可以在分析集群进行文件夹创建、删除，文件导入、导出、删除操作。

背景信息

MRS集群处理的数据源来源于OBS或HDFS，OBS为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用。

导入数据

MRS目前只支持将OBS上的数据导入至HDFS中。上传文件速率会随着文件大小的增大而变慢，适合数据量小的场景下使用。

支持导入文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。

   “bd_app1”目录仅为示例，可以是界面上的任何目录，也可以通过“新建”创建新的文件夹。

   新建文件夹时需要满足以下要求：
   - 文件夹名称小于等于255字符，全路径长度小于等于1023字符。
   - 不允许为空。
   - 不能包含：/:*?"<>|\;&,'`!{}[]$%+特殊字符。
   - 不能以“.”开头或结尾。
   - 开头和末尾的空格会被忽略。
6. 单击“导入数据”，正确配置HDFS和OBS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。
   - OBS路径
     - 必须以“obs://”开头。
     - 不支持导入KMS加密的文件或程序。
     - 不支持导入空的文件夹。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。

- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- OBS全路径长度小于等于1023字符。

– HDFS路径

- 默认以“/user”开头。
- 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- HDFS全路径长度小于等于1023字符。

7. 单击“确定”。

   文件上传进度可在“文件操作记录”中查看。MRS将数据导入操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

导出数据

数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。

支持导出文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群基本信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。
6. 单击“导出数据”，配置OBS和HDFS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。

   – OBS路径

   - 必须以“obs://”开头。
   - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
   - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
   - OBS全路径长度小于等于1023字符。

   – HDFS路径

   - 默认以“/user”开头。
   - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
   - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
   - HDFS全路径长度小于等于1023字符。

说明

当导出文件夹到OBS系统时，在OBS路径下，将增加一个标签文件，文件命名为“folder name_$folder$”。请确保导出的文件夹为非空文件夹，如果导出的文件夹为空文件夹，OBS无法显示该文件夹，仅生成一个命名为“folder name_$folder$”的文件。

7. 单击“确定”。

文件上传进度可在“文件操作记录”中查看。MRS将数据导出操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

2.4 添加作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。

本章节以MapReduce作业为例指导您在MRS集群页面如何提交一个新的作业。MapReduce作业用于提交jar程序快速并行处理大量数据，是一种分布式数据处理模式和执行环境。

若在集群详情页面不支持“作业管理”和“文件管理”功能，请通过后台功能来提交作业。

用户创建作业前需要将本地数据上传至OBS系统用于计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，并使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

说明

IAM用户名存在空格时（如admin 01），不支持添加作业。

步骤6 “作业类型”选择“MapReduce”，并配置其他作业信息。

- 若集群为MRS 2.1.0及之后版本，请参考**表2-1**配置MapReduce作业信息。
- 若集群为MRS 2.1.0之前版本，请参考**表2-3**配置MapReduce作业信息。

表 2-1 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。
执行程序参数	可选参数，程序执行的关键参数。多个参数间使用空格隔开。 配置方法：*程序类名 数据输入路径 数据输出路径* ● 程序类名：由用户程序内的函数指定，MRS只负责参数的传入。 ● 数据输入路径：通过单击“HDFS”或者“OBS”选择或者直接手动输入正确路径。 ● 数据输出路径：输出路径请手动输入一个不存在的目录。最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表2-2**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 2-2 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

表 2-3 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。示例：s3a://wordcount/program/xxx.jar – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript需要以“.sql”结尾，MapReduce和Spark需要以“.jar”结尾。sql、jar不区分大小写。
执行程序参数	程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 配置方法：包名.类名。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **说明** 用户输入带有敏感信息（如登录密码）的参数时，可通过在参数名前添加“@”的方式，为该参数值加密，以防止敏感信息被明文形式持久化。在MRS管理控制台查看作业信息时，敏感信息显示为“*”。 例如：username=admin @password=admin_123
数据输入路径	数据输入地址。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

参数	参数说明
输出路径	数据输出地址。 **说明** • 配置此参数时，单击“OBS”或“HDFS”，并选择文件目录，或者手动输入文件目录，然后单击“确定”。 • 若添加hadoop-mapreduce-examples-x.x.x.jar样例程序或和hadoop-mapreduce-examples-x.x.x.jar类似的程序，请手动输入一个不存在的目录。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 • OBS：以“s3a://”开头。（仅MRS 1.8.10及之前版本支持。） • HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。
日志路径	作业日志存储地址，该日志信息记录作业运行状态。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 • OBS：以“s3a://”开头。 • HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录任意一个Master节点，具体请参考**登录集群节点**。

步骤2 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤3 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证，则无需执行该步骤。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 执行如下命令拷贝OBS文件系统中的程序到集群的Master节点。

hadoop fs -Dfs.obs.access.key=AK -Dfs.obs.secret.key=SK -copyToLocal source_path.jar target_path.jar

例如：**hadoop fs -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX -copyToLocal "obs://mrs-word/program/hadoop-mapreduce-examples-XXX.jar" "/home/omm/hadoop-mapreduce-examples-XXX.jar"**

AK/SK可登录OBS控制台，请在集群控制台页面右上角的用户名下拉框中选择“我的凭证 > 访问密钥”页面获取。

步骤5 执行如下命令提交wordcount作业，如需从OBS读取或向OBS输出数据，需要增加AK/SK参数。

source /opt/Bigdata/client/bigdata_env;hadoop jar execute_jar wordcount input_path output_path

例如：**source /opt/Bigdata/client/bigdata_env;hadoop jar /home/omm/hadoop-mapreduce-examples-XXX.jar wordcount -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX "obs://mrs-word/input/*" "obs://mrs-word/output/"**

input_path为OBS上存放作业输入文件的路径。output_path为OBS上存放作业输出文件地址，请设置为一个不存在的目录。

----结束

2.5 快速使用 Kerberos 认证集群

本章节提供从零开始使用安全集群并执行MapReduce程序、Spark程序和Hive程序的操作指导。

MRS 3.x版本Presto组件暂不支持开启Kerberos认证。

本指导的基本内容如下所示：

1. **创建安全集群并登录其Manager**
2. **创建角色和用户**
3. **执行MapReduce程序**
4. **执行Spark程序**
5. **执行Hive程序**

创建安全集群并登录其 Manager

步骤1 创建安全集群，请参见**创建自定义集群**页面，开启“Kerberos认证”参数开关，并配置“密码”、“确认密码”参数。该密码用于登录Manager，请妥善保管。

步骤2 登录MRS管理控制台页面。

步骤3 单击“集群列表”，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤4 单击“集群管理页面”后的“前往Manager”，打开Manager页面。

- 若用户创建集群时已经绑定弹性公网IP。
  a. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

  说明

  - 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
  - 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问Manager服务。

b. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

- 若用户创建集群时暂未绑定弹性公网IP。

a. 在弹性公网IP下拉框中选择可用的弹性公网IP或单击“管理弹性公网IP”创建弹性公网IP。

b. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请点击“管理安全组规则”。

说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问Manager服务。

c. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

步骤5 单击“确定”，进入Manager登录页面，如需给其他用户开通访问Manager的权限，请参见**访问MRS Manager（MRS 2.x及之前版本）**章节，添加对应用户访问公网的IP地址为可信范围。

步骤6 输入创建集群时默认的用户名“admin”及设置的密码，单击“登录”进入Manager页面。

----结束

当集群版本为**MRS 1.8.0**以前版本时，执行如下步骤：

步骤1 创建安全集群，开启“Kerberos认证”参数开关，并配置“密码”、“确认密码”参数。该密码用于登录Manager，请妥善保管。

步骤2 在MRS管理控制台，单击“集群列表”。

说明

- 访问支持Kerberos认证的Manager方法请参考**步骤2**-**步骤7**，或者请参考。
- 分析集群和流式集群访问支持Kerberos认证的Manager方法相同。

步骤3 在 “现有集群” 列表中，单击已创建的安全集群名称。

记录集群的“可用区”、“虚拟私有云”、“集群控制台地址”，以及Master节点的“默认安全组”。

步骤4 在弹性云服务器管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果弹性云服务器的安全组和Master节点的“默认安全组”不同，用户可以选择以下任一种方法修改配置：

- 将弹性云服务器的安全组修改为Master节点的默认安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组添加两条安全组规则使弹性云服务器可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤5 在虚拟私有云管理控制台，申请一个弹性IP地址，并与弹性云服务器绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤6 登录弹性云服务器。

登录弹性云服务器需要Windows系统的账号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤7 在Windows的远程桌面中，打开浏览器访问Manager。

例如Windows 2012操作系统可以使用Internet Explorer 11。

Manager访问地址形式为**https://*集群控制台地址*:28443/web**。**集群控制台地址**为步骤3中记录的“集群控制台地址”，访问时需要输入MRS集群的用户名和密码，例如“admin”用户，密码为步骤1创建集群时开启“Kerberos认证”参数开关时输入的密码。

说明

- 如果使用其他MRS集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

----结束

创建角色和用户

开启Kerberos认证的集群，必须通过以下步骤创建一个用户并分配相应权限来允许用户执行程序。

步骤1 在Manager界面选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，详情请参见创建角色。

填写如下信息：

- 填写角色的名称，例如mrrole。
- 在“配置资源权限”选择待操作的集群，然后选择“Yarn > 调度队列 > root”，勾选“权限”列中的“提交”和“管理”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。
- 选择“HBase > HBase Scope”，勾选global的“权限”列的“创建”、“读”、“写”和“执行”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。

- 选择“HDFS > 文件系统 > hdfs://hacluster/”，勾选“权限”列的“读”、“写”和“执行”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。
- 选择“Hive > Hive读写权限”，勾选“权限”列的“查询”、“删除”、“插入”和“建表”，单击“确定”，完成角色的创建。

步骤3 选择“系统 > 权限 > 用户组 > 添加用户组”，为样例工程创建一个用户组，例如mrgroup，详情请参见**创建用户组**。

步骤4 选择“系统 > 权限 > 用户 > 添加用户”，为样例工程创建一个用户，详情请参见**创建用户**。

- 填写用户名，例如**test**，当需要执行Hive程序时，请设置用户名为“hiveuser”。
- 用户类型为“人机”用户。
- 输入密码（特别注意该密码在后面运行程序时要用到）。
- 加入用户组mrgroup和supergroup。
- 设置其“主组”为supergroup，并绑定角色mrrole取得权限。

  单击“确定”完成用户创建。

步骤5 选择“系统 > 权限 > 用户”，选择新建用户**test**，选择“更多 > 下载认证凭据”，保存后解压得到用户的keytab文件与krb5.conf文件。

----结束

执行 MapReduce 程序

本小节提供执行MapReduce程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如mapreduce-examples-1.0.jar、input_data1.txt和input_data2.txt。

操作步骤

步骤1 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

步骤2 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

步骤3 使用上传工具（比如：WinScp）将mapreduce-examples-1.0.jar、input_data1.txt和input_data2.txt复制到test目录下，将“创建角色和用户”中的步骤**步骤5**获得的keytab文件和krb5.conf文件复制到conf目录。

步骤4 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回（首次登录需按照系统提示修改密码），则完成了用户的kerberos认证。

步骤5 执行如下命令将数据导入到HDFS中：

```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

步骤6 执行如下命令运行程序：

```
yarn jar mapreduce-examples-1.0.jar com.xx.bigdata.mapreduce.examples.FemaleInfoCollector /tmp/
input /tmp/mapreduce_output
```

其中：

/tmp/input指HDFS文件系统中input的路径。

/tmp/mapreduce_output指HDFS文件系统中output的路径，该目录必须不存在，否则会报错。

步骤7 程序运行成功后，执行 hdfs dfs -ls /tmp/mapreduce_output会显示如下：

图 2-1 查看程序运行结果

```
[root@node-master1-SsjQd test]# hdfs dfs -ls /tmp/mapreduce_output
Found 2 items
-rw-r--r--+  2 test hadoop          0 2018-08-20 20:53 /tmp/mapreduce_output/_
SUCCESS
-rw-r--r--+  2 test hadoop         23 2018-08-20 20:53 /tmp/mapreduce_output/p
art-r-00000
[root@node-master1-SsjQd test]#
```

----结束

执行 Spark 程序

本小节提供执行Spark程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如FemaleInfoCollection.jar、input_data1.txt和input_data2.txt。

操作步骤

步骤1 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

步骤2 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

步骤3 使用上传工具（比如：WinScp）将样FemaleInfoCollection.jar、input_data1.txt和input_data2.txt复制到test目录下，将“创建角色和用户”中的步骤**步骤5**获得的keytab文件和krb5.conf文件复制到conf目录。

步骤4 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回，则完成了用户的kerberos认证。

步骤5 执行如下命令将数据导入到HDFS中：

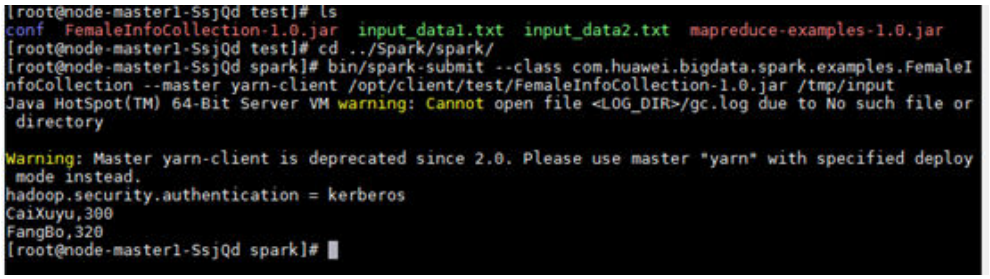
```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

步骤6 执行如下命令运行程序：

```
cd /opt/Bigdata/client/Spark/spark
bin/spark-submit --class com.xxxx.bigdata.spark.examples.FemaleInfoCollection --master yarn-client /opt/Bigdata/client/test/FemaleInfoCollection-1.0.jar /tmp/input
```

步骤7 程序运行成功后，会显示如下：

图 2-2 程序运行结果



----结束

执行 Hive 程序

本小节提供执行Hive程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如hive-examples-1.0.jar、input_data1.txt和input_data2.txt。

操作步骤

步骤1 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

步骤2 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

步骤3 使用上传工具（比如：WinScp）将样FemaleInfoCollection.jar、input_data1.txt和input_data2.txt复制到test目录下，将“创建角色和用户”中的步骤**步骤5**获得的keytab文件和krb5.conf文件复制到conf目录。

步骤4 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回，则完成了用户的kerberos认证。

步骤5 执行如下命令运行程序：

```
chmod +x /opt/hive_examples -R   cd /opt/hive_examples   java -cp .:hive-examples-1.0.jar:/opt/
hive_examples/conf:/opt/Bigdata/client/Hive/Beeline/lib/*:/opt/Bigdata/client/HDFS/hadoop/lib/*
com.xxx.bigdata.hive.example.ExampleMain
```

步骤6 程序运行成功后，会显示如下：

图 2-3 程序运行的结果

```
[root@node-master1-iYpxp hive_examples]# java -cp .:hive-examples-mrs-1.7.0.jar:
/opt/hive_examples/conf:/opt/client/Hive/Beeline/lib/*:/opt/client/HDFS/hadoop/l
ib/* com.huawei.bigdata.hive.example.ExampleMain
log4j:WARN No appenders could be found for logger (com.huawei.bigdata.security.L
oginUtil).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Create table success!
_c0
0
Delete table success!
[root@node-master1-iYpxp hive_examples]#
```

----结束

2.6 删除集群

如果作业执行结束后不需要集群， 可以删除MRS集群。

背景信息

一般在数据完成分析和存储后或集群异常无法提供服务时才执行集群删除操作。当MRS集群部署失败时，集群会被自动删除。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”。

步骤3 在需要删除的集群对应的“操作”列中，单击“删除”。

集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。

----结束

3 准备用户

3.1 创建 MRS 操作用户

如果您需要对您所拥有的MapReduce服务（MapReduce Service）进行精细的权限管理，您可以使用统一身份认证服务（Identity and Access Management，简称IAM），通过IAM，您可以：

- 根据企业的业务组织，在您的云账号中，给企业中不同职能部门的员工创建IAM用户，让员工拥有唯一安全凭证，并使用MRS资源。
- 根据企业用户的职能，设置不同的访问权限，以达到用户之间的权限隔离。
- 将MRS资源委托给更专业、高效的其他云账号或者云服务，这些账号或者云服务可以根据权限进行代运维。

如果云账号已经能满足您的要求，不需要创建独立的IAM用户，您可以跳过本章节，不影响您使用MRS服务的其它功能。

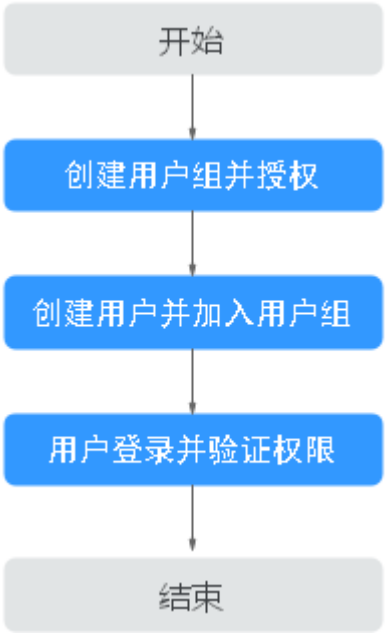
本章节为您介绍对用户授权的方法，操作流程如**图3-1**所示。

前提条件

给用户组授权之前，请您了解用户组可以添加的MRS权限，并结合实际需求进行选择。若您需要对除MRS之外的其它服务授权，IAM支持服务的所有策略请参见**权限策略**。

示例流程

图 3-1 给用户授权 MRS 权限流程



1. **创建用户组并授权**
   在IAM控制台创建用户组，并授予MRS服务对应权限。
2. **创建用户并加入用户组**
   在IAM控制台创建用户，并将其加入**1.创建用户组并授权**中创建的用户组。
3. 用户登录并验证权限
   新创建的用户登录控制台，切换至授权区域，验证权限：
   - 在“服务列表”中选择MRS服务，进入MRS主界面，单击右上角“创建集群”，尝试创建MRS集群，如果无法创建MRS集群（假设当前权限仅包含MRS ReadOnlyAccess），表示“MRS ReadOnlyAccess”已生效。
   - 在“服务列表”中选择除MRS服务外（假设当前策略仅包含MRS ReadOnlyAccess）的任一服务，若提示权限不足，表示“MRS ReadOnlyAccess”已生效。

MRS 权限说明

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

MRS部署时通过物理区域划分，为项目级服务。授权时，“作用范围”需要选择“区域级项目”，然后在指定区域对应的项目中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问MRS时，需要先切换至授权区域。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于各服务之间存在业

务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。

- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对MRS服务，管理员能够控制IAM用户仅能对集群进行指定的管理操作。如不允许某用户组删除集群，仅允许操作MRS集群基本操作，如创建集群、查询集群列表等。多数细粒度策略以API接口为粒度进行权限拆分。

如表3-1所示，包括了MRS的所有系统策略。

表 3-1 MRS 系统策略

策略名称	描述	策略类别
MRS FullAccess	MRS管理员权限，拥有该权限的用户可以拥有MRS所有权限。	细粒度策略
MRS CommonOperations	MRS服务普通用户权限，拥有该权限的用户可以拥有MRS服务使用权限，无新增、删除资源权限。	细粒度策略
MRS ReadOnlyAccess	MRS服务只读权限，拥有该权限的用户仅能查看MRS的资源。	细粒度策略
MRS Administrator	操作权限： • 对MRS服务的所有执行权限。 • 拥有该权限的用户必须同时拥有Tenant Guest和Server Administrator权限。	RBAC策略

表3-2列出了MRS常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 3-2 常用操作与系统策略的授权关系

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
创建集群	√	x	x	√
调整集群	√	x	x	√
升级节点规格	√	x	x	√
删除集群	√	x	x	√
查询集群详情	√	√	√	√

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
查询集群列表	√	√	√	√
设置弹性伸缩策略	√	x	x	√
查询主机列表	√	√	√	√
查询操作日志	√	√	√	√
创建并执行作业	√	√	x	√
停止作业	√	√	x	√
删除单个作业	√	√	x	√
批量删除作业	√	√	x	√
查询作业详情	√	√	√	√
查询作业列表	√	√	√	√
新建文件夹	√	√	x	√
删除文件	√	√	x	√
查询文件列表	√	√	√	√
批量操作集群标签	√	√	x	√
创建单个集群标签	√	√	x	√
删除单个集群标签	√	√	x	√
按照标签查询资源列表	√	√	√	√
查询集群标签	√	√	√	√
访问Manager页面	√	√	x	√

操作	MRS FullAccess	MRS CommonOperations	MRS ReadOnlyAccess	MRS Administrator
查询补丁列表	√	√	√	√
安装补丁	√	√	x	√
卸载补丁	√	√	x	√
运维通道授权	√	√	x	√
运维通道日志共享	√	√	x	√
查询告警列表	√	√	√	√
订阅告警消息提醒	√	√	x	√
提交SQL语句	√	√	x	√
查询SQL结果	√	√	x	√
取消SQL执行任务	√	√	x	√

3.2 创建 MRS 自定义策略

如果系统预置的MRS权限，不满足您的授权要求，可以创建自定义策略。自定义策略中可以添加的授权项（Action）请参考MRS服务API参考中“权限策略和授权项 > 策略及授权项说明”章节。

目前支持以下两种方式创建自定义策略：

- 可视化视图创建自定义策略：无需了解策略语法，按可视化视图导航栏选择云服务、操作、资源、条件等策略内容，可自动生成策略。
- JSON视图创建自定义策略：可以在选择策略模板后，根据具体需求编辑策略内容；也可以直接在编辑框内编写JSON格式的策略内容。

本章为您介绍常用的MRS自定义策略样例。

MRS 自定义策略样例

- 示例1：授权用户仅有创建MRS集群的权限

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
```

```
            "mrs:cluster:create",
            "ecs:*:*",
            "bms:*:*",
            "evs:*:*",
            "vpc:*:*",
            "smn:*:*"
        ]
    }
  ]
}
```

- 示例2：授权用户调整MRS集群

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "mrs:cluster:resize"
      ]
    }
  ]
}
```

- 示例3：授权用户创建集群、创建并执行作业、删除单个作业，但不允许用户删除集群的权限

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "mrs:cluster:create",
        "mrs:job:submit",
        "mrs:job:delete"
      ]
    },
    {
      "Effect": "Deny",
      "Action": [
        "mrs:cluster:delete"
      ]
    }
  ]
}
```

- 示例4：授权用户最小权限，创建ECS规格的集群

  说明

  - 创建集群时如果使用密钥对，增加权限：ecs:serverKeypairs:get和ecs:serverKeypairs:list
  - 创集群时使用数据盘加密，增加权限：kms:cmk:list
  - 创建集群时开启告警功能，增加权限：mrs:alarm:subscribe
  - 创建集群时使用外置数据源，增加权限：rds:instance:list

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "mrs:cluster:create"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "ecs:cloudServers:updateMetadata",
```

```
            "ecs:cloudServerFlavors:get",
            "ecs:cloudServerQuotas:get",
            "ecs:servers:list",
            "ecs:servers:get",
            "ecs:cloudServers:delete",
            "ecs:cloudServers:list",
            "ecs:serverInterfaces:get",
            "ecs:serverGroups:manage",
            "ecs:servers:setMetadata",
            "ecs:cloudServers:get",
            "ecs:cloudServers:create"
        ]
    },
    {
        "Effect": "Allow",
        "Action": [
            "vpc:securityGroups:create",
            "vpc:securityGroupRules:delete",
            "vpc:vpcs:create",
            "vpc:ports:create",
            "vpc:securityGroups:get",
            "vpc:subnets:create",
            "vpc:privateIps:delete",
            "vpc:quotas:list",
            "vpc:networks:get",
            "vpc:publicIps:list",
            "vpc:securityGroups:delete",
            "vpc:securityGroupRules:create",
            "vpc:privateIps:create",
            "vpc:ports:get",
            "vpc:ports:delete",
            "vpc:publicIps:update",
            "vpc:subnets:get",
            "vpc:publicIps:get",
            "vpc:ports:update",
            "vpc:vpcs:list"
        ]
    },
    {
        "Effect": "Allow",
        "Action": [
            "evs:quotas:get",
            "evs:types:get"
        ]
    },
    {
        "Effect": "Allow",
        "Action": [
            "bms:serverFlavors:get"
        ]
    }
  ]
}
```

- 示例5：授权用户最小权限，创建BMS规格的集群

  说明

  - 创建集群时如果使用密钥对，增加权限：ecs:serverKeypairs:get和ecs:serverKeypairs:list
  - 创集群时使用数据盘加密，增加权限：kms:cmk:list
  - 创建集群时开启告警功能，增加权限：mrs:alarm:subscribe
  - 创建集群时使用外置数据源，增加权限：rds:instance:list

```
{
  "Version": "1.1",
  "Statement": [
    {
        "Effect": "Allow",
```

```
            "Action": [
                "mrs:cluster:create"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "ecs:servers:list",
                "ecs:servers:get",
                "ecs:cloudServers:delete",
                "ecs:serverInterfaces:get",
                "ecs:serverGroups:manage",
                "ecs:servers:setMetadata",
                "ecs:cloudServers:create",
                "ecs:cloudServerFlavors:get",
                "ecs:cloudServerQuotas:get"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "vpc:securityGroups:create",
                "vpc:securityGroupRules:delete",
                "vpc:vpcs:create",
                "vpc:ports:create",
                "vpc:securityGroups:get",
                "vpc:subnets:create",
                "vpc:privateIps:delete",
                "vpc:quotas:list",
                "vpc:networks:get",
                "vpc:publicIps:list",
                "vpc:securityGroups:delete",
                "vpc:securityGroupRules:create",
                "vpc:privateIps:create",
                "vpc:ports:get",
                "vpc:ports:delete",
                "vpc:publicIps:update",
                "vpc:subnets:get",
                "vpc:publicIps:get",
                "vpc:ports:update",
                "vpc:vpcs:list"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "evs:quotas:get",
                "evs:types:get"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "bms:servers:get",
                "bms:servers:list",
                "bms:serverQuotas:get",
                "bms:servers:updateMetadata",
                "bms:serverFlavors:get"
            ]
        }
    ]
}
```

- 示例6：授权用户最小权限，创建ECS和BMS混合集群

说明

- 创建集群时如果使用密钥对，增加权限：ecs:serverKeypairs:get和ecs:serverKeypairs:list
- 创集群时使用数据盘加密，增加权限：kms:cmk:list
- 创建集群时开启告警功能，增加权限：mrs:alarm:subscribe
- 创建集群时使用外置数据源，增加权限：rds:instance:list

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "mrs:cluster:create"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "ecs:cloudServers:updateMetadata",
        "ecs:cloudServerFlavors:get",
        "ecs:cloudServerQuotas:get",
        "ecs:servers:list",
        "ecs:servers:get",
        "ecs:cloudServers:delete",
        "ecs:cloudServers:list",
        "ecs:serverInterfaces:get",
        "ecs:serverGroups:manage",
        "ecs:servers:setMetadata",
        "ecs:cloudServers:get",
        "ecs:cloudServers:create"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "vpc:securityGroups:create",
        "vpc:securityGroupRules:delete",
        "vpc:vpcs:create",
        "vpc:ports:create",
        "vpc:securityGroups:get",
        "vpc:subnets:create",
        "vpc:privateIps:delete",
        "vpc:quotas:list",
        "vpc:networks:get",
        "vpc:publicIps:list",
        "vpc:securityGroups:delete",
        "vpc:securityGroupRules:create",
        "vpc:privateIps:create",
        "vpc:ports:get",
        "vpc:ports:delete",
        "vpc:publicIps:update",
        "vpc:subnets:get",
        "vpc:publicIps:get",
        "vpc:ports:update",
        "vpc:vpcs:list"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "evs:quotas:get",
        "evs:types:get"
      ]
    },
    {
      "Effect": "Allow",
```

```
            "Action": [
                "bms:servers:get",
                "bms:servers:list",
                "bms:serverQuotas:get",
                "bms:servers:updateMetadata",
                "bms:serverFlavors:get"
            ]
        }
    ]
}
```

3.3 IAM 用户同步 MRS 说明

IAM用户同步是指将绑定MRS相关策略的IAM用户同步至MRS系统中，创建同用户名、不同密码的账号，用于集群管理。同步之后，用户可以使用IAM用户名（密码需要Manager的管理员admin重置后方可使用）登录Manager管理集群。也可以在开启Kerberos认证的集群中，通过界面方式提交作业。

IAM用户权限策略及同步MRS后权限对比请参考**表3-3**，Manager对应默认权限说明请参考**MRS集群中的用户与权限**。

表 3-3 IAM 权限策略与 MRS 权限同步映射

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
细粒度	MRS ReadOnlyAccess	Manager_viewer	否	否
	MRS CommonOperations	• Manager_viewer • default • launcher-job	否	是

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
	MRS FullAccess	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	是	是
RBAC	MRS Administrator	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	否	是

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
	Server Administrator、Tenant Guest和MRS Administrator	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	是	是
	Tenant Administrator	• Manager_administrator • Manager_auditor • Manager_operator • Manager_tenant • Manager_viewer • System_administrator • default • launcher-job	是	是

策略类别	IAM策略	同步后用户在MRS对应默认权限	是否有权限执行同步操作	是否有权限提交作业
自定义	Custom policy（自定义策略）	• Manager_viewer • default • launcher-job	• 自定义策略以RBAC策略为模板则参考RBAC策略。 • 自定义策略以细粒度策略为模板则参考细粒度策略，建议使用细粒度策略。	是

说明

为了更方便进行用户权限管理，请尽可能使用细粒度策略，减少RBAC策略的使用，细粒度策略判断action时以deny优先原则。

- 只有具有Tenant Administrator或同时具有Server Administrator、Tenant Guest、MRS Administrator角色才在MRS集群中拥有同步IAM用户的权限。
- 只要拥有**action:mrs:cluster:syncUser**策略就在MRS集群中拥有同步IAM用户的权限。

操作步骤

步骤1 创建用户并授权使用MRS服务，具体请参考**创建MRS操作用户**。

步骤2 登录MRS控制台并创建集群，具体请参考**创建自定义集群**。

步骤3 在左侧导航栏中选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤4 在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

步骤5 同步请求下发后，返回MRS控制台在左侧导航栏中选择“操作日志”页面查看同步是否成功，日志相关说明请参考**查看MRS服务操作日志**。

步骤6 同步成功后，即可使用IAM同步用户进行后续操作。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。
- 单击“IAM用户同步”右侧的“同步”后，集群详情页面会出现短时间空白，这是由于正在进行用户数据同步中，请耐心等待，数据同步完成后，页面将会正常显示。

- 安全集群提交作业：安全集群中用户可通过界面“作业管理”功能提交作业，具体请参考**运行MapReduce作业**。
- 集群详情页面页签显示完整（包含“组件管理”，“租户管理”和“备份恢复”）。
- 登录Manager页面。
  a. 使用admin账号登录Manager，具体请参考**访问集群Manager**。
  b. 初始化IAM同步用户密码，具体请参考**初始化系统用户密码**。
  c. 修改用户所在用户组绑定的角色，精确控制Manager下用户权限，具体请参考**相关任务**修改用户组绑定的角色，如需创建修改角色请参考**创建角色**。用户所在用户组绑定的组件角色修改后，权限生效需要一定时间，请耐心等待。
  d. 使用IAM同步用户及**步骤6.b**初始化后的密码登录Manager。

  说明

  当IAM用户权限发生变化时，需要执行**步骤4**进行二次同步。对于系统用户，二次同步后用户的权限为IAM系统策略定义的权限和用户在Manager自行添加角色的权限的并集。对于自定义用户，二次同步后用户的权限以Manager配置的权限为准。
  - 系统用户：如果IAM用户所在用户组全部都绑定系统策略（RABC策略和细粒度策略均属于系统策略），则该用户为系统用户。
  - 自定义用户：如果IAM用户所在用户组只要有绑定任何自定义策略，则该用户为自定义用户。

----结束

4 配置集群

4.1 创建方式简介

本节介绍创建MRS服务的方式。

- **快速创建Hadoop分析集群**：快速创建Hadoop分析集群为您提高了配置效率，可以在几分钟之内快速创建Hadoop集群，更加方便快捷的进行海量数据分析与查询。
- **快速创建HBase查询集群**：快速创建HBase查询集群为您提高了配置效率，可以在几分钟之内快速创建HBase集群，更加方便快捷的进行海量数据存储以及分布式计算。
- **快速创建Kafka流式集群**：快速创建Kafka流式集群为您提高了配置效率，可以在几分钟之内快速创建Kafka集群，更加方便快捷的进行流式数据采集，实时数据处理存储等。
- **快速创建ClickHouse集群**：快速创建一个ClickHouse集群，ClickHouse是一个用于联机分析的列式数据库管理系统，具有极致压缩率和极速查询性能。
- **快速创建实时分析集群**：快速创建一个实时分析集群为您提高了配置效率，可以在几分钟之内快速创建实时分析集群，更加方便快捷的进行海量的数据采集、数据的实时分析和查询。
- **创建自定义集群**：自定义创建可以灵活地选择配置项，针对不同的应用场景，可以选择不同规格的弹性云服务器，全方位贴合您的业务诉求。

4.2 快速集群

4.2.1 快速创建 Hadoop 分析集群

本章节为您介绍如何快速创建一个Hadoop分析集群，Hadoop完全使用开源Hadoop生态，采用YARN管理集群资源，提供Hive、Spark离线大规模分布式数据存储和计算，SparkStreaming、Flink流式数据计算，Presto交互式查询，Tez有向无环图的分布式计算框架等Hadoop生态圈的组件，进行海量数据分析与查询。

Hadoop分析集群包含如下组件：

- MRS 1.8.9版本：Hadoop 2.8.3, Hive 1.2.1, Presto 0.215, Spark 2.2.1 and Flink 1.7.0。
- MRS 2.0.1版本：Hadoop 3.1.1, Hive 3.1.0, Spark 2.3.2 and Tez 0.9.1。
- MRS 3.1.0-LTS.1版本：Hadoop 3.1.1, Hive 3.1.0, Spark2x 2.4.5, Flink 1.12.0, ZooKeeper 3.5.6, Ranger 2.0.0 and Tez 0.9.2。
- MRS 3.1.2-LTS.3版本：Hadoop 3.1.1, Hive 3.1.0, Spark2x 3.1.1, Flink 1.12.2, ZooKeeper 3.6.3, Ranger 2.0.0 and Tez 0.9.2。
- MRS 3.2.0-LTS.1版本：Hadoop 3.3.1, Hive 3.1.0, Spark2x 3.1.1, Flink 1.15.0, ZooKeeper 3.6.3, Ranger 2.0.0 and Tez 0.9.2。

快速创建 Hadoop 分析集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“Hadoop分析集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.2.0-LTS.1及之后版本支持该参数。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。MRS 3.2.0-LTS.1及之后版本支持root用户。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。
- 密钥对：从下拉框中选择密钥对用于登录ECS机器，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。MRS 3.1.2-LTS.3及之前版本支持该参数。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.2 快速创建 HBase 查询集群

本章节为您介绍如何快速创建一个HBase查询集群，HBase集群使用Hadoop和HBase组件提供一个稳定可靠，性能卓越、可伸缩、面向列的分布式云存储系统，适用于海量数据存储以及分布式计算的场景，用户可以利用HBase搭建起TB至PB级数据规模的存储系统，对数据轻松进行过滤分析，毫秒级得到响应，快速发现数据价值。

HBase分析集群包含如下组件：

- MRS 1.8.9版本：Hadoop 2.8.3、HBase 1.3.1。
- MRS 2.0.1版本：Hadoop 3.1.1、HBase 2.1.1。
- MRS 3.1.0-LTS.1版本：Hadoop 3.1.1, HBase 2.2.3, ZooKeeper 3.5.6 and Ranger 2.0.0。
- MRS 3.1.2-LTS.3版本：Hadoop 3.1.1, HBase 2.2.3, ZooKeeper 3.6.3 and Ranger 2.0.0。
- MRS 3.2.0-LTS.1版本：Hadoop 3.3.1, HBase 2.2.3, ZooKeeper 3.6.3 and Ranger 2.0.0。

快速创建 HBase 查询集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“HBase查询集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。

- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.2.0-LTS.1及之后版本支持该参数。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。MRS 3.2.0-LTS.1及之后版本支持root用户。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。
- 密钥对：从下拉框中选择密钥对用于登录ECS机器，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。MRS 3.1.2-LTS.3及之前版本支持该参数。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.3 快速创建 Kafka 流式集群

本章节为您介绍如何快速创建一个Kafka流式集群，Kafka集群使用Kafka和Storm组件提供一个开源高吞吐量，可扩展性的消息系统。广泛用于日志收集、监控数据聚合等场景，实现高效的流式数据采集，实时数据处理存储等。

Kafka流式集群包含如下组件：

- MRS 1.8.9版本：Kafka 1.1.0、Storm 1.2.1。
- MRS 2.0.1版本：Kafka 1.1.0、Storm 1.2.1。

快速创建 Kafka 流式集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20200321”。
- 集群版本：不同版本集群提供的组件有所不同，请根据需要选择集群版本。
- 组件选择：选择“Kafka流式集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“admin”，用于登录集群管理页面。
- 密码：设置admin用户密码。
- 确认密码：再次输入设置的admin用户密码。
- 密钥对：从下拉框中选择密钥对用于登录ECS机器，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.4 快速创建 ClickHouse 集群

本章节为您介绍如何快速创建一个ClickHouse集群，ClickHouse是一个用于联机分析的列式数据库管理系统，具有极致压缩率和极速查询性能。被广泛的应用于互联网广告、App和Web流量、电信、金融、物联网等众多领域。

ClickHouse集群包含的组件：

- MRS 3.1.0-LTS.1版本：ClickHouse 21.3.4.25 and ZooKeeper 3.5.6。
- MRS 3.1.2-LTS.3版本：ClickHouse 21.3.4.25 and ZooKeeper 3.6.3。
- MRS 3.2.0-LTS.1版本：ClickHouse 22.3.2.2 and ZooKeeper 3.6.3。

CPU架构为鲲鹏计算的ClickHouse集群表引擎不支持使用HDFS和Kafka。

快速创建 ClickHouse 集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考**创建自定义集群**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20201121”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“ClickHouse集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.2.0-LTS.1及之后版本支持该参数。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。MRS 3.2.0-LTS.1及之后版本支持root用户。
- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。
- 密钥对：从下拉框中选择密钥对用于登录ECS机器，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。MRS 3.1.2-LTS.3及之前版本支持该参数。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见表5-4中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.2.5 快速创建实时分析集群

本章节为您介绍如何快速创建一个实时分析集群，实时分析集群使用Hadoop、Kafka、Flink和ClickHouse组件提供一个海量的数据采集、数据的实时分析和查询的系统。

集群包含的组件信息实时分析：

- MRS 3.1.0-LTS.1版本：Hadoop 3.1.1, Kafka 2.11-2.4.0, Flink 1.12.0, ClickHouse 21.3.4.25, ZooKeeper 3.5.6 and Ranger 2.0.0。
- MRS 3.1.2-LTS.3版本：Hadoop 3.1.1, Kafka 2.11-2.4.0, Flink 1.12.2, ClickHouse 21.3.4.25, ZooKeeper 3.6.3 and Ranger 2.0.0。
- MRS 3.2.0-LTS.1版本：Hadoop 3.3.1, Kafka 2.11-2.4.0, Flink 1.15.0, ClickHouse 22.3.2.2, ZooKeeper 3.6.3 and Ranger 2.0.0。

快速创建实时分析集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“快速创建”页签。

步骤4 参考下列参数说明配置集群基本信息，参数详细信息请参考创建自定义集群。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20201130”。
- 集群版本：默认选择最新版本即可（不同版本集群提供的组件有所不同，请根据需要选择集群版本）。
- 组件选择：选择“实时分析集群”。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.2.0-LTS.1及之后版本支持该参数。
- 集群节点：请根据自身需要选择集群节点规格数量等。MRS 3.x及之后版本集群Master节点规格不能小于64GB。
- 集群高可用：默认即可。MRS 3.x版本暂时不支持该参数。
- Kerberos认证：选择是否开启Kerberos认证。
- 用户名：默认为“root/admin”，root用于远程登录ECS机器，admin用于登录集群管理页面。MRS 3.2.0-LTS.1及之后版本支持root用户。

- 密码：设置root用户和admin用户密码。
- 确认密码：再次输入设置的root用户和admin用户密码。
- 密钥对：从下拉框中选择密钥对用于登录ECS机器，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。MRS 3.1.2-LTS.3及之前版本支持该参数。

步骤5 勾选“确认授权”开通通信安全授权，通信安全授权详情请参考**授权安全通信**。

步骤6 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。单击“访问集群”，可以查看集群详情。

集群创建的状态过程请参见**表5-4**中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

4.3 创建自定义集群

使用MRS的首要操作就是创建集群，本章节为您介绍如何在MRS管理控制台自定义创建一个新的MRS集群。

注册账号后，如果需要对云上的资源进行精细管理，请使用IAM服务创建IAM用户及用户组，并授权，以使得IAM用户获得具体的操作权限，具体请参考**创建MRS操作用户**。

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

说明

创建集群时需要注意配额提醒。当资源配额不足时，建议按照提示申请足够的资源，再创建集群。

步骤3 在创建集群页面，选择“自定义创建”页签。

步骤4 参考**软件配置**配置集群信息后，单击“下一步”。

步骤5 参考**硬件配置**配置集群信息后，单击“下一步”。

步骤6 参考**高级配置（可选）**配置集群信息后，单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤7 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建的状态过程请参见表5-4中的“状态”参数说明。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

MRS系统界面支持同一时间并发创建10个集群，且最多支持管理100个集群。

----结束

软件配置

表 4-1 MRS 集群软件配置

参数	参数说明
区域	选择区域。 不同区域的云服务产品之间内网互不相通。请就近选择靠近您业务的区域，可减少网络时延，提高访问速度。
集群名称	集群名称不允许重复。 只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 默认名称为mrs_xxxx，xxxx为字母和数字的四位随机组合数，系统自动组合。
集群版本	MRS目前支持MRS 1.8.9、MRS 2.0.1、MRS 3.1.0-LTS.1、MRS 3.1.2-LTS.3、MRS 3.2.0-LTS.1版本。
集群类型	提供几种集群类型： ● 分析集群：用来做离线数据分析，提供的是Hadoop体系的组件。 ● 流式集群：用来做流处理任务，提供的是流式处理组件。 ● 混合集群：既可以用来做离线数据分析，也可以用来做流处理任务，提供的是Hadoop体系的组件和流式处理组件。建议同时需要做离线数据分析和流处理任务时使用混合集群。 ● 自定义：用户可按照业务需求调整集群服务的部署方式，具体请参见创建自定义拓扑集群。（目前仅MRS 3.x版本支持） **说明** ● MRS流式集群不支持“作业管理”和“文件管理”功能。 ● 如需在集群中安装全部组件，请选择“自定义”类型集群。

参数	参数说明
组件选择	MRS配套的组件如下： **分析集群组件** ● Presto：开源、分布式SQL查询引擎。 ● Hadoop：分布式系统基础架构。 ● Spark：内存分布式系统框架。（ MRS 3.x版本不支持 ） ● Spark2x：Spark2x是一个对大规模数据处理的快速和通用引擎,基于开源Spark2.x版本开发。（ 仅MRS 3.x版本支持 ） ● Hive：建立在Hadoop上的数据仓库框架。 ● Opentsdb：可扩展的分布式时间序列数据库，可以存储和服务于大量时间序列数据，而不会丢失粒度。（ MRS 3.x版本不支持 ） ● HBase：分布式列数据库。 ● Tez：提供有向无环图的分布式计算框架。 ● Hue：提供Hadoop UI能力，让用户通过浏览器分析处理Hadoop集群数据。 ● Loader：基于开源sqoop 1.99.7开发，专为Apache Hadoop和结构化数据库（如关系型数据库）设计的高效传输大量数据的工具。（ MRS 3.x版本不支持 ） Hadoop为必选组件，且Spark与Hive组件需要配套使用。请根据业务选择搭配组件。 ● Flink：分布式大数据处理引擎，可对有限数据流和无限数据流进行有状态计算。 ● Oozie：Hadoop作业调度系统。（ 仅MRS 3.x版本支持 ） ● HetuEngine：HetuEngine是一个同异构大数据集的分布式SQL查询引擎。（ 仅MRS 3.1.x LTS版本支持 ） ● Ranger：一个基于Hadoop平台监控和管理数据安全的框架。 ● Impala：一种处理大量数据的SQL查询引擎。 ● ClickHouse：ClickHouse是一个用于联机分析(OLAP)的列式数据库管理系统(DBMS)。CPU架构为鲲鹏计算的ClickHouse集群表引擎不支持使用HDFS和Kafka。 ● Kudu：一种列存储管理器。 **流式集群组件** ● Kafka：提供分布式消息订阅的系统。 ● Flume：提供分布式、高可用、高可靠的海量日志采集、聚合和传输系统。 ● ZooKeeper：一个集中的服务，它用于维护配置信息、命名、提供分布式的同步和提供分组服务。（ 仅MRS 3.x版本支持 ） ● Ranger：一个基于Hadoop平台监控和管理数据安全的框架。（ 仅MRS 3.x版本支持 ）
组件端口	MRS 3.1.2-LTS.3机器之后

硬件配置

表 4-2 MRS 集群硬件配置

参数	参数说明
跨可用区部署 MRS 3.1.0-LTS.1及其之后版本支持。	MRS提供跨可用区部署集群的能力，当某一可用区故障时其他可用区依然可以正常提供服务，实现了强大的容灾能力，保障客户系统稳定运行。 如需使用该功能请联系技术支持。当开启该功能且集群中部署全部组件时，Master节点的规格内存至少为64G。 ● 可用区间均等：当每个可用区内资源充足时，集群角色均等部署在各个可用区中，集群的容灾能力更强，在某个可用区出现故障时，可以更有效的保障系统的稳定运行。 ● 2个主可用区+仲裁可用区：当某个可用区中可用资源较少时，可以指定其为仲裁可用区，仅部署需要三个可用区部署的支持跨AZ高可用的角色。两个主可用区用于安装数据存储与计算的角色，减少跨AZ的数据传递流量，降低成本，提高运算效率。
可用区	选择集群工作区域下关联的可用区。 可用区是使用独立电源和网络资源的物理区域。通过内部网络互联，再以物理方式进行隔离，提高了应用程序的可用性。建议您在不同的可用区下创建集群。
虚拟私有云	VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。 选择需要创建集群的VPC，单击“查看虚拟私有云”进入VPC服务查看已创建的VPC名称和ID。如果没有VPC，需要创建一个新的VPC。
子网	通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。 选择需要创建集群的子网，单击“查看子网”可查看所选子网的详细信息，若VPC下未创建子网，请在VPC服务控制台单击“创建子网”进行创建。网络ACL出规则配置请参考**如何配置网络ACL出规则？** 。 **说明** 创建MRS 集群需要的IP数量和集群节点和组件个数相关，集群类型不影响IP数量。 MRS集群部署默认需要的IP数量为：集群节点数量+2（Manager+DB），如果部署集群时选择Hadoop、Hue、Sqoop或Loader、Presto组件，则每一个组件需要再加一个IP。若单独创建ClickHouse集群则需要的IP数量为：集群节点数量+1（Manager）。

参数	参数说明
安全组	安全组是一组对弹性云服务器的访问规则的集合，为同一个VPC内具有相同安全保护需求并相互信任的弹性云服务器提供访问策略。 用户创建集群时，可自动创建安全组，也可选择下拉框中已有的安全组。 **说明** 选择用户自己创建的安全组时，请确保入方向规则中有一条全部协议，全部端口，源地址为可信任的IP访问范围的规则，源地址请勿使用0.0.0.0/0，否则会有安全风险。若用户不清楚可信任的IP访问范围，请选择自动创建。
弹性公网IP	通过将弹性公网IP与MRS集群绑定，实现使用弹性公网IP访问Manager的目的。 用户创建集群时，可选择下拉框中已有的弹性公网IP进行绑定。若下拉框中没有可选的弹性公网IP，可以单击“管理弹性公网IP”进入弹性公网IP服务进行创建。 **说明** 该参数仅在MRS 1.8.0及以后版本有效。 弹性公网IP必须和集群在同一区域。
企业项目	选择集群所属的企业项目，如需使用企业项目，请先通过“企业 > 项目管理”服务创建。 企业项目所在的企业管理控制台以面向企业资源管理为出发点，帮助企业以公司、部门、项目等分级管理方式实现企业云上的人员、资源、权限、财务的管理。

表 4-3 集群节点信息

参数	参数说明
CPU架构	MRS提供的CPU架构类型。MRS 3.2.0-LTS.1及之后版本支持该参数。 ● x86计算：x86 CPU架构采用复杂指令集（CISC），CISC指令集的每个小指令可以执行一些较低阶的硬件操作，指令数目多而且复杂，每条指令的长度并不相同。由于指令执行较为复杂所以每条指令花费的时间较长。 ● 鲲鹏计算：鲲鹏CPU架构采用精简指令集（RISC），RISC是一种执行较少类型计算机指令的微处理器，它能够以更快的速度执行操作，使计算机的结构更加简单合理地提高运行速度，相对于x86 CPU架构具有更加均衡的性能功耗比。鲲鹏的优势是高密度低功耗，可以提供更高的性价比。
常用模板	当“集群类型”选择“自定义”时该参数有效，请参考**自定义集群模板说明**选择。

参数	参数说明
集群节点	MRS提供节点类型： ● Master：指集群Master节点，负责管理集群，协调将集群可执行文件分配到Core节点。此外，还会跟踪每个作业的执行状态，监控DataNode的运行状况。 ● Core：指集群Core节点，处理数据并在HDFS中存储过程数据。分析集群将创建分析Core节点，流式集群将创建流式Core节点，混合集群分别创建分析Core和流式Core节点。 ● Task：指集群Task节点，主要用于计算，不存放持久数据。主要安装Yarn、Storm组件。Task节点为可选节点，数目可以是零。分析集群将创建分析Task节点，流式集群将创建流式Task节点，混合集群分别创建分析Task和流式Task节点。 当集群数据量变化不大而集群业务处理能力需求变化比较大，大的业务处理能力只是临时需要，此时选择添加Task节点。 – 临时业务量增大，如年底报表处理。 – 需要在短时间内处理完原来需要处理很久的任务，如一些紧急分析任务。
实例规格	选择主节点和核心节点的实例规格。MRS当前支持主机规格的配型由CPU+内存+Disk共同决定。单击 ，配置集群节点的实例规格、系统盘和数据盘参数。 **说明** ● 节点的实例规格配置越高，数据处理分析能力越强。 ● MRS 3.x及之后版本集群Master节点规格不能小于64GB。
系统盘	节点系统盘的存储类型和存储空间。 存储类型： ● SATA：普通IO ● SAS：高IO ● SSD：超高IO ● GPSSD：通用型SSD

参数	参数说明
数据盘	节点数据磁盘存储空间。为增大数据存储容量，创建集群时可同时添加磁盘，有如下应用场景： ● 数据存储和计算分离，数据存储在OBS中，集群存储成本低，存储量不受限制，并且集群可以随时删除，但计算性能取决于OBS访问性能，相对HDFS有所下降，建议在数据计算不频繁场景下使用。 ● 数据存储和计算不分离，数据存储在HDFS中，集群成本较高，计算性能高，但存储量受磁盘空间限制，删除集群前需将数据导出保存，建议在数据计算频繁场景下使用。 目前的存储类型： ● SATA：普通IO ● SAS：高IO ● SSD：超高IO ● GPSSD：通用型SSD **说明** 创建的节点个数越多，对管理节点（即master节点）的硬盘容量要求越高。为了保证集群能够健康地运行，当创建的节点个数达到300时，建议将master的硬盘容量配置成600GB以上；当创建的节点个数达到500时，建议将master的硬盘容量配置成1TB以上。
实例数量	配置主节点和核心节点的个数。 Master： ● 开启“集群高可用”时，Master实例数量固定为2个。 ● 关闭“集群高可用”时，Master实例数量固定为1个。 Core节点至少存在一个，Core节点和Task节点的数量之和不能超过500个。 Task：单击添加Task节点。单击修改Task节点额实例规格和磁盘配置。单击删除已添加的Task节点。 **说明** ● Core节点默认的最大值为500，如果用户需要的Core节点数大于500，请联系技术支持人员。 ● 过小的节点容量会导致您的集群运行缓慢，而过大的节点容量会产生不必要的成本，请根据您要处理的数据对集群节点数量进行调整。
拓扑调整	当常用模板中的部署方式不满足需求，请设置“拓扑调整”为“开启”，然后根据业务需要调整实例部署方式，具体说明请参见**自定义集群拓扑调整说明**。当集群类型为“自定义”时该参数有效。

高级配置（可选）

表 4-4 MRS 集群高级配置拓扑

参数	参数说明
标签	具体请参考**添加集群标签**。
主机名前缀	用作集群中ECS机器主机名的前缀。
弹性伸缩	请在“硬件配置”页签指定Task节点的规格，然后参考**配置弹性伸缩规则**配置。
引导操作	具体请参考**添加引导操作**。MRS 3.x版本暂时不支持该参数。
委托	通过绑定委托，ECS或BMS云服务将有权限来管理您的部分资源，请根据实际业务场景需求确认是否需要配置委托。 例如通过配置ECS委托可自动获取AK/SK访问OBS，具体请参见**配置存算分离集群（委托方式）**。 **MRS_ECS_DEFAULT_AGENCY**委托拥有对象存储服务的OBSOperateAccess权限和在集群所在区域拥有CESFullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。
指标共享	用于采集大数据组件的监控指标，当用户使用集群过程中出现问题时，供支持人员定位问题。MRS 3.x版本暂时没有该参数。
OBS权限控制	开启细粒度权限控制的用户可以通过该功能实现不同的MRS用户对OBS文件系统下的不同目录有不同的权限。具体请参见**配置MRS多用户访问OBS细粒度权限**。MRS 3.x版本暂时没有该参数。
数据盘加密	是否对集群挂载的数据盘中的数据进行加密，默认关闭。如需使用该功能，当前用户必须拥有“Security Administrator”和“KMS Administrator”权限。MRS 3.x版本暂时没有该参数。 加密数据盘使用的密钥由密码安全中心（Data Encryption Workshop，DEW）中的密钥管理（KMS，Key Management Service）功能提供，无需您自行构建和维护密钥管理基础设施，安全便捷。 通过单击“数据盘加密”开启或关闭数据盘加密功能。
密钥ID	当“数据盘加密”功能开启时，显示该参数。用于显示已选择的密钥名称对应的密钥ID。MRS 3.x版本暂时没有该参数。

参数	参数说明
密钥名称	当“数据盘加密”功能开启时，需要配置该参数。选择用来加密数据盘的密钥名称，默认选择密钥名称为“evs/default”的默认主密钥，在下拉框中可以选择其他用户主密钥。MRS 3.x版本暂时没有该参数。 使用用户主密钥加密云硬盘，若对用户主密钥执行禁用、计划删除等操作，将会导致云硬盘不可读写，甚至数据永远无法恢复，请谨慎操作。 单击“查看密钥列表”，进入密钥管理页面可以创建及管理密钥。
告警	开启告警功能可在集群运行异常或系统故障时，及时通知集群维护人员定位问题。
规则名称	用户自定义发送告警消息的规则名称，只能包含数字、英文字符、中划线和下划线。
主题名称	选择已创建的主题，也可以单击“创建主题”重新创建。新创建的主题请参考**向主题添加订阅**向该主题添加订阅者才能接收发布至主题的消息。 主题是发送消息和订阅通知的信道，为发布者和订阅者提供一个可以相互交流的通道。
日志记录	集群创建失败时，是否收集失败日志。 开启日志记录开关之后将自动收集集群创建失败、扩/缩容失败等场景下的系统日志及相关组件运行日志到OBS文件系统中，该日志用于运维人员快速定位问题。该日志信息将最多保留7天。
Kerberos认证	登录Manager管理页面时是否启用Kerberos认证。 • ：“Kerberos认证”关闭时，普通用户可使用MRS集群的所有功能。建议单用户场景下使用。 • ：“Kerberos认证”开启时，普通用户无权限使用MRS集群的“文件管理”和“作业管理”功能，并且无法查看Hadoop、Spark的作业记录以及集群资源使用情况。如果需要使用集群更多功能，需要找Manager的管理员分配权限。建议在多用户场景下使用。
用户名	Manager管理员用户，目前默认为**admin**用户。

参数	参数说明
密码	配置Manager管理员用户的密码。 需要满足： ● 密码长度应在8～26个字符之间 ● 必须包含如下4种字符的组合 – 至少一个小写字母 – 至少一个大写字母 – 至少一个数字 – 至少一个特殊字符：！?,.: -_{} [ ]@ $% ^ + = / ● 不能和用户名或倒序的用户名相同 安全程度：颜色条红、橙、绿分别表示密码安全强度弱、中、强。
确认密码	再次输入Manager管理员用户的密码。
登录方式 MRS 3.2.0-LTS.1及之后的版本支持。	● 密码 使用密码方式登录ECS节点。 密码设置约束如下： 1. 字符串类型，可输入的字符串长度为8~26。 2. 至少包含四种字符组合，如大写字母，小写字母，数字，特殊字符（！?,.: -_{} [ ]@ $% ^ + = /）。 3. 不能与用户名或倒序用户名相同。 ● 密钥对 使用密钥方式登录集群ECS节点。从下拉框中选择密钥对，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。 密钥对即SSH密钥，包含SSH公钥和私钥。您可以新建一个SSH密钥，并下载私钥用于远程登录身份认证。为保证安全，私钥只能下载一次，请妥善保管。 您可以通过以下两种方式中的任意一种使用SSH密钥。 1. 创建SSH密钥：创建SSH密钥，同时会创建公钥和私钥，公钥保存在ECS系统中，私钥保存在用户本机。当登录弹性云服务器时，使用公钥和私钥进行鉴权。 2. 导入SSH密钥：当用户已有公钥和私钥，可以选择将公钥导入系统。当登录弹性云服务器时，使用公钥和私钥进行鉴权。

参数	参数说明
密钥对	使用密钥方式登录集群ECS节点。从下拉框中选择密钥对，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。 密钥对即SSH密钥，包含SSH公钥和私钥。您可以新建一个SSH密钥，并下载私钥用于远程登录身份认证。为保证安全，私钥只能下载一次，请妥善保管。 您可以通过以下两种方式中的任意一种使用SSH密钥。 1. 创建SSH密钥：创建SSH密钥，同时会创建公钥和私钥，公钥保存在ECS系统中，私钥保存在用户本机。当登录弹性云服务器时，使用公钥和私钥进行鉴权。 2. 导入SSH密钥：当用户已有公钥和私钥，可以选择将公钥导入系统。当登录弹性云服务器时，使用公钥和私钥进行鉴权。
通信安全授权	MRS集群通过管理控制台为用户发放、管理和使用大数据组件，大数据组件部署在用户的VPC内部，MRS管理控制台需要直接访问部署在用户VPC内的大数据组件时需要开通相应的安全组规则，而开通相应的安全组规则需要获取用户授权，此授权过程称为通信安全授权。具体请参考**授权安全通信**。 若不开启通信安全授权，MRS将无法创建集群。

集群创建失败

如果集群创建失败后，失败任务会自动转入“失败任务管理”页面。选择“集群列表 > 现有集群”，单击**图4-1**中进入“失败任务管理”页面，在“任务状态”列中，将鼠标移动到上可以查看到失败原因。可以参见**查看失败的集群操作任务**章节删除失败任务。

图 4-1 失败任务管理

MRS集群创建失败错误码列表如**表4-5**所示。

表 4-5 错误码

错误码	说明
MRS.101	用户请求配额不足，请联系客服提升配额。
MRS.102	用户Token为空或不合法，请稍后重试或联系客服。
MRS.103	用户请求不合法，请稍后重试或联系客服。

错误码	说明
MRS.104	用户资源不足，请稍后重试或联系客服。
MRS.105	现子网IP不足，请稍后重试或联系客服。
MRS.201	因ECS服务导致失败，请稍后重试或联系客服。
MRS.202	因IAM服务导致失败，请稍后重试或联系客服。
MRS.203	因VPC服务导致失败，请稍后重试或联系客服。
MRS.400	MRS内部出错，请稍后重试或联系客服。

4.4 创建自定义拓扑集群

MRS当前提供的“分析集群”、“流式集群”和“混合集群”采用固定模板进行部署集群的进程，无法满足用户自定义部署管理角色和控制角色在集群节点中的需求。如需自定义集群部署方式，可在创建集群时的“集群类型”选择“自定义”，实现用户自主定义集群的进程实例在集群节点中的部署方式。仅MRS 3.x及之后版本支持创建自定义拓扑集群。

自定义集群可实现以下功能：

- 管控分离部署，管理角色和控制角色分别部署在不同的Master节点中。
- 管控合设部署，管理角色和控制角色共同部署在Master节点中。
- ZooKeeper单独节点部署，增加可靠性。
- 组件分开部署，避免资源争抢。

MRS集群中角色类型：

- 管理角色：Management Node(MN)，安装Manager，即MRS集群的管理系统，提供统一的访问入口。Manager对部署在集群中的节点及服务进行集中管理。
- 控制角色：Control Node(CN)，控制监控数据角色执行存储数据、接收数据、发送进程状态及完成控制节点的公共功能。 MRS的控制节点包括HMaster、HiveServer、ResourceManager、NameNode、JournalNode、SlapdServer等。
- 数据角色：Data Node(DN)，执行管理角色发出的指示，上报任务状态、存储数据，以及执行数据节点的公共功能。MRS的数据节点包括DataNode、RegionServer、NodeManager等。

创建自定义集群

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“自定义创建”页签。

步骤4 参考下列参数说明配置集群软件信息，参数详细信息请参考**软件配置**。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs_20180321”。

- 集群版本：目前仅MRS 3.x版本支持。
- 集群类型：选择“自定义”并根据需要勾选对应组件。

步骤5 单击“下一步”，并配置硬件信息。

- 跨可用区部署：默认即可。MRS 3.1.0-LTS.1及其之后版本支持。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 安全组：选择“自动创建”。
- 弹性公网IP：选择“暂不绑定”。
- 企业项目：默认即可。
- CPU架构：默认即可。MRS 3.2.0-LTS.1及之后版本支持该参数。
- 常用模板：具体说明请参见**自定义集群模板说明**。
- 实例规格：单击 配置实例规格、系统盘和数据盘存储类型和存储空间。
- 实例数量：请根据业务量调整集群实例数量。具体可参考**表4-7**。
- 拓扑调整：若常用模板中的部署方式不满足需求或者需要手动安装部分默认安装不部署的实例或者需要手动安装部分实例时，请设置“拓扑调整”为“开启”，然后根据业务需要调整实例部署方式，具体说明请参见**自定义集群拓扑调整说明**。

步骤6 单击“下一步”进入高级配置页签。

参数说明请参见**高级配置（可选）**。

步骤7 单击“立即创建”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

步骤8 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

----结束

自定义集群模板说明

表 4-6 自定义集群常用模板说明

常用模板	说明	节点数量范围
管控合设	管理角色和控制角色共同部署在Master节点中，数据实例合设在同一节点组。该部署方式适用于100个以下的节点，可以减少成本。	• Master节点数量大于等于3个，小于等于11个。 • 节点组数量总和小于等于10个，非Master节点组中节点数量总和小于等于10000个。
管控分设	管理角色和控制角色分别部署在不同的Master节点中，数据实例合设在同一节点组。该部署方式适用于100-500个节点，在高并发负载情况下表现更好。	• Master节点数量大于等于5个，小于等于11个。 • 节点组数量总和小于等于10个，非Master节点组中节点数量总和小于等于10000个。
数据分设	管理角色和控制角色分别部署在不同的Master节点中，数据实例分设在不同节点组。该部署方式适用于500个以上的节点，可以将各组件进一步分开部署，适用于更大的集群规模。	• Master节点数量大于等于9个，小于等于11个。 • 节点组数量总和小于等于10个，非Master节点组中节点数量总和小于等于10000个。

表 4-7 MRS 自定义集群节点部署方案

节点部署原则		适用场景	组网规则
管理节点、控制节点和数据节点分开部署（此方案至少需要8个节点）	MN × 2 + CN × 9 + DN × n	（推荐）数据节点数500-2000时采用此方案	• 集群节点数超过200时，各节点划分到不同子网，各子网通过核心交换机三层互联，每个子网的节点数控制在200个以内，不同子网中节点数量请保持均衡。 • 集群节点数低于200时，各节点部署在同一子网，集群内通过汇聚交换机二层互联。
	MN × 2 + CN × 5 + DN × n	（推荐）数据节点数100-500时采用此方案	
	MN × 2 + CN × 3 + DN × n	（推荐）数据节点数30-100时采用此方案	

节点部署原则		适用场景	组网规则
管理节点和控制节点合并部署，数据节点单独部署	(MN+CN) × 3 + DN × n	（推荐）数据节点数3-30时采用此方案	集群内节点部署在同一子网，集群内通过汇聚交换机二层互联。
管理节点、控制节点和数据节点合并部署		• 节点数小于6的集群使用此方案 • 此方案至少需要3个节点 **说明** 生产环境或商用环境不推荐使用此场景： • 管理节点、控制节点和数据节点合并部署时，集群性能和可靠性都会产生较大影响。 • 如节点数量满足需求，建议将数据节点单独部署。 • 如节点数量不满足将数据节点单独部署的要求，必须使用此场景时，需要使用双平面组网方式。将管理网络与业务网络流量隔离，防止业务平面的数据量过大，导致管理操作不能正常下发。	集群内节点部署在同一子网，集群内通过汇聚交换机二层互联。

自定义集群拓扑调整说明

表 4-8 拓扑调整说明

服务名称	依赖关系	角色名称	角色业务部署建议	说明
OMSServer	-	OMSServer	部署在Master节点上，不支持修改。	-
ClickHouse	依赖ZooKeeper	CHS（ClickHouseServer）	所有节点均可部署。 角色实例部署数量范围：偶数个，2~256。	部署了该角色的非Master节点组会被认为是Core节点类型。
		CLB（ClickHouseBalancer）	所有节点均可部署。 角色实例部署数量范围：2~256。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
ZooKeeper	-	QP(quorumpeer)	只能部署在Master节点上。 角色实例部署数量范围：3~9，步长为2。	-
Hadoop	依赖ZooKeeper	NN(NameNode)	只能部署在Master节点上。 角色实例部署数量范围：2。	NameNode与Zkfc进程共机部署用于集群高可用
		HFS（HttpFS）	只能部署在Master节点上。 角色实例部署数量范围：0~10。	-
		JN(JournalNode)	只能部署在Master节点上。 角色实例部署数量范围：3~60，步长为2。	-
		DN(DataNode)	所有节点均可部署。 角色实例部署数量范围：3~10000。	部署了该角色的非Master节点组会被认为是Core节点类型。
		RM(ResourceManager)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		NM(NodeManager)	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
		JHS(JobHistoryServer)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
		TLS(TimelineServer)	只能部署在Master节点上。 角色实例部署数量范围：0~1。	-
Presto	依赖Hive	PCD(Coordinator)	只能部署在Master节点上。 角色实例部署数量范围：2。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
		PWK(Worker)	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
Spark2x	• 依赖Hadoop • 依赖Hive • 依赖ZooKeeper	JS2X(JDBCServer2x)	只能部署在Master节点上。 角色实例部署数量范围：2~10。	-
		JH2X(JobHistory2x)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		SR2X(SparkResource2x)	只能部署在Master节点上。 角色实例部署数量范围：2~50。	-
		IS2X(IndexServer2x)	（可选）只能部署在Master节点上。 角色实例部署数量范围：0~2，步长为2。	-
HBase	依赖Hadoop	HM(HMaster)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
		TS(ThriftServer)	所有节点均可部署。 角色实例部署数量范围：0~10000。	-
		RT(RESTServer)	所有节点均可部署。 角色实例部署数量范围：0~10000。	-
		RS(RegionServer)	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
		TS1(Thrift1Server)	所有节点均可部署。 角色实例部署数量范围：0~10000。	若集群安装了Hue服务并且需要在Hue WebUI使用HBase，HBase服务需安装此实例。

服务名称	依赖关系	角色名称	角色业务部署建议	说明
Hive	• 依赖Hadoop • 依赖DBService	MS(MetaStore)	只能部署在Master节点上。 角色实例部署数量范围：2~10。	-
		WH（WebHCat）	只能部署在Master节点上。 角色实例部署数量范围：1~10。	-
		HS(HiveServer)	只能部署在Master节点上。 角色实例部署数量范围：2~80。	-
Hue	依赖DBService	H(Hue)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
Sqoop	依赖Hadoop	SC(SqoopClient)	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
Kafka	依赖ZooKeeper	B(Broker)	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
Flume	-	MS(MonitorServer)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
		F(Flume)	所有节点均可部署。 角色实例部署数量范围：1~10000。	部署了该角色的非Master节点组会被认为是Core节点类型。
Tez	• 依赖Hadoop • 依赖DBService • 依赖ZooKeeper	TUI(TezUI)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
Flink	• 依赖ZooKeeper • 依赖Hadoop	FR(FlinkResource)	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
		FS(FlinkServer)	所有节点均可部署。 角色实例部署数量范围：0~2。	-
Oozie	• 依赖Hadoop • 依赖DBService • 依赖ZooKeeper	O(oozie)	只能部署在Master节点上。 角色实例部署数量范围：2。	-
Impala	• 依赖Hadoop • 依赖Hive • 依赖DBService • 依赖ZooKeeper	StateStore	只能部署在Master节点上。 角色实例部署数量范围：1。	-
		Catalog	只能部署在Master节点上。 角色实例部署数量范围：1。	-
		Impalad	所有节点均可部署。 角色实例部署数量范围：1~10000。	-
Kudu	-	KuduMaster	只能部署在Master节点上。 角色实例部署数量范围：3或者5。	-
		KuduTserver	所有节点均可部署。 角色实例部署数量范围：3~10000。	-
Ranger	依赖DBservice	RA(RangerAdmin)	只能部署在Master节点上。 角色实例部署数量范围：1~2。	-
		USC(UserSync)	只能部署在Master节点上。 角色实例部署数量范围：1。	-

服务名称	依赖关系	角色名称	角色业务部署建议	说明
		TSC（TagSync）	所有节点均可部署。 角色实例部署数量范围：0~1。	-

4.5 添加集群标签

标签是集群的标识。为集群添加标签，可以方便用户识别和管理拥有的集群资源。

您可以在创建集群时添加标签，也可以在集群创建完成后，在集群的详情页添加标签，您最多可以给集群添加10个标签。

标签共由两部分组成：“标签键”和“标签值”，其中，“标签键”和“标签值”的命名规则如**表4-9**所示。

表 4-9 标签命名规则

参数	规则	样例
标签键	不能为空。 对于同一个集群，Key值唯一。 长度不超过36个字符。 不能包含“=”，“*”，“<”，“>”，“\”，“,”，“|”，“/”，且首尾字符不能为空格。	Organization
标签值	长度不超过43个字符。 不能包含“=”，“*”，“<”，“>”，“\”，“,”，“|”，“/”，且首尾字符不能为空格。value可以为空。	Apache

为集群增加标签

在申请集群页，为集群增加标签。

1. 登录MRS管理控制台。
2. 单击“创建集群”，进入创建集群页面。
3. 在创建集群页面，选择“自定义创建”。
4. 参考**创建自定义集群**配置集群软件配置和硬件配置信息。
5. 在“高级配置”页签的标签栏。
   输入新添加标签的键和值。

系统支持添加多个标签，最多可添加10个标签，并取各个标签的交集，对目标集群进行搜索。

说明

您也可对现有集群增加标签，详见**管理标签**。

搜索目标集群

在现有集群列表页，按标签键或标签值搜索目标集群。

1. 登录MRS管理控制台。
2. 单击现有集群列表右上角的“标签搜索”，展开查询页。
3. 输入待查询集群的标签。

   标签键或标签值可以通过下拉列表中选择，当标签键或标签值全匹配时，系统可以自动查询到目标集群。当有多个标签条件时，会取各个标签的交集，进行集群查询。
4. 单击“搜索”。

   系统根据标签键或标签值搜索目标集群。

管理标签

在现有集群的标签页，执行标签的增、删、改、查操作。

1. 登录MRS管理控制台。
2. 在现有集群列表中，单击待管理标签的集群名称。

   系统跳转至该集群详情页面。
3. 选择“标签管理”页签，对集群的标签执行增、删、改、查。
   - 查看

     在“标签”页，可以查看当前集群的标签详情，包括标签个数，以及每个标签的键和值。
   - 添加

     单击左上角的“添加标签”，在弹出的“添加标签”窗口，输入新添加标签的键和值，并单击“确定”。
   - 修改

     单击标签所在行“操作”列下的“编辑”，在弹出的“编辑标签”窗口，输入修改后标签的值，并单击“确定”。
   - 删除

     单击标签所在行“操作”列下的“删除”，如果确认删除，在弹出的“删除标签”窗口，单击“确定”。

     说明

     MRS标签更新会同步到集群中的每台ECS上，为了使所有ECS标签与MRS标签保持一致，不建议在ECS服务控制台上单独修改MRS集群的ECS标签。当集群中某个ECS的标签数量达到上限时，集群将不能再创建标签。

4.6 授权安全通信

MRS集群通过管理控制台为用户发放、管理和使用大数据组件，大数据组件部署在用户的VPC内部，MRS管理控制台需要直接访问部署在用户VPC内的大数据组件时需要开

通相应的安全组规则，而开通相应的安全组规则需要获取用户授权，此授权过程称为通信安全授权。

若不开启通信安全授权，MRS将无法创建集群。集群创建成功后若关闭通信将导致集群状态为“网络通道未授权”且如下功能将受到影响：

- 大数据组件安装、集群扩容、集群缩容、升级Master节点规格功能不可用。
- 集群的运行状态、告警、事件无法监控。
- 集群详情页的节点管理、组件管理、告警管理、文件管理、作业管理、补丁管理、租户管理功能不可用。
- Manager页面、各组件的Web站点无法访问。

再次开启通信安全授权，集群状态会恢复为“运行中”，以上功能将恢复为可用。具体操作请参见**为关闭安全通信的集群开启安全通信**。

当集群中授权的安全组规则不足以支撑MRS集群管理控制台为用户发放、管理和使用大数据组件的操作时，“通信安全授权”右侧出现 的提示，请单击“一键修复”按钮进行修复，具体请参考**一键修复**。

创建集群时开启安全通信

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入创建集群页面。

步骤3 在创建集群页面，选择“快速创建”或“自定义创建”。

步骤4 参考**快速集群**或**创建自定义集群**配置集群信息。

步骤5 在“通信安全授权”栏，勾选“确认授权”。

步骤6 单击“立即创建”创建集群。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

----结束

集群创建成功后关闭安全通信

步骤1 登录MRS管理控制台。

步骤2 在现有集群列表中，单击待关闭安全通信的集群名称。

系统跳转至该集群详情页面。

步骤3 单击“通信安全授权”右侧的开关关闭授权，在弹出窗口单击“确定”。

关闭授权后将导致集群状态变更为“网络通道未授权”，集群部分功能不可用，请谨慎操作。

----结束

为关闭安全通信的集群开启安全通信

步骤1 登录MRS管理控制台。

步骤2 在现有集群列表中，单击待开启安全通信的集群名称。

系统跳转至该集群详情页面。

步骤3 单击“通信安全授权”右侧的开关开启授权。

开启授权后集群状态变更为“运行中”。

----结束

一键修复

当集群中授权的安全组规则不足以支撑MRS集群管理控制台为用户发放、管理和使用大数据组件的操作时，“通信安全授权”右侧出现 的提示，请单击“一键修复”按钮进行修复。

步骤1 登录MRS管理控制台。

步骤2 在现有集群列表中，单击待修复安全通信的集群名称。

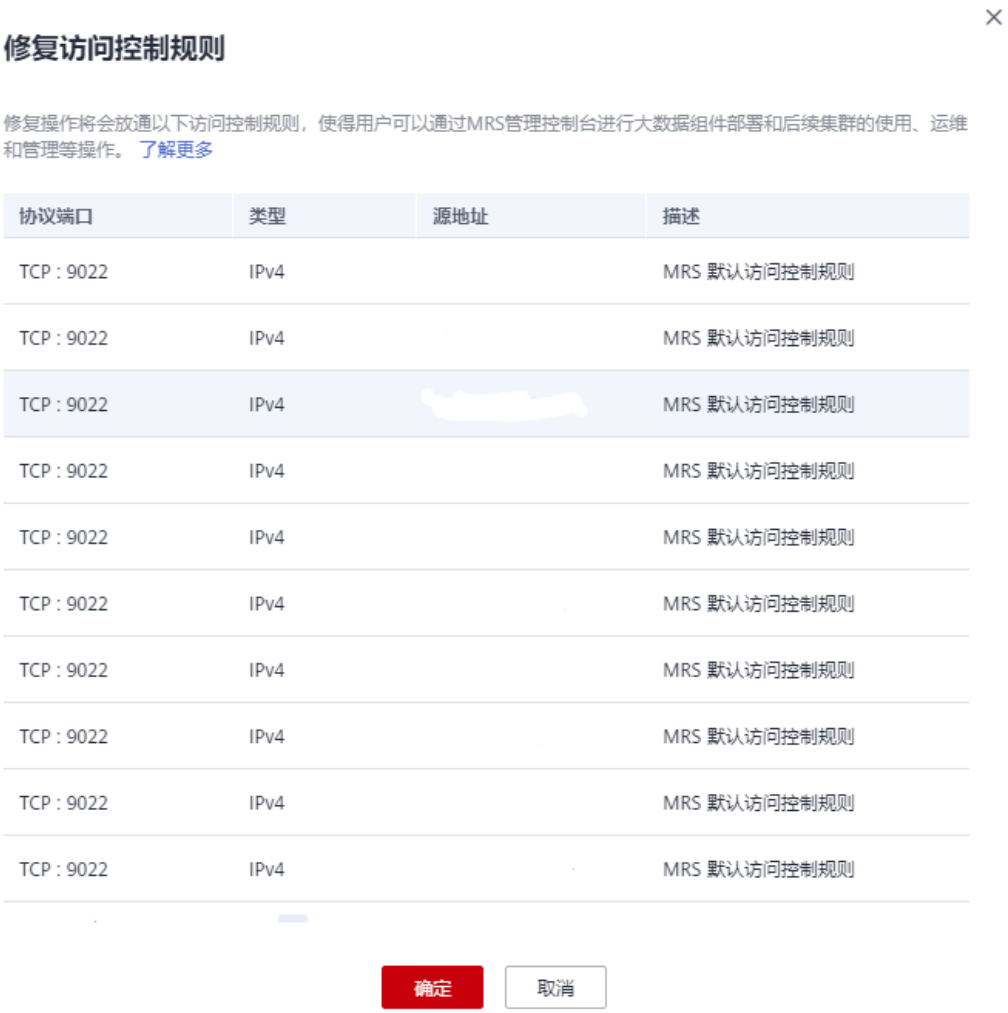
系统跳转至该集群详情页面。

步骤3 单击“通信安全授权”右侧的“一键修复”。

图 4-2 一键修复

步骤4 单击“确定”，完成修复。

图 4-3 修复访问控制策略



----结束

4.7 配置弹性伸缩规则

4.7.1 简介

在大数据应用，尤其是实时分析处理数据的场景中，常常需要根据数据量的变化动态调整集群节点数量以增减资源。MRS的弹性伸缩规则功能支持根据集群负载对集群的Task节点进行弹性伸缩。如果数据量是按照周期进行有规律的变化，用户可以按照固定时间段来自动调整Task节点数量范围，从而在数据量变化前提前完成集群的扩缩容。

- 弹性伸缩规则：根据集群实时负载指标对Task节点数量进行调整，数据量变化后触发扩缩容，有一定的延后性。
- 资源计划：按时间段设置Task节点数量范围，若数据量变化存在周期性规律，则可通过资源计划在数据量变化前提前完成集群的扩缩容，避免出现增加或减少资源的延后。

弹性伸缩规则与资源计划均可触发弹性伸缩，两者必须至少配置其中一种，也可以叠加使用。资源计划与基于负载的弹性伸缩规则叠加使用可以使得集群节点的弹性更好，足以应对偶尔超出预期的数据峰值出现。

当某些业务场景要求在集群扩缩容之后，根据节点数量的变化对资源分配或业务逻辑进行更改时，手动扩缩容的场景客户可以登录集群节点进行操作。对于弹性伸缩场景，MRS支持通过自定义弹性伸缩自动化脚本来解决。自动化脚本可以在弹性伸缩前后执行相应操作，自动适应业务负载的变化，免去了人工操作。同时，自动化脚本给用户实现个性需求提供了途径，完全自定义的脚本与多个可选的执行时机基本可以满足用户的各项需求，使弹性伸缩更具灵活性。

- 弹性伸缩规则：
  - 用户对于一个集群，可以同时设置扩容、缩容最多各5条弹性伸缩规则。
  - 系统根据用户的配置顺序从前到后依次判断规则，先扩容，后缩容。请尽量把重要的策略放在前面，以防一次扩容或缩容无法达到预期效果而进行反复触发。
  - 比对因子包括大于、大于等于、小于、小于等于。
  - 集群连续5n（n默认值为1）分钟持续满足配置的指标阈值后才能触发扩容或者缩容。
  - 每次扩容或者缩容后，存在一个冷却时间，冷却时间默认为20分钟，最小值为0。
  - 单次扩容或者缩容的节点数，最小1个节点，最大100个节点。
  - 集群Task节点数量会被限定在用户配置的默认节点数量范围内或者当前时间段生效的资源计划的节点数量范围内。当前时间段生效的资源计划的节点数量范围优先级更高。
- 资源计划（按时间段设置Task节点数量范围）：
  - 用户可以按时间段设置集群Task节点的最大数量和最小数量，当集群Task节点数不满足当前时间资源计划节点范围要求时，系统触发扩容或缩容。
  - 用户最多可以为一个集群设置5条资源计划。
  - 资源计划周期以天为单位，起始时间与结束时间可以设置为00:00-23:59之间的任意时间点。起始时间早于结束时间至少30分钟。不同资源计划配置的时间段不可交叉。
  - 资源计划触发扩容或缩容后，存在10分钟的冷却时间，冷却时间内不会再次触发弹性伸缩。
  - 当启用资源计划时，在除配置资源计划配置时间段的其他时间内，集群Task节点数量会被限定在用户配置的默认节点数量范围内。
- 自动化脚本：
  - 用户可以设置自定义脚本，当弹性伸缩触发时，在集群节点上自动运行。
  - 用户最多可以为一个集群设置10个自动化脚本。
  - 可以指定自动化脚本某种或多种类型的节点上执行。
  - 脚本执行时机可以是扩容前、扩容后、缩容前或缩容后。
  - 使用自动化脚本前，请先将脚本上传到集群虚拟机或与集群同region的OBS文件系统中。集群虚拟机上的脚本只能在已有节点上执行，若脚本需要在新扩容的节点上执行，请将脚本上传到OBS。

4.7.2 创建集群时配置弹性伸缩

创建集群时，弹性伸缩功能可以通过在高级配置参数中进行配置。

说明

仅分析集群、流式集群和混合集群支持创建集群时配置弹性伸缩策略。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在创建包含有Task类型节点组件的集群时，参考**创建自定义集群**配置集群软件配置和硬件配置信息后，在“高级配置”页签的弹性伸缩栏，打开对应Task节点类型后的开关按钮，即可进行弹性伸缩规则及资源计划的配置或修改。

说明

您可以参考以下场景进行配置：

- **场景1：单独配置弹性伸缩规则**
- **场景2：单独使用资源计划**
- **场景3：弹性伸缩规则与资源计划叠加使用**

----结束

4.7.3 为已有集群新增弹性伸缩策略

集群创建成功后，可以通过节点组维度或资源池维度对集群内的Task节点组配置相关规则。

由于节点组维度策略和资源池维度策略互斥，用户根据实际需要选择其中一个配置即可。

指定资源池维度策略在MRS 3.1.5及其之后版本支持。

对比项	节点组维度	资源池维度
弹性伸缩对象	Task节点组内的所有节点	弹性伸缩策略指定资源池内的Task节点
扩容节点资源池归属	default资源池	弹性伸缩策略指定资源池
缩容对象	Task节点组内随机缩容	弹性伸缩策略指定资源池内随机缩容

前提条件

- 弹性伸缩仅用于Task节点组，当集群内没有Task节点时，需先**添加Task节点**后再继续配置。
- 资源池维度时必须先添加资源池，具体请参考**添加资源池**。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 选择“弹性伸缩”页签，进入配置弹性伸缩界面。

用户根据实际需要选择资源池维度或者节点组维度进行策略配置。

说明

- 不同节点组之间的弹性伸缩策略互斥，即只能存在一个节点组有开启的弹性伸缩策略。

步骤4 单击“新增弹性伸缩”按钮，即可添加弹性伸缩策略。

说明

您可以参考以下场景进行配置：

- **场景1：单独配置弹性伸缩规则**
- **场景2：单独使用资源计划**
- **场景3：弹性伸缩规则与资源计划叠加使用**

----结束

4.7.4 场景 1：单独配置弹性伸缩规则

单独配置弹性伸缩规则场景：需要根据Yarn资源使用情况动态调整节点数，在Yarn可用内存低于20%时扩容5个节点，可用内存高于70%时缩容5个节点。Task节点组最高不超过10个节点，最低不少于1个节点。

操作步骤

步骤1 进入弹性伸缩配置界面后，配置弹性伸缩规则。

- 配置节点默认范围

  输入Task实例弹性伸缩的数量范围，此约束应用于所有扩容缩容规则，取值范围为0～500。

  例如本业务场景中，配置为1~10。

- 配置弹性伸缩规则

  需要配置扩容或者缩容规则，否则弹性伸缩将不会启用。

  a. 规则类型选择“扩容”或者“缩容”。

  b. 单击“添加规则”，进入规则编辑页面。

  c. 配置“规则名称”、“如果”、“持续”、“添加”、“冷却时间”，具体弹性伸缩指标含义可以参考**配置弹性伸缩指标**。

  d. 单击“确定”。

  您可以在弹性伸缩页面，扩容或者缩容区域查看、编辑或删除配置的规则。您可以继续添加并配置多条规则。

步骤2 单击“确定”，完成弹性伸缩规则设置。

说明

如果是为已有集群配置弹性伸缩的场景，需勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

----结束

4.7.5 场景 2：单独使用资源计划

当数据量以天为周期有规律的变化，并且希望在数据量变化前提前完成集群的扩缩容，可以使用MRS的资源计划配置在规定时间内按计划调整Task节点数量。

背景信息

某项实时处理业务数据量在每天7:00~13:00出现高峰，其他时间保持平稳低水平。假设我们使用MRS流式集群来处理该业务数据，在7:00~13:00时，为应对数据量高峰需要5个Task节点的资源，其他时间只需要2个Task节点。

操作步骤

步骤1 进入弹性伸缩配置界面后，配置资源计划。

步骤2 节点数量范围的“默认范围”设置为“2-2”，表示除资源计划规定时间范围外，其他时间Task节点数量固定为2个。

步骤3 单击默认范围下方的“配置指定时间段的节点数量范围”或者“添加资源计划”。

步骤4 配置“生效日期”、“时间范围”和“节点数量范围”。

例如此处“生效日期”设置为周一、周二和周六，“时间范围”设置为“07:00-13:00”，“节点数量范围”设置为“5-5”，表示在该时间范围内，Task节点数量固定为5个。

单击“配置指定时间段的节点数量范围”配置多条资源计划。

说明

- 生效日期默认是每日生效，也可以选择周一至周日任意一天或几天生效。
- 如果没有配置指定时间段的节点数量范围，则节点数量范围以“默认范围”为准。
- 如果配置了指定时间段的节点数量范围，则在这个时间范围内，以配置的“节点数量范围”为准。不在配置的时间范围时，则以“默认范围”为准。

----结束

4.7.6 场景 3：弹性伸缩规则与资源计划叠加使用

假如数据量并非非常平稳，有可能出现超出预期的波动，因此并不能保证固定Task节点范围一定可以满足业务场景，我们需要在资源计划的基础上根据实时负载对Task节点数量进行调整。

背景信息

某项实时处理业务数据量在每天7:00-13:00出现规律性变化，但是数据量变化并非非常平稳。假设在7:00-13:00期间，需要Task节点的数量范围是5~8个，其他时间需要Task节点数量是2~4个。因此可以在资源计划的基础上，设置基于负载的弹性伸缩规则，以实现当数据量超出预期后，Task节点数量可以在资源计划规定的范围内根据负载情况进行浮动，但不会超出该规定范围。资源计划触发时，会以变化最小的方式使节点数量满足计划规定范围，即如果需要扩容则扩容到计划节点数量范围的下限，如果需要缩容则缩容到计划节点数量范围的上限。

操作步骤

步骤1 进入弹性伸缩配置界面后，配置弹性伸缩规则。

- 节点数量范围的默认范围：

  输入Task实例弹性伸缩的数量范围，此约束应用于所有扩容缩容规则。

  例如本场景中，配置为2~4个。

- 伸缩规则：

  需要配置扩容或者缩容，否则弹性伸缩将不会启用。

  a. 规则类型选择“扩容”或者“缩容”。

  b. 单击“添加规则”，进入“添加规则”页面。

  c. 配置“规则名称”、“如果”、“持续”、“添加”、“冷却时间”。

  d. 单击“确定”。

  您可以在弹性伸缩页面，扩容或者缩容区域查看配置的规则。

步骤2 配置资源计划。

1. 单击节点默认范围下方的“配置指定时间段的节点数量范围”或者“添加资源计划”。
2. 配置“生效日期”、“时间范围”和“节点数量范围”。

   例如此处“生效日期”设置为周一、周二和周六，“时间范围”设置为“07:00-13:00”，“节点数量范围”设置为“5~8”。

   单击“配置指定时间段的节点数量范围”或者“添加资源计划”按钮配置多条资源计划。

说明

- 生效日期默认是每日生效，也可以选择周一至周日任意一天或几天生效。
- 如果没有配置指定时间段的节点数量范围，则节点数量范围以“默认范围”为准。
- 如果配置了指定时间段的节点数量范围，则在这个时间范围内，以配置的“节点数量范围”为准。不在配置的时间范围时，则以“默认范围”为准。

----结束

4.7.7 修改已有弹性伸缩策略

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 选择“弹性伸缩”页签。

步骤4 在已有弹性伸缩策略右边单击“编辑”按钮。

----结束

4.7.8 删除已有弹性伸缩策略

步骤1 登录MRS管理控制台。

步骤2 选择“现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 选择“弹性伸缩”页签。

步骤4 在已有弹性伸缩策略右边单击“删除”按钮。

----结束

4.7.9 开启或关闭已有弹性伸缩策略

步骤1 登录MRS管理控制台。

步骤2 选择“现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 选择“弹性伸缩”页签。

步骤4 在已有弹性伸缩策略上单击“启用/关闭弹性伸缩策略”。

----结束

4.7.10 查看已有弹性伸缩策略

步骤1 登录MRS管理控制台。

步骤2 选择“现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 选择“弹性伸缩”页签。

步骤4 在已有弹性伸缩策略右边单击“查看”按钮。

----结束

4.7.11 配置自动化脚本

步骤1 登录MRS管理控制台。

步骤2 选择“现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 选择“弹性伸缩”页签。

步骤4 单击“配置自动化脚本”按钮。

步骤5 单击“添加”。

步骤6 配置“名称”、“脚本路径”、“执行节点类型”、“参数”、“执行时机”、“失败操作”。参数详情请参考表**表4-13**。

步骤7 单击“确定”保存自动化脚本配置。

----结束

4.7.12 配置弹性伸缩指标

节点组维度策略

在添加规则时，可以参考**表4-10**配置相应的指标。

表 4-10 弹性伸缩指标列表

集群类型	指标名称	数值类型	说明
流式集群	StormSlotAvailable	整型	Storm组件的可用slot数。 取值范围为[0～2147483646]。
	StormSlotAvailablePercentage	百分比	Storm组件可用slot百分比。是可用slot数与总slot数的比值。 取值范围为[0～100]。
	StormSlotUsed	整型	Storm组件的已用slot数。 取值范围为[0～2147483646]。
	StormSlotUsedPercentage	百分比	Storm组件已用slot百分比。是已用slot数与总slot数的比值。 取值范围为[0～100]。
	StormSupervisorMemAverageUsage	整型	Storm组件Supervisor的内存平均使用量。 取值范围为[0～2147483646]。
	StormSupervisorMemAverageUsagePercentage	百分比	Storm组件Supervisor进程使用的内存占系统总内存的平均百分比。 取值范围[0 ~ 100]。
	StormSupervisorCPUAverageUsagePercentage	百分比	Storm组件Supervisor进程使用的CPU占系统总CPU的平均百分比。 取值范围[0 ~ 6000]。
分析集群	YARNAppPending	整型	YARN组件挂起的任务数。 取值范围为[0～2147483646]。
	YARNAppPendingRatio	比率	YARN组件挂起的任务数比例。是YARN挂起的任务数与YARN运行中的任务数比值。 取值范围为[0～2147483646]。
	YARNAppRunning	整型	YARN组件运行中的任务数。 取值范围为[0～2147483646]。

集群类型	指标名称	数值类型	说明
	YARNContainerAllocated	整型	YARN组件中已分配的container个数。 取值范围为[0～2147483646]。
	YARNContainerPending	整型	YARN组件挂起的container个数。 取值范围为[0～2147483646]。
	YARNContainerPendingRatio	比率	YARN组件挂起的container比率。是挂起的container数与运行中的container数的比值。 取值范围为[0～2147483646]。
	YARNCPUAllocated	整型	YARN组件已分配的虚拟CPU核心数。 取值范围为[0～2147483646]。
	YARNCPUAvailable	整型	YARN组件可用的虚拟CPU核心数。 取值范围为[0～2147483646]。
	YARNCPUAvailablePercentage	百分比	YARN组件可用虚拟CPU核心数百分比。是可用虚拟CPU核心数与总虚拟CPU核心数比值。 取值范围为[0～100]。
	YARNCPUPending	整型	YARN组件挂起的虚拟CPU核心数。 取值范围为[0～2147483646]。
	YARNMemoryAllocated	整型	YARN组件已分配内存大小。单位为MB。 取值范围为[0～2147483646]。
	YARNMemoryAvailable	整型	YARN组件可用内存大小。单位为MB。 取值范围为[0～2147483646]。
	YARNMemoryAvailablePercentage	百分比	YARN组件可用内存百分比。是YARN组件可用内存大小与YARN组件总内存大小的比值。 取值范围为[0～100]。
	YARNMemoryPending	整型	YARN组件挂起的内存大小。 取值范围为[0～2147483646]。

说明

- 表4-10中指标数值类型为百分比或比率时，有效数值可精确到百分位。其中百分比类型指标数值为去除百分号（%）后的小数值，如16.80即代表16.80%。
- 混合集群的支持分析集群和流式集群的所有指标。

资源池维度策略

在添加规则时，可以参考表4-11配置相应的指标。

说明

MRS 3.1.5及其之后的版本支持配置资源池维度策略。

表 4-11 规则配置项说明

集群类型	指标名称	数值类型	说明
分析集群/自定义集群	ResourcePoolMemoryAvailable	整型	资源池YARN组件可用内存大小。单位为MB。 取值范围为[0～2147483646]。
	ResourcePoolMemoryAvailablePercentage	百分比	资源池YARN组件可用内存百分比。是YARN组件可用内存大小与YARN组件总内存大小的比值。 取值范围为[0～100]。
	ResourcePoolCPUAvailable	整型	资源池YARN组件可用的虚拟CPU核心数。 取值范围为[0～2147483646]。
	ResourcePoolCPUAvailablePercentage	百分比	资源池YARN组件可用虚拟CPU核心数百分比。是可用虚拟CPU核心数与总虚拟CPU核心数比值。 取值范围为[0～100]。

在添加资源计划时，可以参考表4-12配置相应的参数。

表 4-12 资源计划配置项说明

配置项	说明
生效日期	资源计划的生效日期。默认是每日生效，也可以选择周一至周日任意一天或几天生效。

配置项	说明
时间范围	资源计划的起始时间和结束时间，精确到分钟，取值范围[00:00, 23:59]。例如资源计划开始于早上8:00，结束于10:00，则配置为8:00-10:00。结束时间必须晚于开始时间至少30分钟。
节点数量范围	资源计划内的节点数量上下限，取值范围[0,500]，在资源计划时间内，集群Task节点数量小于最小节点数时，弹性伸缩会将集群Task节点一次性扩容到最小节点数。在资源计划时间内，集群Task节点数量大于最大节点数时，弹性伸缩会将集群Task节点一次性缩容到最大节点数。最小节点数必须小于或等于最大节点数。

说明

- 当启用资源计划时，弹性伸缩配置中的“默认节点数量范围”将在资源计划外的时间段内强制生效。例如“默认节点数量范围”配置为1-2，配置资源计划：08:00-10:00之间节点数量范围为4-5，则在一天中的非资源计划时间段（0:00-8:00以及10:00-23:59）内，Task节点会被强制限制在1个到2个中间，若节点数量大于2则触发自动缩容，若节点数量小于1则触发自动扩容。
- 当不启用资源计划时，节点数量范围的“默认范围”会在全部时间范围生效，如果节点数量不在“节点数量范围”的默认范围，主动增减Task节点数量到默认范围内。
- 资源计划间时间段不可交叉，时间段交叉意为某个时间点存在两个生效的资源计划，例如配置资源计划1在08:00-10:00生效，资源计划2在09:00-11:00生效，则两个资源计划存在时间段交叉，交叉时间段09:00-10:00。
- 资源计划不能跨天配置，例如如果要配置23:00至次日01:00的资源计划，请配置时间段为23:00-00:00和00:00-01:00的两个资源计划。

自动化脚本

在添加自动化脚本时，可以参考表4-13配置相应参数。

表 4-13 自动化脚本配置说明

配置项	说明
名称	自动化脚本的名称。 只能由数字、英文字符、空格、中划线和下划线组成，且不能以空格开头。 可输入的字符串长度为1～64个字符。 **说明** 同一集群内，不允许配置相同的名称。不同集群之间，可以配置相同的名称。

配置项	说明
脚本路径	脚本的路径。路径可以是OBS文件系统的路径或虚拟机本地的路径。 ● OBS文件系统的路径，必须以s3a://开头，以.sh结尾。例如：s3a://mrs-samples/xxx.sh ● 虚拟机本地的路径，脚本所在的路径必须以‘/’开头，以.sh结尾。例如，安装Zepelin的示例脚本路径如下：/opt/bootstrap/zepelin/zepelin_install.sh
执行节点类型	选择自动化脚本所执行的节点类型。 **说明** ● 如果选择Master节点，您可以通过开关选择是否只在Active Master节点执行此脚本。 ● 如果选择开启此功能，表示只在Active Master节点上执行。如果选择关闭，表示在所有Master节点执行。默认关闭。
参数	自动化脚本参数，支持通过传入以下预定义变量获得弹性伸缩相关信息： ● ${mrs_scale_node_num}：弹性伸缩节点数量，总是正数 ● ${mrs_scale_type}：弹性伸缩类型，扩容为“scale_out”，缩容为“scale_in” ● ${mrs_scale_node_hostnames}：弹性伸缩节点的主机名，多个主机名之间以“,”隔开 ● ${mrs_scale_node_ips}：弹性伸缩节点的IP，多个IP之间以“,”隔开 ● ${mrs_scale_rule_name}：触发弹性伸缩的规则名，如果是资源计划则为“resource_plan”
执行时机	选择自动化脚本执行的时间。支持“扩容前”、“扩容后”、“缩容前”、“缩容后”四种类型。 **说明** 假设执行节点类型中包含Task节点： ● 执行时机为扩容前的脚本不会在将要扩容出的Task节点上执行。 ● 执行时机为扩容后的脚本会在扩容出的Task节点上执行。 ● 执行时机为缩容前的脚本会在即将被删除的Task节点上执行。 ● 执行时机为缩容后的脚本不会在已经被删除的Task节点上执行。
失败操作	该脚本执行失败后，是否继续执行后续脚本和扩缩容操作。 **说明** ● 建议您在调试阶段设置为“继续”，无论此脚本是否执行成功，则集群都能继续扩缩容操作。 ● 若脚本执行失败，请到集群虚拟机机的“/var/log/Bootstrap”路径下查看失败日志。 ● 由于缩容成功不可回滚，缩容后执行的脚本失败操作只能选择“继续”。

说明

自动化脚本只在弹性伸缩时触发，手动调整集群节点时不会运行。

4.8 管理数据连接

4.8.1 配置数据连接

MRS的数据连接是用来管理集群中组件使用的外部源连接，如Hive的元数据使用外部的关系型数据库，可以通过数据连接来关联Hive组件实现。

- 本地元数据：元数据存储于集群内的本地GaussDB中，当集群删除时元数据同时被删除，如需保存元数据，需提前前往数据库手动保存元数据。
- 数据连接：可选择关联与当前集群同一虚拟私有云和子网的RDS服务中的PostgresDB或MySQL数据库，元数据将存储于关联的数据库中，不会随当前集群的删除而删除，多个MRS集群可共享同一份元数据。

说明

不同集群间Hive元数据切换时，MRS当前只对Hive组件自身的元数据数据库中的权限进行同步。这是由于当前MRS上的权限模型是在Manager上维护的，所以不同集群间的Hive元数据切换，不能自动把用户/用户组的权限同步到另一个集群的Manager上。

数据连接前置操作

步骤1 登录RDS管理控制台。

步骤2 选择“实例管理”，选择MRS数据连接使用的RDS实例所在行的操作列的“更多 > 登录”，以root用户登录该实例。

步骤3 在实例“首页”即可单击“新建数据库”创建新的数据库。

步骤4 在页面顶部选择“账号管理 > 用户管理”。

说明

当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**步骤4**-**步骤6**操作。

步骤5 单击“新建用户”，创建一个非root用户。

步骤6 在页面顶部选择“SQL操作 > SQL查询”，在“库名”处切换对应数据库，然后执行如下SQL命令为该数据库用户进行赋权，其中${db_name}与${db_user}为MRS待连接的数据库名和新建的用户名。

```
grant SELECT, INSERT on mysql.* to '${db_user}'@'%' with grant option;
grant all privileges on ${db_name}.* to '${db_user}'@'%' with grant option;
grant reload on *.* to '${db_user}'@'%' with grant option;
flush privileges;
```

步骤7 参考**创建数据连接**创建数据连接。

----结束

创建数据连接

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 单击“新建数据连接”。

步骤3 参考**表4-14**配置相关参数。

表 4-14 数据连接

参数	说明
类型	选择外部源连接的类型。 • RDS服务MySQL数据库，支持Hive或Ranger组件的集群支持连接该类型数据库。
名称	数据连接的名称。
数据库实例	RDS服务数据库实例，该实例需要先在RDS服务创建后在此处引用，且已创建数据库，具体请参考**数据连接前置操作**。单击“查看数据库实例”查看已创建的实例。 **说明** • 为了保证集群和PostgreSQL数据库的网络访问，建议该实例与MRS集群的虚拟私有云和子网一致。 • 该实例的安全组入方向规则需要放通3306端口（可通过在RDS控制台单击实例名称进入实例基本信息页面，在“连接信息”区域单击“安全组”右侧的安全组名称进入安全组控制台，在入方向规则页签中添加一个“协议端口”为TCP 3306，“源地址”为Hive的MetaStore实例所在的所有节点IP的规则）。 • 当前MRS支持的RDS上Postgres数据库版本号为**PostgreSQL9.5/PostgreSQL9.6**。 • 当前MRS仅支持RDS上MySQL数据库版本为**MySQL 5.7.x**。
数据库	待连接的数据库的名称。
用户名	登录待连接的数据库的用户名。
密码	登录待连接的数据库的密码。

说明

当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**操作。

步骤4 单击“确定”完成创建。

----结束

编辑数据连接

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 在数据连接列表的“操作列”，单击待编辑数据连接所在行的“编辑 ”。

步骤3 参考表4-14修改参数。

如果选择的数据连接已经关联了集群，编辑后会将修改后的配置同步到对应的集群中。

----结束

删除数据连接

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 在数据连接列表的操作列，单击待删除数据连接所在行的“删除 ”。

如果选择的数据连接已经关联了集群，删除动作不会影响对应的集群。

----结束

创建集群时配置数据连接

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“自定义创建”。

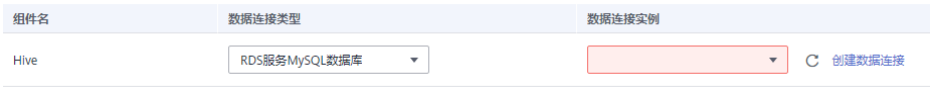
步骤4 在软件配置中，参考表4-15配置“元数据 ”，其他参数请参考创建自定义集群进行配置并创建集群。

表 4-15 数据连接参数说明

参数	参数说明
元数据	是否使用外部数据源存储元数据。 ● 本地元数据： 元数据存储在集群本地。 ● 数据连接：使用外部数据源元数据，若集群异常或删除时将不影响元数据，适用于存储计算分离的场景。 支持Hive或Ranger组件的集群支持该功能。
组件名	当“数据连接”功能开启时，该参数有效。用于表示可以设置外部数据源的组件类型。 ● Hive ● Ranger

参数	参数说明
数据连接类型	当“数据连接”功能开启时，该参数有效。用于表示外部数据源的类型。 ● Hive组件支持的数据连接类型： – RDS服务MySQL数据库 – 本地数据库 ● Ranger组件支持的数据连接类型： – RDS服务MySQL数据库 – 本地数据库
数据连接实例	当“数据连接类型”选择“RDS服务PostgreSQL数据库”或“RDS服务MySQL数据库”时，该参数有效。用于表示MRS集群与RDS服务数据库连接的名称，该实例必须先创建才能在此处引用。可单击“创建数据连接”进行创建，具体请参考**数据连接前置操作**和**创建数据连接**进行操作。

图 4-4 创建集群时配置数据连接



----结束

4.8.2 配置 RDS 数据连接操作

数据连接前置操作

步骤1 登录RDS管理控制台。

步骤2 购买RDS实例。

步骤3 在RDS管理控制台，选择“实例管理”，选择MRS数据连接使用的RDS实例所在行的操作列的“更多 > 登录”，以root用户登录该实例。

步骤4 在实例“首页”即可单击“新建数据库”创建新的数据库。

说明

如果没有创建新的数据库，会导致MRS配置数据连接失败。

步骤5 在页面顶部选择“账号管理 > 用户管理”。

说明

当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**步骤5**-**步骤7**操作。

步骤6 单击“新建用户”，创建一个非root用户。

步骤7 在页面顶部选择“SQL操作 > SQL查询”，在“库名”处切换对应数据库，然后执行如下SQL命令为该数据库用户进行赋权，其中${db_name}与${db_user}为MRS待连接的数据库名和新建的用户名。

```
grant SELECT, INSERT on mysql.* to '${db_user}'@'%' with grant option;
grant all privileges on ${db_name}.* to '${db_user}'@'%' with grant option;
grant reload on *.* to '${db_user}'@'%' with grant option;
flush privileges;
```

步骤8 参考**创建RDS数据连接**创建数据连接。

----结束

创建 RDS 数据连接

该步骤指导用户为当前已有的MRS集群，创建RDS数据连接。

步骤1 登录MRS控制台，在导航栏选择“数据连接”。

步骤2 单击“新建数据连接”。

步骤3 参考**表4-16**配置相关参数。

表 4-16 数据连接

参数	说明
类型	选择外部源连接的类型。 • RDS服务MySQL数据库，支持Hive或Ranger组件的集群支持连接该类型数据库。
名称	数据连接的名称。
数据库实例	RDS服务数据库实例，该实例需要先在RDS服务创建后在此处引用，且已创建数据库，具体请参考**数据连接前置操作**。单击“查看数据库实例”查看已创建的实例。 **说明** • 为了保证集群和PostgreSQL数据库的网络访问，建议该实例与MRS集群的虚拟私有云和子网一致。 • 该实例的安全组入方向规则需要放通3306端口（可通过在RDS控制台单击实例名称进入实例基本信息页面，在“连接信息”区域单击“安全组”右侧的安全组名称进入安全组控制台，在入方向规则页签中添加一个“协议端口”为TCP 3306，“源地址”为Hive的MetaStore实例所在的所有节点IP的规则）。 • 当前MRS支持的RDS上Postgres数据库版本号为**PostgreSQL9.5/PostgreSQL9.6**。 • 当前MRS仅支持RDS上MySQL数据库版本为**MySQL 5.7.x**。
数据库	待连接的数据库的名称。
用户名	登录待连接的数据库的用户名。
密码	登录待连接的数据库的密码。

说明

当用户选择的数据连接为“RDS服务MySQL数据库”或“云数据库GaussDB(for MySQL)”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**操作。

步骤4 单击“确定”完成创建。

----结束

创建集群时配置 RDS 数据连接

该操作指导用户在创建MRS集群时配置RDS数据连接。

步骤1 登录MRS管理控制台。

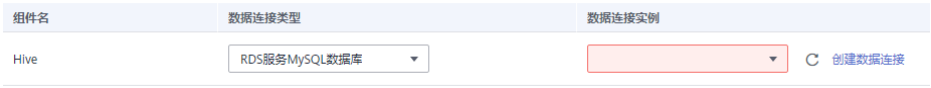
步骤2 在创建集群页面，选择“自定义创建”。

步骤3 在软件配置中，参考**表4-17**配置“元数据 ”，其他参数请参考**创建自定义集群**进行配置并创建集群。

表 4-17 数据连接参数说明

参数	参数说明
元数据	选择“外置数据连接”。使用外部数据源元数据，若集群异常或删除时将不影响元数据，适用于存储计算分离的场景。 支持Hive或Ranger组件的集群支持该功能。
组件名	当“外置数据连接”功能开启时，该参数有效。用于表示可以设置外部数据源的组件类型。 ● Hive ● Ranger
数据连接类型	当“外置数据连接”功能开启时，该参数有效。用于表示外部数据源的类型。 ● Hive组件支持的数据连接类型： – RDS服务MySQL数据库 – 本地数据库 ● Ranger组件支持的数据连接类型： – RDS服务MySQL数据库 – 本地数据库
数据连接实例	当“数据连接类型”选择“本地数据库”时无需配置该参数。用于表示MRS集群与RDS服务数据库连接的名称，该实例必须先创建才能在此处引用。可单击“创建数据连接”进行创建，具体请参考**数据连接前置操作**和**创建RDS数据连接**进行操作。

图 4-5 创建集群时配置数据连接



----结束

4.8.3 配置 Ranger 数据连接

本指导旨在指导用户将现有集群的Ranger元数据切换为RDS数据库中存储的元数据。该操作可以使多个MRS集群共用同一份元数据，且元数据不随集群的删除而删除。也能够避免集群迁移时Ranger元数据的迁移。

前置条件

已创建RDS服务MySQL数据库的实例，请参考**创建数据连接**。

说明

- 对于MRS 3.x之前版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。
- 对于MRS 3.x及之后版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，数据库用户不允许为root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。

Ranger 元数据外置到 Mysql 前置操作

该前置操作仅在**MRS 3.1.0及之后版本**需要执行。

步骤1 登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。选择“集群 > 服务 > *服务名称*”。

当前MRS 3.1.x集群支持Ranger鉴权的组件为: HDFS、HBase、Hive、Spark、Impala、Storm、Kafka组件。

步骤2 在服务“概览”页面右上角单击“更多 > 停用Ranger鉴权”，如果“停用Ranger鉴权”是灰色，则表示未开启Ranger鉴权无需停用Ranger鉴权，如**图4-6**所示。

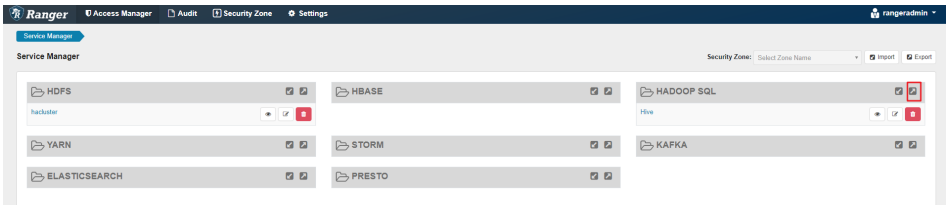
图 4-6 停用 Ranger 鉴权

步骤3 （可选）如需使用已有鉴权策略请执行该步骤在Ranger Web页面导出已有组件的鉴权策略，切换Ranger元数据完成后可重新导入已有的鉴权策略。此处以Hive为例，导出后会生成本地的json格式的策略文件。

1. 登录FusionInsight Manager页面。
2. 选择“集群 > 服务 > Ranger”，进入Ranger服务概览页面。
3. 单击“基本信息”区域中的“RangerAdmin”，进入Ranger WebUI界面。

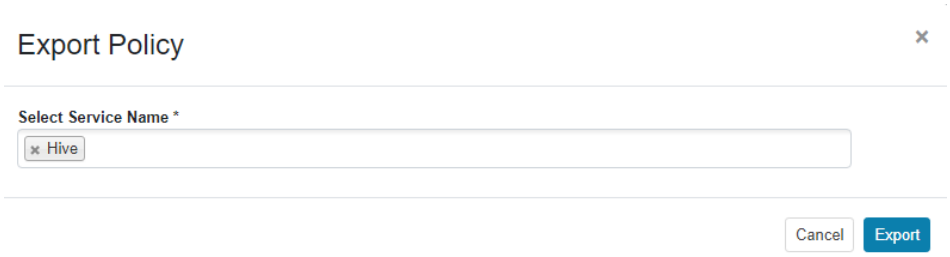
   admin用户在Ranger中的用户类型为“User”，如需查看所有管理页面，可单击右上角用户名后，选择“Log Out”，退出当前用户。
4. 使用**rangeradmin**用户（默认密码为Rangeradmin@123）或者其他具有Ranger管理员权限用户重新登录。
5. 单击Hive组件对应行的导出按钮，导出鉴权策略。

   图 4-7 导出鉴权策略

6. 单击“Export”，导出后会生成本地的json格式的策略文件。

图 4-8 导出 Hive 鉴权策略



----结束

为 MRS 集群配置数据连接

步骤1 登录MRS控制台。

步骤2 单击集群名称进入集群详情页面。

步骤3 单击“数据连接”右侧的“单击管理”，进入数据连接配置界面。

步骤4 单击“配置数据连接”，并配置相关参数。

- 组件名称：Ranger
- 模块类型：Ranger元数据
- 连接类型：RDS服务MySQL数据库
- 连接实例：请选择已创建的到RDS服务MySQL数据库的实例，如需创建新的数据连接，请参考**创建数据连接**。

步骤5 勾选“我已经阅读上述信息，并了解具体影响。”并单击“测试”。

步骤6 测试成功后，单击“确定”完成数据连接配置。

步骤7 登录FusionInsight Manager页面。

步骤8 选择“集群 > 服务 > Ranger”，进入Ranger服务概览页面。

步骤9 单击“更多 > 重启服务”或“更多 > 滚动重启服务”。

重启服务会造成业务中断，滚动重启可以尽量减少或者不影响业务运行。

重启Ranger组件会影响所有受Ranger控制组件的权限，可能影响业务的正常运行，请在集群空闲或业务量较少时时执行重启。重启Ranger组件前，Ranger中的策略依然生效。

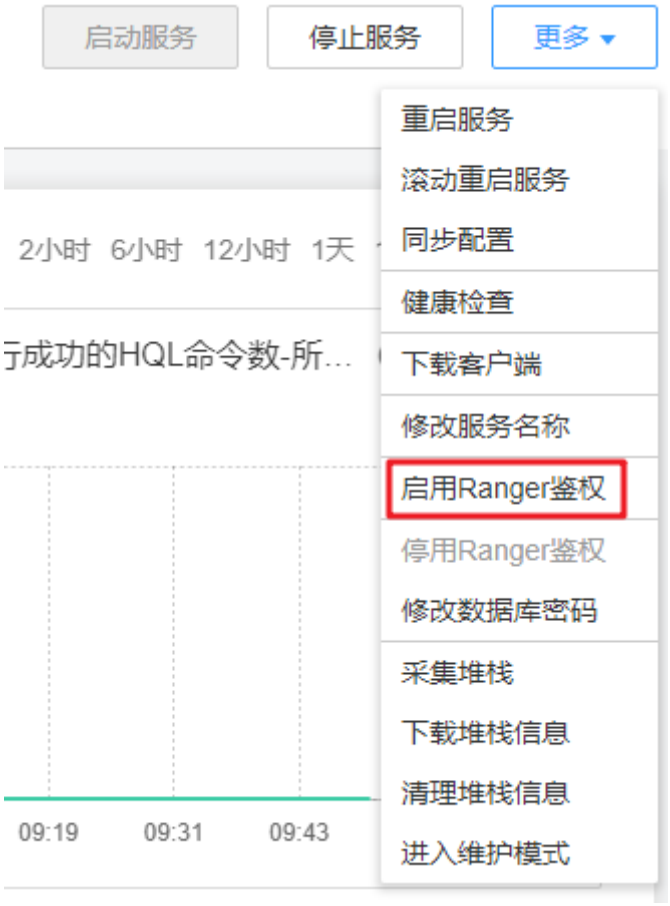
图 4-9 重启服务

步骤10 启用需要鉴权的组件的Ranger鉴权。此处以Hive组件为例。

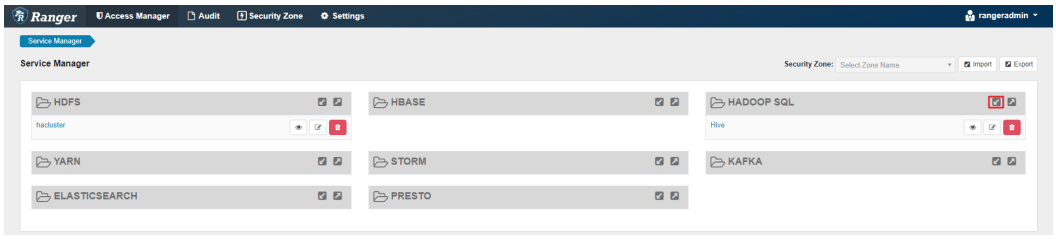
当前MRS 3.1.x集群支持Ranger鉴权的组件为: HDFS、HBase、Hive、Spark、Impala、Storm、Kafka组件。

1. 登录FusionInsight Manager页面，选择“集群 > 服务 > *服务名称*”。
2. 在服务“概览”页面右上角单击“更多 > 启用Ranger鉴权”。

图 4-10 启用 Ranger 鉴权



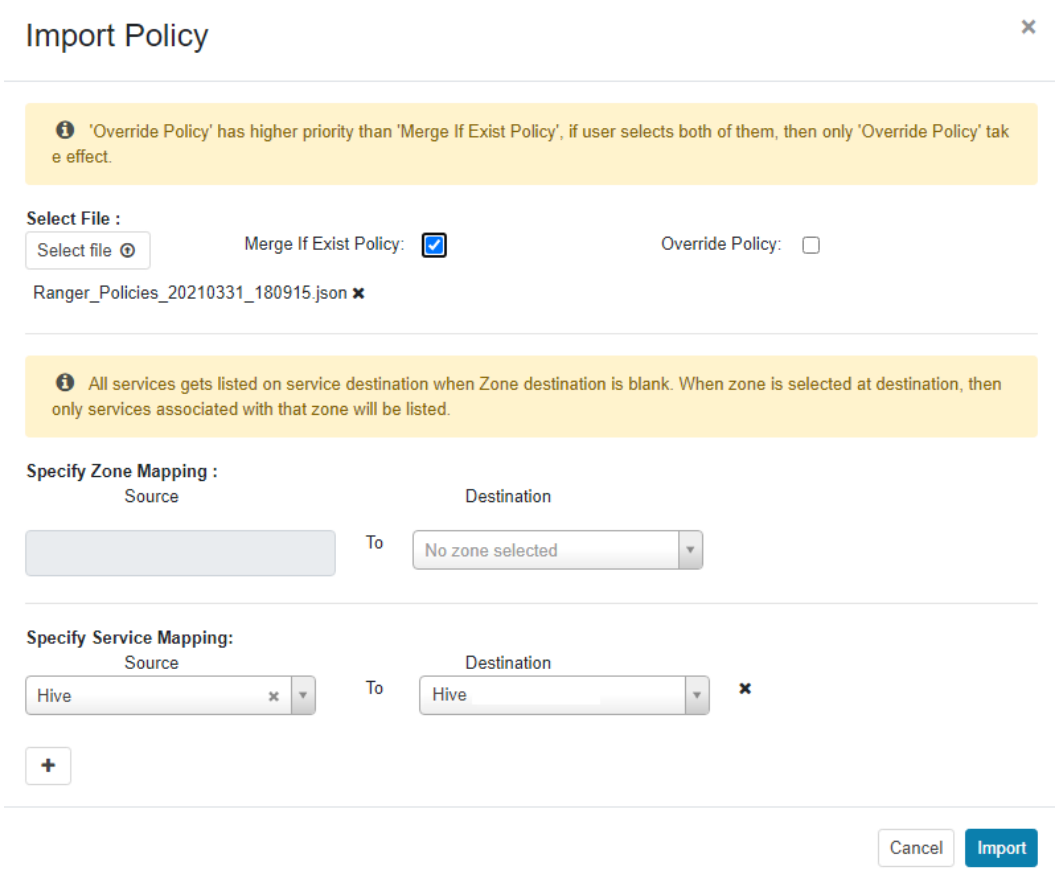
步骤11 登录Ranger WebUI界面，单击Hive组件对应行的导入按钮。



步骤12 配置导入参数。

- Select file：选择**步骤3.6**中下载的鉴权策略文件。
- Merge If Exist Policy：勾选。

图 4-11 导入鉴权策略



步骤13 重启启用Ranger鉴权的组件。

1. 登录FusionInsight Manager页面。
2. 选择“集群 > 服务 > Hive”，进入Hive服务概览页面。
3. 单击“更多 > 重启服务”或“更多 > 滚动重启服务”。
   重启服务会造成业务中断，滚动重启可以尽量减少或者不影响业务运行。

----结束

4.8.4 配置 Hive 数据连接

本章节指导用户在创建后，将现有集群的Hive元数据切换为本地数据库或者RDS数据库中存储的元数据。该操作可以使多个MRS集群共用同一份元数据，且元数据不随集群的删除而删除。也能够避免集群迁移时Hive元数据的迁移。

说明

- 不同集群间Hive元数据切换时，MRS当前只对Hive组件自身的元数据数据库中的权限进行同步。这是由于当前MRS上的权限模型是在Manager上维护的，所以不同集群间的Hive元数据切换，不能自动把用户/用户组的权限同步到另一个集群的Manager上。
- 对于MRS 3.x之前版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，请确保使用的数据库用户为root用户。如果为非root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。
- 对于MRS 3.x及之后版本，当用户选择的数据连接为“RDS服务MySQL数据库”时，数据库用户不允许为root用户，请参考**数据连接前置操作**新建用户并为该用户进行赋权。

配置 Hive 数据连接

该功能在MRS 3.0.5版本暂不支持。

步骤1 登录MRS控制台，在导航栏选择“集群列表 > 现有集群”。

步骤2 单击集群名称，进入集群详情页面。

步骤3 在集群详情页的“概览”页签，单击“数据连接”右侧的“单击管理”。

步骤4 在“数据连接”页面显示集群已关联的数据连接，单击“编辑”或“删除”可对数据连接进行编辑或删除。

步骤5 若“数据连接”页面没有关联连接，单击“配置数据连接”进行增加。

说明

一种模块类型只能配置一个数据连接，如在Hive元数据上配置了数据连接后，不能再配置其他的数据连接。当没有可用的模块类型时，“配置数据连接”按钮不可用。

表 4-18 配置 Hive 数据连接

参数	说明
组件名称	Hive
模块类型	Hive元数据
连接类型	• RDS服务MySQL数据库 • 本地数据库
连接实例	当“连接类型”参数选择“RDS服务PostgreSQL数据库”或“RDS服务MySQL数据库”时有效。选择MRS集群与RDS服务数据库连接名称，该连接必须先创建才能在此处引用。可单击“创建数据连接”进行创建，具体请参考**创建数据连接**。

步骤6 单击“测试”，测试此数据连接和集群的连通性。

步骤7 连接成功后单击“确定” 完成配置数据连接。

说明

- 配置了Hive元数据后，请重启Hive服务，Hive会在指定的数据库下创建Hive必须的数据库表（如表已经存在则不会创建）。
- 重启Hive服务前，请确保已安装对应驱动包到所有MetaStore实例所在节点中。
  - Postgres：使用开源驱动包替换集群已有的驱动包。将postgres驱动包postgresql-42.2.5.jar上传至所有MetaStore实例节点${BIGDATA_HOME}/third_lib/Hive目录下（开源驱动包下载地址：**https://repo1.maven.org/maven2/org/postgresql/postgresql/42.2.5/**）。
  - MySQL：进入MySQL官网（https://www.mysql.com/），选择“Downloads > Community > MySQL Connectors > Connector/J”下载对应版本的驱动包，将MySQL对应版本的驱动包上传至所有Metastore实例节点“/opt/Bigdata/FusionInsight_HD_*/install/FusionInsight-Hive-*/hive-*/lib/”目录下。

----结束

4.9 通过引导操作安装第三方软件

该操作适用于MRS 3.x之前版本集群。

MRS 3.x版本暂不支持在创建集群时添加引导操作。

前提条件

已准备好引导操作脚本，详情请参见**准备引导操作脚本**。

在创建集群时添加引导操作

步骤1 登录MRS管理控制台。

步骤2 单击“创建集群”，进入“创建集群”页面。

步骤3 在创建集群页面，选择“自定义创建”。

步骤4 参考**创建自定义集群**配置集群软件配置和硬件配置信息。

步骤5 在“高级配置”区域的引导操作栏，单击“添加”。

表 4-19 参数描述

参数	说明
名称	引导操作脚本的名称。 只能由数字、英文字符、空格、中划线和下划线组成，且不能以空格开头。 可输入的字符串长度为1～64个字符。 **说明** 同一集群内，不允许配置相同的名称。不同集群之间，可以配置相同的名称。
脚本路径	脚本的路径。路径可以是OBS文件系统的路径或虚拟机本地的路径。 • OBS文件系统的路径，必须以s3a://开头，以.sh结尾。例如：s3a://mrs-samples/xxx.sh • 虚拟机本地的路径，脚本所在的路径必须以‘/’开头，以.sh结尾。 **说明** 同一集群内，不允许配置相同的路径。不同集群之间，可以配置相同的路径。
参数	引导操作脚本参数。
执行节点	选择引导操作脚本所执行的节点类型。
执行时机	选择引导操作脚本执行的时间。 • 组件首次启动前 • 组件首次启动后

参数	说明
失败操作	该脚本执行失败后，是否继续执行后续脚本和创建集群。 **说明** 建议您设置为“继续”，无论此引导操作是否执行成功，则集群都能继续创建。

步骤6 单击“确定”。

添加成功后，可以通过“操作”列进行编辑、克隆和删除。

----结束

在弹性伸缩集群页面添加自动化脚本

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称。进入集群详情页面。

步骤3 在“节点管理”页签Task节点组的“操作”列单击“弹性伸缩”，进入“弹性伸缩”页面。

当集群没有Task节点时，先单击“配置Task节点”增加Task节点，再执行该步骤。

说明

对于MRS 3.x及之后版本，“配置Task节点”仅适用于分析集群、流式集群和混合集群。

步骤4 配置资源计划。

配置操作：

1. 在弹性伸缩页面，开启弹性伸缩功能。
2. 节点数量范围的“默认范围”设置为“2-2”，表示除资源计划规定时间范围外，其他时间Task节点数量固定为2个。
3. 单击默认范围下方的“配置指定时间段的节点数量范围”。
4. 配置“时间范围”和“节点数量范围”。此处“时间范围”设置为“07:00-13:00”，“节点数量范围”设置为“5-5”，表示在该时间范围内，Task节点数量固定为5个。参数详情请参考**表4-12**。

   可以单击“配置指定时间段的节点数量范围”配置多条资源计划。

步骤5 （可选）配置自动化脚本。

1. 在“高级配置”项选择“现在配置”。
2. 单击“添加”，进入“自动化脚本”配置页面。
3. 配置“名称”、“脚本路径”、“执行节点类型”、“参数”、“执行时机”、“失败操作”。参数详情请参考**表4-13**。
4. 单击“确定”保存自动化脚本配置。

步骤6 勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

步骤7 单击“确定”，完成弹性伸缩集群设置。

----结束

4.10 查看失败的集群操作任务

本章节介绍如何查看并删除失败的MRS任务。

背景信息

当集群创建失败、集群删除失败、集群扩容失败和集群缩容失败后，失败任务会转入“失败任务管理”页面，其中仅集群删除失败的任务会同步转入“历史集群”页面。当不需要失败的任务时，可以删除。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”。

步骤3 单击“失败任务管理”右侧的或数字，进入“失败任务管理”页面。

步骤4 在需要删除的任务对应的“操作”列中，单击“删除”。

此处只能删除单个失败的任务。

步骤5 单击任务列表左上方的“删除所有失败任务”可以删除全部任务。

----结束

4.11 查看历史集群信息

选择“集群列表 > 历史集群”，选中一集群并单击集群名，进入集群基本信息页面。用户可查看集群的配置信息、部署的节点信息。

参考下列表格查看集群信息参数说明。

表 4-20 集群基本信息

参数	参数说明
集群名称	集群的名称，创建集群时设置。
集群状态	集群状态信息。
集群版本	集群的版本信息。
集群类型	创建集群时的集群类型。
集群ID	集群的唯一标识，创建集群时系统自动赋值，不需要用户设置。
创建时间	显示集群创建的时间。
可用区	集群工作区域下的可用区，创建集群时设置。

参数	参数说明
默认生效子网	子网信息，创建集群时所选。 通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。
虚拟私有云	VPC信息，创建集群时所选。 VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。
OBS权限控制	单击“单击管理”，修改MRS用户与OBS权限的映射关系，具体请参考**配置MRS多用户访问OBS细粒度权限**。
数据连接	单击“单击管理”，查看集群关联的数据连接类型，具体请参考**配置数据连接**。
委托	单击“管理委托”，为集群绑定或修改委托。 通过绑定委托，您可以将部分资源共享给ECS或BMS云服务来管理，例如通过配置ECS委托可自动获取AK/SK访问OBS，具体请参见**配置存算分离集群（委托方式）**。 **MRS_ECS_DEFAULT_AGENCY**委托拥有对象存储服务的OBSOperateAccess权限和在集群所在区域拥有CESFullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。
密钥对	密钥对名称，创建集群时设置。 如果创建集群时设置的登录方式为密码，则不显示。
Kerberos认证	登录Manager管理页面时是否启用Kerberos认证。 **说明** Kerberos认证模式不支持手动修改，集群创建成功后将无法开启和关闭此功能，需要在创建MRS服务集群的时候选择开启或者关闭Kerberos服务，建议重新创建集群。
安全组	集群的安全组名称。
数据盘密钥名称	用于加密数据盘的密钥名称。如需对已使用的密钥进行管理，请登录密钥管理控制台进行操作。
数据盘密钥ID	用于加密数据盘的密钥ID。
组件版本	集群安装各组件的版本信息。
委托	通过绑定委托，ECS或BMS云服务将有权限来管理您的部分资源。

返回到历史集群列表页面，用户可使用如下按钮进行操作，参考下列表格查看按钮说明。

表 4-21 按钮说明

按钮	说明
	单击 ，手动刷新节点信息。
	在搜索框中输入集群名称或ID，单击 ，搜索集群。

5 管理集群

5.1 登录集群

5.1.1 MRS 集群节点简介

介绍远程登录的概念、MRS集群的节点类型和节点功能。

MRS集群节点支持用户远程登录，远程登录包含界面登录和SSH登录两种方式：

- 界面登录：直接通过弹性云服务器管理控制台提供的远程登录功能，登录到集群Master节点的Linux界面。操作步骤，请参见**登录集群节点**。
- SSH登录：仅适用于Linux弹性云服务器。您可以使用远程登录工具（例如PuTTY），登录弹性云服务器。此时，需要该弹性云服务器绑定弹性IP地址。

  Master节点申请和绑定弹性IP，请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

  可以使用密钥方式也可以使用密码方式登录Linux弹性云服务器。

须知

当您使用密钥方式访问MRS1.6.2版本（不包含）前的集群节点，需要以Linux用户登录，详细步骤请参见**登录弹性云服务器（SSH密钥方式）**。

当您使用密钥方式访问MRS1.6.2版本（包含）及以后的集群节点，需要以root用户登录，详细步骤请参见**登录弹性云服务器（SSH密钥方式）**。

MRS集群中每个节点即为一台弹性云服务器，节点类型及节点功能如**表5-1**所示。

表 5-1 集群节点分类

节点类型	功能
Master节点	MRS集群管理节点，负责管理和监控集群。在MRS管理控制台选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群信息页面。在“节点管理”中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。 Master节点可以通过弹性云服务器界面的VNC方式登录，也可以通过SSH方式登录，并且Master节点可以免密码登录到Core节点。 系统自动将Master节点标记为主备管理节点，并支持MRS集群管理的高可用特性。如果主管理节点无法提供服务，则备管理节点会自动切换为主管理节点并继续提供服务。 查看Master1节点是否为主管理节点，请参见**如何确认Manager的主备管理节点**。
Core节点	MRS集群工作节点，负责处理和分析数据，并存储过程数据。
Task节点	计算节点，用于弹性伸缩，集群计算资源不足时扩容至集群中。

5.1.2 登录集群节点

本章节介绍如何使用弹性云服务器管理控制台上提供的远程登录（VNC方式）和如何使用密钥或密码方式（SSH方式）登录MRS集群中的节点，远程登录主要用于紧急运维场景，远程登录弹性云服务器进行相关维护操作。其他场景下，建议用户采用SSH方式登录。

说明

如果需要使用SSH方式登录集群节点，需要在集群的安全组规则中手动添加入方向规则：其中源地址为“客户端IPV4地址/32(或者客户端IPV6地址/128)”，端口为22，具体请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

登录弹性云服务器（VNC 方式）

步骤1 登录MapReduce服务管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群基本信息页面。

步骤3 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。

步骤4 单击右上角的“远程登录”。

步骤5 根据界面提示，输入Master节点的用户名和密码。

1. 创建集群时登录方式选择了“密码”。此时，你需要输入的用户名、密码分别是root和创建集群时设置的密码。

2. 创建集群时登录方式选择了密钥对，则使用如下方式登录：
   a. 创建集群成功后，为集群的Master节点绑定一个弹性IP地址。
   b. 使用root用户名和密钥文件，SSH方式远程登录Master节点。
   c. 执行**passwd root**命令，设置root用户密码。
   d. 设置成功后，返回界面登录方式，输入root用户名和**步骤5.2.c**设置的密码，登录节点。

步骤6 参考**远程登录（VNC方式）**进行后续操作。

----结束

登录弹性云服务器（SSH 密钥方式）

本地使用Windows操作系统

如果您本地使用Windows操作系统登录Linux弹性云服务器，可以按照下面方式登录弹性云服务器。下面步骤以PuTTY为例。

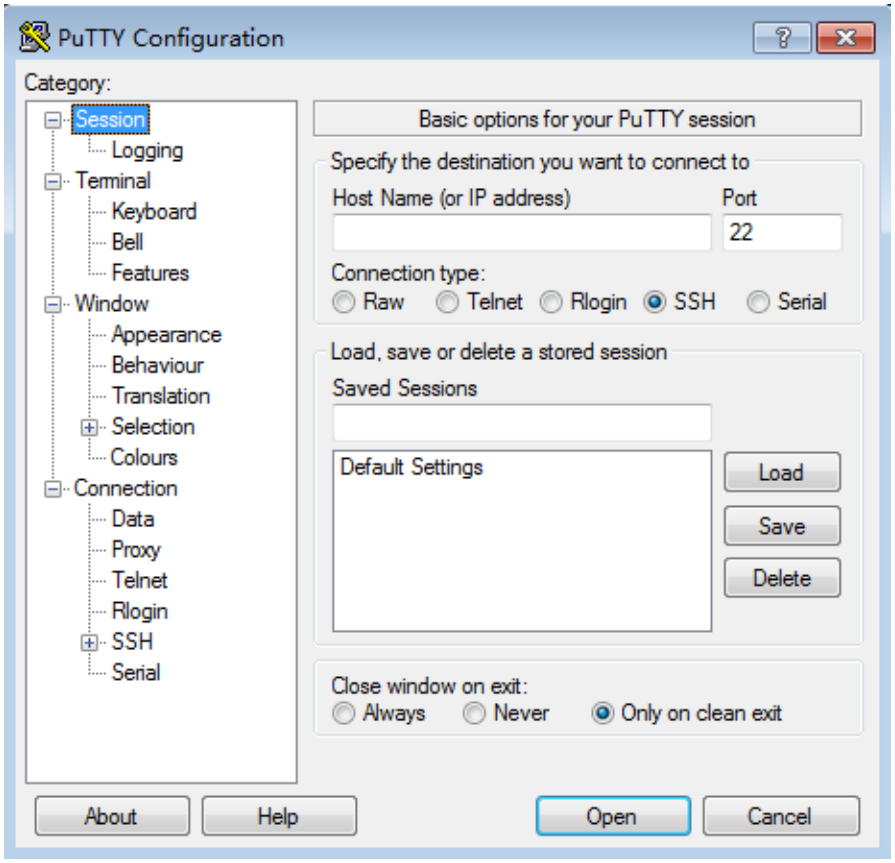
1. 登录MapReduce服务管理控制台。
2. 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群基本信息页面。
3. 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。
4. 选择“弹性公网IP”页签，单击“绑定弹性IP”为弹性云服务器绑定一个弹性公网IP并记录该IP地址，若已绑定弹性公网IP请跳过该步骤。
5. 判断私钥文件是否为.ppk格式。
   - 是，执行**10**。
   - 否，执行**6**。
6. 运行PuTTY。
7. 在“Actions”区域，单击“Load”，并导入创建弹性云服务器时使用的密钥对的私钥文件。

   导入时注意确保导入的格式要求为“All files (*.*)”。
8. 单击“Save private key”。
9. 保存转化后的私钥到本地。例如：kp-123.ppk。
10. 运行PuTTY。
11. 选择“Connection > data”，在Auto-login username处输入镜像的用户名。

    说明

    集群节点镜像的用户名是root。
12. 选择“Connection > SSH > Auth”，在最下面一个配置项“Private key file for authentication”中，单击“Browse”，选择**9**转化的密钥。
13. 单击“Session”。
    a. Host Name (or IP address)：输入弹性云服务器所绑定的弹性公网IP。
    b. Port：输入 22。
    c. Connection Type：选择 SSH。
    d. Saved Sessions：任务名称，在下一次使用putty时就可以单击保存的任务名称，即可打开远程连接。

图 5-1 单击“Session”



14. 单击“Open”登录云服务器。

    如果首次登录云服务器，PuTTY会显示安全警告对话框，询问是否接受服务器的安全证书。单击“是”将证书保存到本地注册表中。

本地使用Linux操作系统

如果您本地使用Linux操作系统登录Linux弹性云服务器，可以按照下面方式登录。下面步骤以私钥文件以kp-123.pem为例进行介绍。

1. 在您的linux计算机的命令行中执行如下命令，变更权限。

   chmod 400 /*path*/kp-123.pem

   说明

   上述令的path为密钥文件的存放路径。

2. 执行如下命令，登录弹性云服务器。

   ssh -i /*path*/kp-123.pem*默认用户名@弹性公网IP*

   假设Linux弹性云服务器的默认用户名是root，弹性公网IP为123.123.123.123，则命令如下：

   ssh -i /*path*/kp-123.pem root@123.123.123.123

说明

- path为密钥文件的存放路径。
- 弹性公网IP地址为弹性云服务器绑定的弹性公网IP地址。
- MRS1.6.2版本（不包含）前的集群节点，镜像的用户名是Linux。
- MRS1.6.2版本（包含）及以后的集群节点，镜像的用户名是root。

登录弹性云服务器（SSH 密码方式）

本地使用Windows操作系统

如果本地主机为Windows操作系统，可以按照下面方式登录弹性云服务器。下面步骤以PuTTY为例。

步骤1 登录MapReduce服务管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群基本信息页面。

步骤3 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。

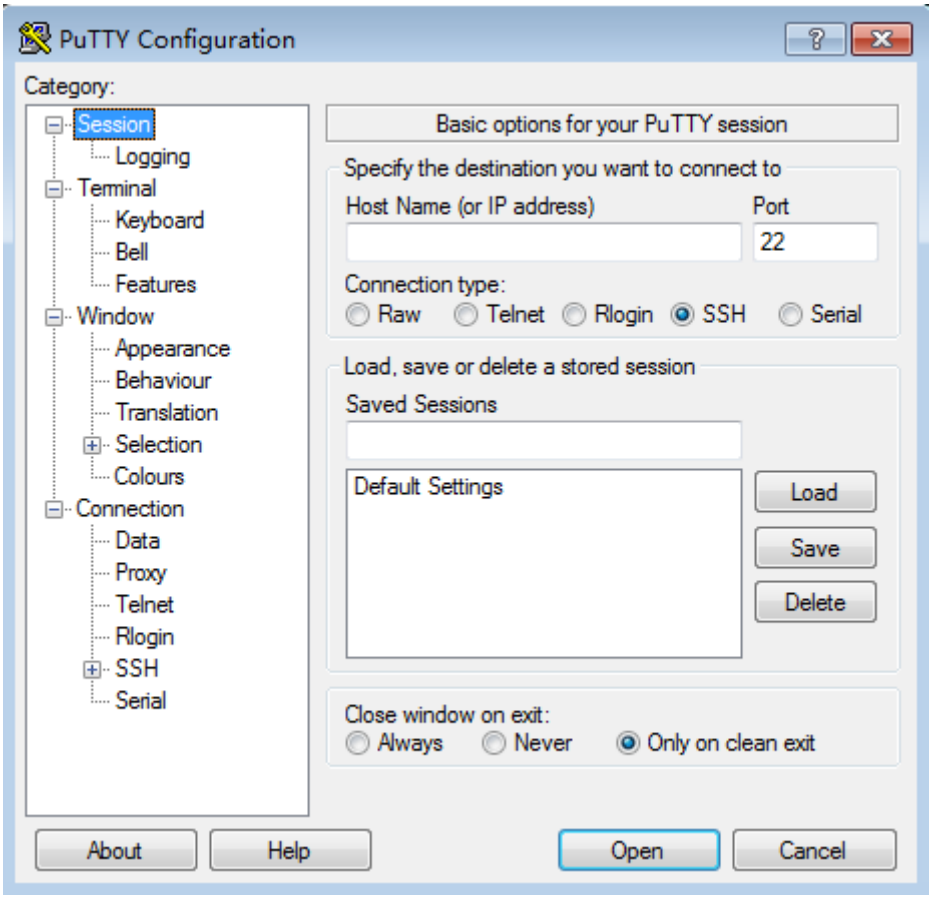
步骤4 选择“弹性公网IP”页签，单击“绑定弹性IP”为弹性云服务器绑定一个弹性公网IP并记录该IP地址，若已绑定弹性公网IP请跳过该步骤。

步骤5 运行PuTTY。

步骤6 单击“Session”。

1. Host Name (or IP address)：输入弹性云服务器所绑定的弹性公网IP。
2. Port：输入 22。
3. Connection Type：选择 SSH。
4. Saved Sessions：任务名称，在下一次使用PuTTY时就可以单击保存的任务名称，即可打开远程连接。

图 5-2 单击 Session



步骤7 单击“Window”，在“Translation”下的“Remote character set:”选择“UTF-8”。

步骤8 单击“Open”登录云服务器。

如果首次登录云服务器，PuTTY会显示安全警告对话框，询问是否接受服务器的安全证书。单击“是”将证书保存到本地注册表中。

步骤9 建立到云服务器的SSH连接后，根据提示输入用户名和密码登录弹性云服务器。

说明

用户名、密码分别是root和创建集群时设置的密码。

----结束

本地使用Linux操作系统

如果本地主机为Linux操作系统，您可以参考步骤1~步骤4为弹性云服务器绑定弹性公网IP后，在计算机的命令行中运行如下命令登录弹性云服务器：**ssh *弹性云服务器绑定的弹性公网IP***

修改操作系统键盘语言

登录的MRS集群节点的操作系统均为Linux操作系统，操作系统键盘语言的修改方法，请参见弹性云服务器《用户指南》中“入门 > 登录弹性云服务器 > 登录弹性云服务器（VNC方式）”。

5.1.3 如何确认 Manager 的主备管理节点

介绍如何在Master1节点中确认Manager的主备管理节点。

背景信息

用户可以在Master节点登录到集群中的其他节点，同时登录Master节点后，可以确认Manager的主备管理节点，并在对应的管理节点中执行命令。

在主备模式下，由于Master1和Master2之间会切换，Master1节点不一定是Manager的主管理节点。

操作步骤

步骤1 确认MRS集群的Master节点。

1. 登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群信息页面。查看指定的集群信息。
2. 在“节点管理”中查看节点名称，名称中包含“master1”的节点为Master1节点，名称中包含“master2”的节点为Master2节点。

步骤2 确认Manager的主备管理节点。

1. 远程登录Master1节点，请参见**登录集群节点**。
2. 执行以下命令切换用户。

   sudo su - root

   su - omm

3. 执行以下命令确认主备管理节点：

   MRS 3.x之前版本集群执行命令：sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh

   MRS 3.x及之后版本集群执行命令：sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh

   回显信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“mgtomsdat-sh-3-01-1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“mgtomsdat-sh-3-01-2”为备管理节点）。

```
Ha mode
double
NodeName            HostName                HAVersion       StartTime           HAActive
HAAllResOK          HARunPhase
192-168-0-30        mgtomsdat-sh-3-01-1     V100R001C01     2014-11-18 23:43:02
active              normal                  Actived
192-168-0-24        mgtomsdat-sh-3-01-2     V100R001C01     2014-11-21 07:14:02
standby             normal                  Deactived
```

   说明

   如果当前登录的Master1节点是备管理节点，且需要登录到主管理节点时，请执行以下命令：

   ssh *Master2节点IP地址*

----结束

5.2 集群概览

5.2.1 集群列表简介

通过概览信息可以快速查看所有集群和作业的状态，您也可以通过MRS控制台左侧导航的“概览”中获取MRS相关帮助文档。

MRS用于海量数据的管理和分析，MRS使用简单，用户创建好一个集群，在集群内可新增MapReduce、Spark和Hive等作业，对用户数据进行分析和处理。数据处理完成后，采用SSL加密传输数据至OBS，保证数据的完整性和机密性。

集群状态

登录MRS管理控制台后，MRS所有集群包含的状态如表5-2所示。

表 5-2 集群状态说明

状态	说明
启动中	集群正在创建，则其状态为“启动中”。
运行中	集群创建成功且集群中所有组件状态均正常，则其状态为“运行中”。
扩容中	集群Core节点或者Task节点正在扩容，则其状态为“扩容中”。 **说明** 如果集群扩容失败，用户可重新进行扩容操作。
缩容中	当对集群节点进行关机、删除、变更OS、重装OS和修改规格的操作时，被变更的集群节点正在删除，则其状态为“缩容中”。
异常	集群中部分组件状态异常，导致集群异常，则其状态为“异常”。
删除中	集群节点正在删除中，则其状态为“删除中”。
已删除	集群已经删除，仅“历史集群”会显示此参数。
Master升级规格中	Master节点规格正在升级，则其状态为“Master升级规格中”。

作业状态

登录MRS的管理控制台后，用户在MRS集群中执行的作业包含的状态如表5-3所示。

表 5-3 作业状态说明

状态	说明
已接受	作业提交成功后的初始状态。
运行中	作业执行过程中，则其状态为“运行中”。
已完成	作业执行完成，并且执行成功，则其状态为“已完成”。
已终止	作业执行过程中，停止执行，则其状态为“已终止”。

状态	说明
异常	作业执行过程中报错，或者作业执行完成，但执行失败，则其状态为“异常”。

5.2.2 查看集群状态

集群列表显示MRS所有的集群，集群数量较多时，可采用翻页显示，您可以查看任何状态下的集群。

MRS作为一个海量数据管理和分析平台，数据处理能力在PB级以上。MRS支持创建多个集群，集群创建数量受弹性云服务器数量限制。

集群列表默认按时间顺序排列，时间最近的集群显示在最前端。集群列表参数说明如表5-4所示。

- 现有集群：包括除了“失败”和“已删除”状态以外的所有集群。
- 历史集群：仅包含“已删除”的集群，目前界面只显示6个月内创建且已删除的集群，若需要查看6个月以前删除的集群，请联系支持人员。
- 失败任务管理：仅包含“失败”状态的任务。
  - 集群创建失败的任务
  - 集群删除失败的任务
  - 集群扩容失败的任务
  - 集群缩容失败的任务
  - 集群安装补丁失败的任务（仅MRS 3.x之前版本支持）
  - 集群卸载补丁失败的任务（仅MRS 3.x之前版本支持）
  - 集群升级规格失败的任务

表 5-4 集群列表参数

参数	参数说明
名称/ID	集群的名称，创建集群时设置。集群的ID是集群的唯一标识，创建集群时系统自动赋值，不需要用户设置。 • ：修改集群名称。 • ：复制集群ID。
集群版本	集群的版本号。
节点数	集群部署的节点个数，创建集群时设置。

参数	参数说明
状态	集群状态、进度信息。 创建集群进度包括： • Verifying cluster parameters：校验集群参数中 • Applying for cluster resources：申请集群资源中 • Creating VMs：创建虚拟机中 • Initializing VMs：初始化虚拟机中 • Installing MRS Manager：安装MRS Manager中 • Deploying the cluster：部署集群中 • Cluster installation failed：集群安装失败 扩容集群进度包括： • Preparing for cluster expansion：准备扩容中 • Creating VM：创建虚拟机中 • Initializing VM：初始化虚拟机中 • Adding node to the cluster：节点加入集群中 • Cluster expansion failed：集群扩容失败 缩容集群进度包括： • Preparing for cluster shrink：正在准备缩容 • Decommissioning instance：实例退服中 • Deleting VM：删除虚拟机中 • Deleting node from the cluster：从集群删除节点中 • Cluster shrink failed：集群缩容失败 集群安装、扩容、缩容失败，会显示失败的原因，详情请参见**表4-5**。
创建时间	集群节点创建成功。
删除时间	集群节点停止时间，也是集群节点开始删除时间。仅“历史集群”会显示此参数。
可用区	集群工作区域下的可用区，创建集群时设置。
操作	删除：如果作业执行结束后不需要集群，可以单击“删除”，集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。当MRS集群部署失败时，集群会被自动删除。 仅“现有集群”会显示此参数。 **说明** 一般在数据完成分析和存储后或集群异常无法提供服务时才执行删除操作。如果数据没有完成处理分析，删除集群会导致数据丢失，请谨慎操作。

表 5-5 按钮说明

按钮	说明
	在下拉框中选择集群状态，筛选现有集群。 ● 现有集群 – 所有状态：表示筛选所有的现有集群 – 启动中：表示筛选“启动中”状态的现有集群 – 运行中：表示筛选“运行中”状态的现有集群 – 扩容中：表示筛选“扩容中”状态的现有集群。 – 缩容中：表示筛选“缩容中”状态的现有集群。 – 异常：表示筛选“异常”状态的现有集群 – 删除中：表示筛选“删除中”状态的现有集群
	选择“集群列表 > 现有集群”，单击 进入“失败任务管理”页面。 *Num*：表示“失败”状态的任务数。
	在搜索框中输入集群名称或ID，单击 ，搜索集群。
标签搜索	单击“标签搜索”输入待查询集群的标签，然后单击“搜索”搜索对应集群。 标签键或标签值可以通过下拉列表中选择，当标签键或标签值全匹配时，系统可以自动查询到目标集群。当有多个标签条件时，会取各个标签的交集，进行集群查询。
	单击 ，手动刷新现有集群列表。

5.2.3 查看集群基本信息

集群创建完成后，可对集群进行监控和管理。选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群详情页面，查看集群的基本配置信息、部署的节点信息。

说明

ECS集群和BMS集群在管理控制台操作基本一致，本文档主要以ECS集群描述为例，如有操作区别则分开描述。

在集群详情页面选择“概览”，参考表5-6查看集群详情概览信息参数说明。

表 5-6 集群基本信息

参数	参数说明
集群名称	集群的名称，创建集群时设置。单击 可对集群名称进行修改。 当MRS集群为MRS 3.x之前版本时，修改集群名称后仅MRS管理控制台界面显示的集群名称修改，MRS Manager中集群名称不会同步修改。
集群状态	集群状态信息，请参见**表5-2**。
集群管理页面	Manager页面入口。 • 针对MRS 3.x及以后版本，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）** • 针对MRS 3.x之前版本，需要根据提示绑定弹性公网IP及添加安全组规则后才能进入MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**。
集群版本	MRS版本信息。
集群类型	支持以下集群类型： • 分析集群：用来做离线数据分析，提供的是Hadoop体系的组件。 • 流式集群：用来做流处理任务，提供的是流式处理组件。 • 混合集群：既可以用来做离线数据分析，也可以用来做流处理任务，提供的是Hadoop体系的组件和流式处理组件。 • 自定义：全量自定义组件组合的MRS集群，MRS 3.x及之后版本支持此类型。
集群ID	集群的唯一标识，创建集群时系统自动赋值，不需要用户设置。
创建时间	显示集群创建的时间。
可用区	集群工作区域下的可用区，创建集群时设置。
默认生效子网	子网信息，创建集群时所选。 当子网IP不足时，单击“切换子网”切换到当前集群相同VPC下的其他子网，实现可用子网IP的扩充。切换子网不会影响当前已有节点的IP地址和子网。 通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。
虚拟私有云	VPC信息，创建集群时所选。 VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。
弹性公网IP	通过将弹性公网IP与MRS集群绑定，实现使用弹性公网IP访问Manager的目的。
OBS权限控制	单击“单击管理”，修改MRS用户与OBS权限的映射关系，具体请参考**配置MRS多用户访问OBS细粒度权限**。

参数	参数说明
数据连接	单击“单击管理”，查看集群关联的数据连接类型，具体请参考**配置数据连接**。
委托	单击“管理委托”，为集群绑定或修改委托。 通过绑定委托，您可以将部分资源共享给ECS或BMS云服务来管理，例如通过配置ECS委托可自动获取AK/SK访问OBS，具体请参见**配置存算分离集群（委托方式）**。 **MRS_ECS_DEFAULT_AGENCY**委托拥有对象存储服务的OBS OperateAccess权限和在集群所在区域拥有CES FullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。
密钥对	密钥对名称，创建集群时设置。 如果创建集群时设置的登录方式为密码，则不显示。
Kerberos认证	登录Manager管理页面时是否启用Kerberos认证。
日志记录 Logging	用于收集集群创建失败及扩缩容失败的日志。
安全组	集群的安全组名称。
数据盘密钥名称	用于加密数据盘的密钥名称。如需对已使用的密钥进行管理，请登录密钥管理控制台进行操作。
数据盘密钥ID	用于加密数据盘的密钥ID。
IAM用户同步	可以将IAM侧用户信息同步至MRS集群，用于集群管理。具体请参见**IAM用户同步MRS说明**。 **说明** 集群详情页的“组件管理”、“租户管理”和“备份恢复”页签需要同步用户后方可使用。MRS 3.x版本集群同步后可使用“组件管理”。
通讯安全授权	展示安全授权状态，通过 可关闭和开启安全授权。关闭安全授权属于高危操作，请谨慎处理。详细信息请参考**授权安全通信**。

表 5-7 组件版本

参数	参数说明
Hadoop 版本	显示Hadoop组件的版本信息。
Spark 版本	显示Spark组件的版本信息，MRS 3.x之前版本集群支持。
HBase 版本	显示HBase组件的版本信息。
Hive 版本	显示Hive组件的版本信息。
Hue 版本	显示Hue组件的版本信息。
Loader 版本	显示Loader组件的版本信息。

参数	参数说明
Kafka 版本	显示Kafka组件的版本信息。
Storm 版本	显示Storm组件的版本信息。
Flume 版本	显示Flume组件的版本信息。
Tez 版本	显示Tez组件的版本信息。
Presto 版本	显示Presto组件的版本信息。
KafkaManager 版本	显示KafkaManager组件的版本信息。
Opentsdb 版本	显示Opentsdb组件的版本信息。
Flink 版本	显示Flink组件的版本信息。
Alluxio版本	显示Alluxio组件的版本信息。
Ranger 版本	显示Ranger组件的版本信息。
Impala版本	显示Impala组件的版本信息。
Kudu版本	显示Kudu组件的版本信息。
Spark2x 版本	显示Spark2x组件的版本信息。仅MRS 3.x及之后版本集群支持。
Oozie 版本	显示Oozie组件的版本信息。仅MRS 3.x及之后版本集群支持。
ClickHouse版本	显示ClickHouse组件的版本信息。仅MRS 3.x及之后版本集群支持。

在集群详情页面选择“节点管理”，参考**表5-8**查看集群节点信息参数说明。

表 5-8 节点信息

参数	参数说明
配置Task节点	用于增加Task节点，请参见**添加Task节点**的相关任务。 对于3.x及之后版本，该操作仅适用于分析集群、流式集群和混合集群。
新增节点组	仅适用于3.x及之后版本，用于增加节点组，请参见**添加节点组**，仅适用自定义集群。
节点组名称	集群节点组名称。

参数	参数说明
节点类型	节点类型： • Master：集群主节点，负责管理集群，协调将MapReduce可执行文件分配到核心节点。此外，还会跟踪每个作业的执行状态，监控DataNode的运行状况。 • Task类型节点组是指仅部署了不存储数据的数据角色的节点组，主要包含：NodeManager、ThriftServer、Thrift1Server、RESTServer、Supervisor、Logviewer、HBaseIndexer、TagSync。 • 如果节点组内除以上角色外还部署了其他角色，则该节点组为Core类型节点组。 单击节点组名称前方的 ，显示该节点组包含的节点。节点参数说明请参见**管理组件和主机监控**。
节点数	对应节点组中包含的节点数量。
操作	• 扩容：请参见**扩容集群**。 • 缩容：请参见**缩容集群**。 • 弹性伸缩：请参见**配置弹性伸缩规则**。 • 查看角色信息：可查看所在节点组部署的角色信息。仅适用于3.x及之后版本的自定义集群。

5.2.4 查看集群补丁信息

查看集群组件的补丁信息。如果集群组件，如Hadoop或Spark等出现了异常，可下载补丁版本，选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群详情页面升级组件，修复问题。

说明

MRS 3.x版本无补丁版本信息，不涉及此章节。

- 补丁名称：补丁包的名称。
- 发布时间：补丁包发布的时间。
- 状态：展示补丁的状态。
- 补丁内容：补丁版本的描述信息。
- 操作：可安装或者卸载补丁。

5.2.5 查看和定制集群监控指标

MRS支持将集群中所有部署角色的节点，按管理节点、控制节点和数据节点进行分类，分别计算关键主机监控指标在每类节点上的变化趋势，并在报表中按用户自定义的周期显示分布曲线图。如果一个主机属于多类节点，那么对应的指标将被统计多次。

该任务指导用户了解MRS集群的概览、及在MRS查看、自定义与导出节点监控指标报表。

说明

集群指标监控采用周期性监控，历史监控平均周期约为5分钟。

方式一：（适用于MRS 3.x之前版本集群）

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 选择“概览”页签，即可在页面下方查看到集群主机健康状态统计。

步骤3 如需查看或导出其它指标的报表，请选择页面左侧基本信息栏的“集群管理页面 > 前往 Manager”登录Manager页面，具体请参见**访问集群Manager**。

步骤4 在Manager页面查看、定制与导出节点监控指标报表，具体请参见**系统概览**。

----结束

方式二：（适用于MRS 3.x之前版本集群）

步骤1 登录MRS控制台。

步骤2 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤3 在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

步骤4 用户同步完成后即可在页面右侧查看到集群的监控指标报表。

步骤5 在时间区间选择需要查看监控数据的时间段。可供选择的选项如下：

- 近1小时
- 近3小时
- 近12小时
- 近24小时
- 近7天
- 近一个月
- 自定义：在时间范围内自行选择需要查看的时间。

步骤6 自定义监控指标报表。

1. 单击“定制”，勾选需要显示的监控指标。

   MRS支持统计的指标共14个，界面最多显示12个定制的监控指标。

   - 集群主机健康状态统计
   - 集群网络读速率统计
   - 主机网络读速率分布
   - 主机网络写速率分布
   - 集群磁盘写速率统计
   - 集群磁盘占用率统计
   - 集群磁盘信息
   - 主机磁盘占用率统计
   - 集群磁盘读速率统计
   - 集群内存占用率统计

- 主机内存占用率分布
- 集群网络写速率统计
- 主机CPU占用率分布
- 集群CPU占用率统计

2. 单击“确定”保存并显示所选指标。

说明

单击“清除”可批量取消全部选中的指标项。

步骤7 导出监控指标报表。

1. 选择报表的时间范围。可供选择的选项如下：
   - 近1小时
   - 近3小时
   - 近12小时
   - 近24小时
   - 近7天
   - 近一个月
   - 自定义：在时间范围内自行选择需要查看的时间。
2. 单击“导出”，MRS将生成指定时间范围内、已勾选的集群监控指标报表文件，请选择一个位置保存，并妥善保管该文件。

----结束

方式三：（适用于MRS 3.x及之后版本集群）

步骤1 登录MRS控制台。

步骤2 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤3 在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

步骤4 用户同步完成后即可在页面右侧查看到集群的监控指标报表。

步骤5 在时间区间选择需要查看监控数据的时间段。可供选择的选项如下：

- 近1小时
- 近3小时
- 近12小时
- 近24小时
- 近7天
- 近1个月
- 自定义：在时间范围内自行选择需要查看的时间。

步骤6 自定义监控指标报表。

1. 单击“定制”，勾选需要显示的监控指标。
   界面最多显示12个定制的监控指标。
2. 单击“确定”保存并显示所选指标。

说明

单击“清除”可批量取消全部选中的指标项。

----结束

5.2.6 管理组件和主机监控

用户在日常使用中，可以在MRS管理所有组件（含角色实例）和主机的状态及指标信息：

- 状态信息，包括运行、健康、配置及角色实例状态统计。
- 指标信息，各组件的主要监控指标项。
- 导出监控指标（MRS 3.x及之后版本暂不支持）。

说明

- MRS 3.x之前版本操作方法请参考**管理服务和主机监控**。
- MRS 3.x及之后版本操作方法请参考**操作方法**。
- 用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。
- 组件管理支持三种参数值：
  - “每30秒刷新一次”：刷新间隔30秒。
  - “每60秒刷新一次”：刷新间隔60秒。
  - “停止”：停止刷新。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作方法

管理组件监控

说明

MRS Manager操作，请参考**管理服务监控**操作。

步骤1 在MRS集群详情页面，单击“组件管理”。

组件列表中标题包含“服务”、“操作状态”、“健康状态”、“配置状态”、“角色数”和“操作”。

- 服务操作状态描述如**表5-9**所示。

表 5-9 服务操作状态

状态	描述
已启动	服务已启动。
已停止	服务已停止。

状态	描述
启动失败	用户启动操作失败。
停止失败	用户停止操作失败。
未知	后台系统重启后，服务的初始状态。

- 服务健康状态如表5-10所示。

  表 5-10 服务健康状态

状态	描述
良好	该服务中所有角色实例正常运行。
故障	至少一个角色实例运行状态为“故障”或被依赖的服务状态不正常。
未知	该服务中所有角色实例状态为“未知”。
正在恢复	后台系统正在尝试自动启动服务。
亚健康	该服务所依赖的服务状态不正常，异常服务的相关接口无法被外部调用。

- 服务配置状态如表5-11所示。

  表 5-11 服务配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
配置超期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
配置失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
配置中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“服务”列按升序排列，单击**服务**、**操作状态**、**健康状态**或**配置状态**可修改排列方式。

步骤2 单击列表中指定服务名称，查看服务状态及指标信息。

步骤3 定制、查看监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。

----结束

管理角色实例监控

说明

针对MRS 3.x之前版本，请参考**管理角色实例监控**操作。

步骤1 在MRS集群详情页面，单击“组件管理”，在组件列表中单击服务指定名称。

步骤2 单击“实例”，查看角色状态。

角色实例列表中包含实例信息的**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**运行状态**及**配置状态**。

- 角色实例的运行状态如**表5-12**所示。

表 5-12 角色实例运行状态

状态	描述
良好	表示实例当前运行正常。
故障	表示实例当前无法正常工作。
已退服	表示实例处于退服状态。
未启动	表示实例已停止。
未知	表示实例的初始状态信息无法检测。
正在启动	表示实例正在执行启动过程。
正在停止	表示实例正在执行停止过程。
正在恢复	表示实例可能存在异常正在自动修复。
正在退服	表示实例正在执行退服过程。
正在入服	表示实例正在执行入服过程。
启动失败	表示实例启动操作失败。
停止失败	表示实例停止操作失败。

- 角色实例的配置状态如**表5-13**所示。

表 5-13 角色实例配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
配置超期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
配置失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
配置中	参数配置进行中。

状态	描述
未知	无法获取当前配置状态。

默认以“角色”列按升序排列，单击**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**运行状态**或**配置状态**可修改排列方式。

支持在“角色”筛选相同角色的全部实例。

单击“高级搜索”，在角色搜索区域中设置搜索条件，单击“搜索”，查看指定的角色信息。单击“重置”清除输入的搜索条件。支持模糊搜索条件的部分字符。

步骤3 单击列表中指定角色实例名称，查看角色实例状态及指标信息。

步骤4 定制、查看监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。

----结束

管理主机监控

说明

针对MRS 3.x之前版本，请参考**管理主机监控**操作。

步骤1 在MRS集群详情页面，单击“节点管理”并展开节点组信息，查看所有主机状态。

主机列表中包括**节点名称**、**IP**、**机架**、**操作状态**、**健康状态**、**CPU使用率**、**内存使用率**、**磁盘使用率**、**网络速度**、**规格名**、**规格**、**可用区**。

- 主机操作状态如**表5-14**所示。

表 5-14 主机操作状态

状态	描述
正常	主机及主机上的服务角色正常运行。
已隔离	主机被用户隔离，主机上的服务角色停止运行。

- 主机健康状态描述如**表5-15**所示。

表 5-15 主机健康状态

状态	描述
良好	主机心跳检测正常。
故障	主机心跳超时未上报。
未知	执行添加操作时，主机的初始状态。

默认以“节点名称”列按升序排列，单击**节点名称**、**IP**、**机架**、**操作状态**、**健康状态**、**CPU使用率**、**内存使用率**、**磁盘使用率**、**网络速度**、**规格名**或**规格**可修改排列方式。

步骤2 单击列表中指定的节点名称，查看单个节点状态及指标。

----结束

5.3 集群运维

5.3.1 导入导出数据

用户通过“文件管理”页面可以在分析集群进行文件夹创建、删除，文件导入、导出、删除操作，暂不支持文件创建功能。流式集群暂不支持在界面使用“文件管理”功能。开启Kerberos认证的集群中，根目录下的文件夹有权限限制，如需对其进行读写，请参考**创建角色**内容添加拥有对应文件夹权限的角色，再请参考**相关任务**修改提交作业用户所属的用户组，将新增的组件角色加入到该用户组中。

背景信息

MRS集群处理的数据源来源于OBS或HDFS，HDFS是Hadoop分布式文件系统（Hadoop Distributed File System），OBS即对象存储服务，是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用，同时可以通过REST API接口方式单独或集成到业务程序进行管理和访问数据。

用户创建作业前需要将本地数据上传至OBS系统，MRS使用OBS中的数据进行计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

导入数据

MRS目前只支持将OBS上的数据导入至HDFS中。上传文件速率会随着文件大小的增大而变慢，适合数据量小的场景下使用。

支持导入文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。

   “bd_app1”目录仅为示例，可以是界面上的任何目录，也可以通过“新建”创建新的文件夹。

   新建文件夹时需要满足以下要求：

   - 文件夹名称小于等于255字符，全路径长度小于等于1023字符。
   - 不允许为空。

- 不能包含：/:*?"<>|\;&,'`!{}[]$%+特殊字符。
- 不能以“.”开头或结尾。
- 开头和末尾的空格会被忽略。

6. 单击“导入数据”，正确配置HDFS和OBS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。
   - OBS路径
     - 必须以“obs://”开头。
     - 不支持导入KMS加密的文件或程序。
     - 不支持导入空的文件夹。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
     - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
     - OBS全路径长度小于等于1023字符。
   - HDFS路径
     - 默认以“/user”开头。
     - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
     - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
     - HDFS全路径长度小于等于1023字符。
7. 单击“确定”。

   文件上传进度可在“文件操作记录”中查看。MRS将数据导入操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

导出数据

数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。

支持导出文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群基本信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd_app1”。
6. 单击“导出数据”，配置OBS和HDFS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。
   - OBS路径
     - 必须以“obs://”开头。

- 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\特殊字符。
- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- OBS全路径长度小于等于1023字符。

– HDFS路径

- 默认以“/user”开头。
- 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'$*?\:特殊字符。
- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- HDFS全路径长度小于等于1023字符。

说明

当导出文件夹到OBS系统时，在OBS路径下，将增加一个标签文件，文件命名为“folder name_$folder$”。请确保导出的文件夹为非空文件夹，如果导出的文件夹为空文件夹，OBS无法显示该文件夹，仅生成一个命名为“folder name_$folder$”的文件。

7. 单击“确定”。

文件上传进度可在“文件操作记录”中查看。MRS将数据导出操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

查看文件操作记录

通过MRS管理控制台导入和导出数据时，可在“文件管理 > 文件操作记录”查看数据导入、导出进度。

文件操作记录参数说明如表5-16所示。

表 5-16 文操作记录参数说明

Parameter	Description
提交时间	数据导入或导出操作的开始时间。
源目录	数据的源路径。 • 数据导入时“源目录”为OBS路径 • 数据导出时“源目录”为HDFS路径
目标目录	数据的目标路径。 • 数据导入时“目标目录”为HDFS路径 • 数据导出时“目标目录”为OBS路径

Parameter	Description
状态	数据导入或导出操作的状态。 • 已提交 • 已接受 • 运行中 • 已完成 • 已终止 • 异常
持续时间（分钟）	数据导入或导出操作的总时间。 单位：分钟
执行结果	数据导入或导出操作的结果。 • 成功 • 失败 • 终止 • 未定
操作	查看日志：查看文件操作日志。

5.3.2 切换集群子网

MRS支持当子网IP不足时，切换子网到当前集群相同VPC下的其他子网，实现可用子网IP的扩充。切换子网不会影响当前已有节点的IP地址和子网。

如需对网络ACL出规则进行配置请参考**如何配置网络ACL出规则？**。

未关联网络 ACL 时切换子网

步骤1 登录MRS控制台。

步骤2 单击集群名称进入集群详情页。

步骤3 在“默认生效子网”右侧单击“切换子网”。

步骤4 选择待切换子网，并单击“确定”完成切换。

若没有可用子网，请单击“创建子网”进入VPC控制台创建子网后，再在此处引用。

----结束

关联网络 ACL 时切换子网

步骤1 登录MRS控制台，单击集群名称进入集群详情页。

步骤2 在MRS集群详情页，查看“虚拟私有云”。

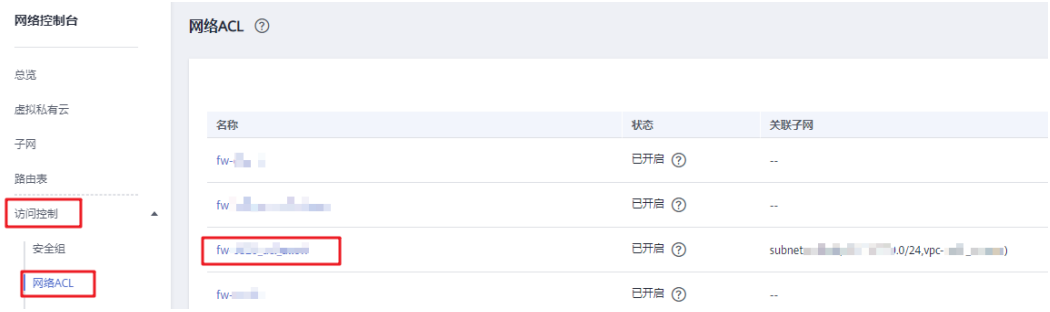
步骤3 进入VPC控制台，在左侧导航处选择“虚拟私有云”，获取**步骤2**中查询的“虚拟私有云”对应的“IPv4网段”。

步骤4 在VPC控制台左侧导航处选择“访问控制 > 网络ACL”，单击MRS集群默认生效子网和切换后子网关联的网络ACL名称，选择“入方向规则”页签。

说明

若集群的默认生效子网和切换后子网均关联网络ACL，则两个子网关联的网络ACL中均需参考**步骤5**-**步骤7**增加入方向规则。

图 5-3 网络 ACL



步骤5 在第一条规则的“操作”列，选择“更多 > 向前插规则”。

步骤6 添加网络ACL规则，其中“策略”选择“允许”，“源地址”填入**步骤3**中获取的“虚拟私有云”对应的“IPv4网段”，其他值保持默认即可。

步骤7 单击“确定”完成网络ACL规则添加。

说明

如果您不想放开VPC对应的全部IPv4网段的规则，请参考**步骤8**-**步骤12**添加切换前后两个子网对应的IPv4网段地址。若已完成VPC对应IPv4网段的规则添加，则无需执行**步骤8**-**步骤12**的操作。

步骤8 登录MRS控制台。

步骤9 单击集群名称进入集群详情页。

步骤10 在“默认生效子网”右侧单击“切换子网”。

步骤11 获取“默认生效子网”和待切换子网对应的IPv4网段。

须知

此时请勿单击切换子网的“确定”按钮，否则默认生效子网将更新为切换后的子网，切换前的子网不易查询，请谨慎操作。

步骤12 参考**步骤4**-**步骤7**添加“默认生效子网”和待切换子网的IPv4网段地址到切换前后子网绑定的网络ACL入方向规则中。

步骤13 登录MRS控制台。

步骤14 单击集群名称进入集群详情页。

步骤15 在“默认生效子网”右侧单击“切换子网”。

步骤16 选择待切换子网，并单击“确定”完成切换。

----结束

如何配置网络 ACL 出规则?

- 方案一：

  放通网络ACL所有出站流量，此方案能保证集群正常创建与使用，优先建议使用此方案。

- 方案二：

  放通保证集群创建成功的最小出规则，此方案可能在后续使用中因出方向规则遗漏导致集群使用问题，不建议使用方案。若出现集群使用问题请联系运维人员支撑处理。

  配置示例：参照方案一中示例，配置策略为“允许”，目的地址为通信安全授权地址、NTP、OBS、Openstack及DNS地址的出方向规则。

5.3.3 配置消息通知

MRS联合消息通知服务(SMN)，采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式（短信和邮件通知）。

操作场景

在MRS管理控制台，按照集群维度，在集群信息页面的告警页签中能够提供选择是否使能通知服务，只有对应集群开关开启以后，才能实现以下场景的功能：

- 在用户订阅了通知服务之后，当集群出现扩容成功/失败、缩容成功/失败、删除成功/失败、弹性升缩成功/失败的场景下，由MRS管理面通过邮件或短信方式通知对应用户。
- 管理面检查大数据集群的告警信息，如果大数据集群的告警信息影响到服务的使用，其告警级别达到致命时，则发送信息通知给对应租户。
- 在用户集群的ECS机器被删除、关机、修改规格、重启、更新OS的行为，会导致大数据集群异常，当检测到用户的虚拟机出现以上状态的时候，发送通知给对应用户。

创建主题

主题是消息发布或客户端订阅通知的特定事件类型。它作为发送消息和订阅通知的信道，为发布者和订阅者提供一个可以相互交流的通道。

1. 登录管理控制台。
2. 单击“服务列表”选择“管理与监管 > 消息通知服务”。

   进入消息通知服务页面。
3. 在左侧导航栏，选择“主题管理 > 主题”。

   进入主题页面。
4. 在主题页面，单击“创建主题”，开始创建主题。

   此时将显示“创建主题”对话框。
5. 在“主题名称”框中，输入主题名称，在“显示名”框中输入相关描述。
6. 在“企业项目”中选择已有的项目，或者单击“新建企业项目”，在“企业项目管理”界面创建好企业项目后再进行添加。
7. 在“标签”填写“标签键”和“标签值”，用于标识云资源，可对云资源进行分类和搜索。

向主题添加订阅

要接收发布至主题的消息，您必须添加一个订阅终端节点到该主题。消息通知服务会发送一条订阅确认的消息到订阅终端，订阅确认的消息将在48小时内有效。如果订阅者在48小时之内确认订阅，将会收到推送至主题的消息。如果订阅者在48小时之内没有确认订阅，则需要再次给订阅者发送订阅确认的消息。

1. 登录管理控制台。
2. 选择“管理与监管 > 消息通知服务”。

   进入消息通知服务页面。
3. 在左侧导航栏，选择“主题管理 > 主题”。

   进入主题页面。
4. 在主题列表中，选择您要向其添加订阅者的主题，在右侧“操作”栏单击“添加订阅”。

   此时将显示“添加订阅”对话框。

   其中：协议参数选项为“短信”、“邮件”、FunctionGraph（函数）HTTP、HTTPS。

   订阅终端参数为订阅的终端地址，短信、邮件终端支持批量输入，批量添加时，每个终端地址占一行。最多可输入10个终端。
5. 单击“确定”。

新增订阅将显示在页面下方的订阅列表中。

向订阅者发送消息

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。
3. 单击“告警管理”。
4. 选择“消息订阅规则 > 添加消息订阅规则”，进入添加消息订阅页面。
5. 配置消息订阅规则相关参数。

表 5-17 消息订阅规则参数说明

参数	说明
规则名称	用户自定义发送订阅消息的规则名称，只能包含数字、英文字符、中划线和下划线。
提醒通知	• 选择开启时，将按照该订阅规则为订阅者发送对应订阅消息。 • 选择关闭时，该规则不会生效，即不会向订阅者发送订阅消息。
主题名称	选择已创建的主题，也可以单击“创建主题”重新创建。

参数	说明
消息类型	选择需要订阅的消息类型。 ● 告警 ● 事件
订阅规则	选择需要订阅的消息规则，可根据需要勾选全部或部分规则。 MRS 3.x及之后版本订阅规则： 告警：紧急，重要，次要 事件：重要，次要，提示 MRS 3.x之前版本订阅规则： ● 致命 ● 严重 ● 一般 ● 提示

6. 单击“确定”完成消息提醒配置。

5.3.4 健康检查

5.3.4.1 使用前须知

本章节指导用户在MRS控制台执行健康检查管理操作。

在MRS控制台执行健康检查管理操作仅适用于**MRS 1.9.2~MRS 2.1.x版本**集群。

在Manager界面执行健康检查管理操作适用于所有版本。

MRS 3.x之前版本请参考**执行健康检查**。

5.3.4.2 执行健康检查

操作场景

该任务指导用户在日常运维中完成集群进行健康检查的工作，以保证集群各项参数、配置以及监控没有异常、能够长时间稳定运行。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

操作步骤

- 手动执行所有服务的健康检查

  在集群详情页，单击页面右上角“管理操作 > 启动集群健康检查”。

  说明

  - 集群健康检查包含了Manager、服务与主机状态的检查。
  - 在MRS Manager界面，选择“系统设置 > 健康检查 > 集群健康检查”，也可以执行集群健康检查。
  - 手动执行健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。

- 手动执行单个服务的健康检查

  a. 在集群详情页，单击“组件管理”。

  b. 在服务列表中单击指定服务名称。

  c. 选择“更多 > 启动服务健康检查”启动指定服务健康检查。

- 手动执行主机健康检查

  a. 在集群详情页，单击“节点管理”。

  b. 展开节点组信息，勾选待检查主机前的复选框。

  c. 选择“节点操作 > 启动主机健康检查”启动指定主机健康检查。

5.3.4.3 查看并导出检查报告

操作场景

为了满足对健康检查结果的进一步具体分析，您可以在MRS中查看以及导出健康检查的结果。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

前提条件

已执行健康检查。

操作步骤

步骤1 在集群详情页，单击页面右上角“管理操作 > 查看集群健康检查报告”。

步骤2 在健康检查的报告面板上单击“导出报告”导出健康检查报告，可查看检查项的完整信息。

----结束

5.3.5 远程运维

5.3.5.1 运维授权

当用户使用集群过程中出现问题需要支持人员协助解决时，用户可先联系支持人员，再通过“运维授权”功能授权支持人员访问用户机器的权限用于定位问题。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在页面右上角单击“运维”，选择“运维授权”，选择授权给支持人员访问本机的权限的“截止时间”。在截止时间之前支持人员有临时访问本机的权限。

步骤4 问题解决后，在页面右上角单击“运维”，选择“取消授权”为支持人员取消访问权限。

----结束

5.3.5.2 日志共享

当用户使用集群过程中出现问题需要支持人员协助解决时，用户可先联系支持人员，再通过“日志共享”功能提供特定时间段内的日志给支持人员以便定位问题。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在页面右上角单击“运维”，选择“日志共享”，进入“日志共享”界面。

步骤4 在“起止时间”的输入框单击选择日期和时间。

说明

- “起止时间”按照支持人员的建议选取。
- 结束时间的选择的时间必须大于开始时间选择的时间，否则，无法根据时间筛选日志。

----结束

5.3.6 查看 MRS 服务操作日志

“操作日志”页面记录用户对集群和作业的操作信息。日志信息常用于集群运行异常时的问题定位，帮助用户快速定位问题原因，以便及时解决问题。

操作类型

目前MRS记录以下操作类型的日志信息，可在搜索框中筛选查询：

- 集群操作
  - 创建集群、删除集群、扩容集群和缩容集群等
  - 创建目录、删除目录和删除文件
- 作业操作：创建作业、停止作业和删除作业
- 数据操作：IAM用户任务、新增用户、新增用户组等操作

日志字段

日志列表默认按时间顺序排列，时间最近的日志显示在最前端。

日志信息中的各字段说明如表5-18所示。

表 5-18 日志说明

参数	参数说明
操作类型	记录执行的操作类型，包括： ● 集群操作 ● 作业操作 ● 数据操作
操作IP	记录执行操作的IP地址。 **说明** 当MRS集群部署失败时，集群会被自动删除，并且自动删除集群的操作日志中不包含用户的“操作IP”信息。
操作内容	记录实际操作内容，不超过2048字符。
时间	记录操作的时间。对于已删除的集群，界面只显示6个月内的日志信息，若需要查看6个月之前的日志信息，请联系支持人员。

表 5-19 按钮说明

按钮	说明
	在下拉框中选择操作类型，筛选日志。 ● 全部：表示筛选所有的日志 ● 集群操作：表示筛选“集群操作”的日志 ● 作业操作：表示筛选“作业操作”的日志 ● 数据操作：表示筛选“数据操作”的日志
	根据时间筛选日志。 1. 单击输入框。 2. 选择日期和时间。 3. 单击“确认”。 左侧框为需要查询的开始时间，右侧框为需要查询的结束时间。右侧的输入框选择的时间必须大于或等于左侧输入框的时间，否则，无法根据时间筛选日志。

按钮	说明
	在搜索框中输入“操作内容”中的关键字，单击，搜索日志。
	单击，手动刷新日志列表。

5.3.7 删除集群

如果作业执行结束后不需要集群， 可以删除MRS集群。

背景信息

一般在数据完成分析和存储后或集群异常无法提供服务时才执行集群删除操作。当MRS集群部署失败时，集群会被自动删除。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 在左侧导航栏中选择“集群列表 > 现有集群”。

步骤3 在需要删除的集群对应的“操作”列中，单击“删除”。

集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。

----结束

5.4 节点管理

5.4.1 扩容集群

MRS的扩容不论在存储还是计算能力上，都可以简单地通过增加Core节点或者Task节点来完成，不需要修改系统架构，降低运维成本。集群Core节点不仅可以处理数据，也可以存储数据。可以在集群中添加Core节点，通过增加节点数量处理峰值负载。集群Task节点主要用于处理数据，不存放持久数据。

背景信息

MRS集群支持Core与Task节点总数最大为500个。如果用户需要的Core/Task节点数大于500，可以联系支持人员或者调用后台接口修改数据库。

目前支持扩容Core节点和Task节点，不支持扩容Master节点。此处扩容的最大Core/Task节点数为（500 - 集群Core/Task节点数）。例如：当前集群Core节点数为3，此处扩容的Core节点数必须小于等于497。如果集群扩容失败，用户可重新进行扩容操作。

如果在创建集群时，没有扩容节点，用户可以在扩容时添加节点个数，但不能指定具体节点扩容。

选择的版本不同，扩容集群的操作也不同。

约束条件

- 对于已安装HBase组件的节点组扩容时：

  如果集群没有开启节点自动注册DNS，扩容该节点组时需选择不启动组件，之后请参考**更新客户端**更新完HBase客户端配置后再启动扩容节点上的HBase组件实例。

- 扩容后，集群内节点已安装的客户端无需更新，集群外节点安装的客户端请参考**更新客户端**进行更新。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择 “集群列表 > 现有集群” ，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 选择“节点管理”页签，在需要扩容的节点组的“操作”列单击“扩容”，进入扩容集群页面。

只有运行中的集群才能进行扩容操作。

步骤4 设置“扩容节点数量”、“启动组件”和“执行引导操作”参数，并单击“确定”。

说明

- 若集群中没有Task节点组，请参考**添加Task节点**配置Task节点。
- 如果创建集群时添加了引导操作，则“执行引导操作”参数有效，开启该功能时扩容的节点会把创建集群时添加的引导操作脚本都执行一遍。
- 如果“新节点规格”参数有效，则表示与原有节点相同的规格已售罄或已下架，新扩容的节点将按照“新节点规格”增加。
- 扩容集群前需要检查集群安全组是否配置正确，要确保集群入方向安全组规则中有一条全部协议，全部端口，源地址为可信任的IP访问范围的规则。

步骤5 进入“扩容节点”窗口，单击“确定”。

步骤6 弹出扩容节点提交成功提示框。

集群扩容过程说明如下：

- 扩容中：集群正在扩容时集群状态为“扩容中”。已提交的作业会继续执行，也可以提交新的作业，但不允许继续扩容和删除集群，也不建议重启集群和修改集群配置。
- 扩容成功：集群扩容成功后集群状态为“运行中”。
- 扩容失败：集群扩容失败时集群状态为“运行中”。用户可以执行作业，也可以重新进行扩容操作。

扩容成功后，可以在集群详情的“节点管理”页签查看集群的节点信息。

----结束

添加 Task 节点

“自定义”类型集群添加Task节点操作步骤：

1. 在集群详情页面，选择“节点管理”页签，单击“新增节点组”，进入“新增节点组”页面。

2. “部署角色”参数仅选择“NM”部署NodeManager角色，则新增节点组为Task节点组，其他参数根据需要配置。

非“自定义”类型集群添加Task节点操作步骤：

1. 在集群详情页面，选择“节点管理”页签，单击“配置Task节点”，进入“配置Task节点”页面。
2. 配置“节点类型”、“节点规格”、“节点数量”、“系统盘”，如开启“添加数据盘”后，还需要配置数据盘的存储类型、大小和数量。
3. 单击“确定”。

添加节点组

说明

用于增加节点组，适用MRS 3.x版本的自定义集群。

1. 在集群详情页面，选择“节点管理”页签，单击“新增节点组”，进入“新增节点组”页面。
2. 根据需求配置参数。

表 5-20 新增节点组参数说明

参数名称	描述
节点规格	选择节点组内主机的规格类型。
节点数量	设置新增节点组内的节点数量。
系统盘	设置新增节点的系统盘的规格与容量。
数据盘/数据盘数量	设置新增节点的数据盘的规格与容量及数量。
部署角色	新增节点组内，各节点的实例部署发布，可手动调节。

3. 单击“确定”。

5.4.2 缩容集群

用户可以根据业务需求量，通过简单的缩减Core节点或者Task节点，对集群进行缩容，以使MRS拥有更优的存储、计算能力，降低运维成本。

当集群正在进行主备同步操作时，不允许进行缩容操作。

背景信息

目前支持缩容Core节点和Task节点，不支持缩容Master节点。对集群进行缩容时，只需要在界面调整节点个数，MRS会自动选择缩容节点，完成缩容任务。

自动选择缩容节点的策略如下：

- 不允许缩容安装了基础组件（Zookeeper，DBService，KrbServer，LdapServer等）的节点，MRS不会选择这些节点进行缩容。因为这些基础组件是集群运行的基础。

- Core节点是存放集群业务数据的节点，在缩容时必须保证待缩容节点上的数据被完整迁移到其他节点，即完成各个组件的退服之后，才会执行缩容的后续操作（节点退出Manager和删除ECS等）。在选择Core节点时，会优先选择存储数据量较小，且可退服实例健康状态良好的节点，避免节点退服失败。例如在分析集群上，Core节点安装了DataNode，缩容时会优先选择DataNode存储数据量较小且健康状态良好的节点。

  Core节点在缩容的时候，会对原节点上的数据进行迁移。业务上如果对数据位置做了缓存，客户端自动刷新位置信息可能会影响时延。缩容节点可能会影响部分HBase on HDFS数据的第一次访问响应时长，可以重启HBase或者对相关的表Disable/Enable来避免。
- Task节点本身不存储集群数据，属于计算节点，不存在节点数据迁移的问题。因此在选择Task节点时，优先选择健康状态为故障、未知、亚健康的节点进行缩容。这些节点实例的健康状态信息可以在MRS上的“实例”管理界面查看。

缩容校验策略

缩容节点选择完成后，为了避免组件退服失败，不同组件提供了不同的退服约束规则，只有满足了所有安装组件的退服约束规则才允许缩容。缩容校验策略如**表5-21**所示。

表 5-21 组件退服约束规则

组件名称	退服约束规则
HDFS/ DataNode	规则：缩容后节点数不小于当前HDFS的副本数且HDFS数据总量不超过缩容后HDFS集群总容量的80%，可以执行缩容操作。 原因：确保缩容后剩余空间足够存放现有数据，并预留一部分空间。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。
HBase/ RegionServer	规则：除缩容节点外，其他节点RegionServer剩余可用内存的总和，大于所选缩容节点RegionServer当前使用内存的1.2倍。 原因：当一个节点退服时，这个节点上的Region会迁移到其他节点，所以其他节点的可用内存必须足够才能负担起退服节点的Region。
Storm/ Supervisor	规则：缩容后集群slot数足够运行当前已提交的任务。 原因：防止缩容后没有充足的资源运行流处理任务。
Flume/ FlumeServer	规则：节点安装了FlumeServer，并且已经配置了Flume任务，则该节点不能删除。 原因：防止误删了已部署的业务程序。

指定数量缩容

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 选择“节点管理”页签，在需要缩容的节点组的“操作”列，单击“缩容”，进入缩容集群页面。

只有运行中的集群且集群中的节点都在运行中才能进行该操作。

步骤4 设置“缩容节点数量”，并单击“确定”。

说明

- 缩容集群前需要检查集群安全组是否配置正确，要确保集群入方向安全组规则中有一条全部协议，全部端口，源地址为可信任的IP访问范围的规则。
- 若HDFS存在损坏的数据块，则缩容集群可能出现失败，请联系支持人员处理。

步骤5 页面右上角弹出缩容节点提交成功提示框。

集群缩容过程说明如下：

- 缩容中：集群正在缩容时集群状态为“缩容中”。已提交的作业会继续执行，也可以提交新的作业，但不允许继续缩容和删除集群，也不建议重启集群和修改集群配置。
- 缩容成功：集群缩容成功后集群状态为“运行中”。
- 缩容失败：集群缩容失败时集群状态为“运行中”。用户可以执行作业，也可以重新进行缩容操作。

缩容成功后，可以在集群详情的“节点管理”页签查看集群的节点信息。

----结束

5.4.3 管理主机（节点）操作

操作场景

当主机（节点）故障异常时，用户可能需要在MRS停止主机（节点）上的所有角色，对主机（节点）进行维护检查。故障清除后，启动主机（节点）上的所有角色恢复主机（节点）业务。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“节点管理”。

步骤2 展开节点组信息，勾选待操作节点前的复选框。

步骤3 选择“节点操作 > 启动所有角色”或“停止所有角色”执行相应操作。

----结束

5.4.4 隔离主机

操作场景

用户发现某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。在为集群安装补丁的场景中，也支持排除指定节点不安装补丁。

该任务指导用户在MRS上根据实际业务或运维规划手工将主机隔离。隔离主机仅支持隔离非管理节点。

对系统的影响

- 主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。
- 主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“节点管理”。

步骤2 展开节点组信息，勾选待隔离主机前的复选框。

步骤3 选择“节点操作 > 隔离主机”。

步骤4 确认待隔离主机信息并单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功隔离，“操作状态”显示为“已隔离”

说明

已隔离的主机，可以取消隔离重新加入集群，请参见**取消隔离主机**。

----结束

5.4.5 取消隔离主机

操作场景

用户已排除主机的异常或故障后，需要将主机隔离状态取消才能正常使用。

该任务指导用户在MRS上取消隔离主机。

前提条件

- 主机状态为“已隔离”。
- 主机的异常或故障已确认修复。

- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“节点管理”。

步骤2 展开节点组信息，勾选待取消隔离主机前的复选框。

步骤3 选择“节点操作 > 取消隔离主机”。

步骤4 确认待取消隔离主机信息并单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功取消隔离，“操作状态”显示为“正常”。

步骤5 勾选已取消隔离的主机，选择“节点操作 > 启动所有角色”。

----结束

5.4.6 升级 Master 节点规格

随着用户业务的增长，Core节点的扩容，CPU使用率变高，而Master节点规格已经不满足用户需求时，则需要升级Master节点规格。本章节介绍Master节点规格升级的操作流程。

前提条件

确认是否开启了企业主机安全（Host Security Service，简称HSS）服务，如果已开启，升级Master节点规格前需要先暂时关闭HSS服务对MRS集群的监测。

使用限制

- 支持2个及以上Master节点的集群升级Master节点规格 。
- 不支持使用BMS类型规格的集群升级Master节点规格 。

集群 Master 节点规格升级（一键升级）

步骤1 登录MRS管理控制台。

步骤2 选择 “集群列表 > 现有集群” ，选中需要升级Master节点规格的集群并单击集群名，进入集群信息页面。

步骤3 在“节点管理”页签Master节点组的“操作”列选择“升级规格”，进入“升级Master规格”页面。

步骤4 选择升级后的规格，单击“提交”成功提交升级Master规格任务。

节点规格升级过程需要时间，升级成功后集群状态更新为“运行中”，请您耐心等待。

说明

- 升级过程中集群会自动关闭升级的虚拟机，升级完成后自动开启该虚拟机 。
- 因用户对组件使用需求不同，节点规格升级成功后不会自动更新组件内存配置，用户可根据实际使用情况自行调整各组件内存配置。

----结束

5.5 作业管理

5.5.1 MRS 作业简介

MRS作业是MRS为用户提供的程序执行平台，用于处理和分析用户数据。作业创建完成后，所有的作业列表信息展示在“作业管理”页面中，您可以查看所有的作业列表，也可以创建和管理作业。若集群详情页面不支持“作业管理”页签，请通过后台方式提交作业。

MRS集群处理的数据源来源于OBS或HDFS，HDFS是Hadoop分布式文件系统（Hadoop Distributed File System），OBS即对象存储服务，是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用，同时可以通过REST API接口方式单独或集成到业务程序进行管理和访问数据。

用户创建作业前需要将本地数据上传至OBS系统，MRS使用OBS中的数据进行计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

作业分类

目前MRS集群支持创建和管理如下几种类型的作业。如果处于“运行中”状态的集群创建作业失败，请查看集群管理页面中相关组件健康情况。操作方法，请参见**查看和定制集群监控指标**。

- MapReduce：提供快速并行处理大量数据的能力，是一种分布式数据处理模式和执行环境。MRS当前支持提交MapReduce Jar程序。
- Spark：基于内存进行计算的分布式计算框架，MRS当前支持提交SparkSubmit、Spark Script和Spark SQL作业。
  - SparkSubmit：支持提交Spark Jar和Spark python程序，执行Spark application，计算和处理用户数据。
  - SparkScript：支持提交SparkScript脚本，批量执行Spark SQL语句。
  - Spark SQL：运用Spark提供的类似SQL的Spark SQL语言，实时查询和分析用户数据。
- Hive：建立在Hadoop基础上的开源的数据仓库。MRS当前支持提交HiveScript脚本，和执行Hive SQL语句。
- Flink：提供一个分布式大数据处理引擎，可对有限数据流和无限数据流进行有状态计算。

作业列表

作业列表默认按时间顺序排列，时间最近的作业显示在最前端。各类作业列表参数说明如**表 1**所示。

表 5-22 作业列表参数

参数	参数说明
作业名称/ID	作业的名称，新增作业时配置。 ID是作业的唯一标识，作业新增后系统自动赋值。
用户名称	提交作业的用户名称。
作业类型	支持的作业类型： • Distcp：导入、导出数据 • MapReduce • Spark • SparkSubmit • SparkScript • Spark SQL • Hive SQL • HiveScript • Flink **说明** • 在“文件管理”页面进行文件的导入导出操作后，您可以在“作业管理”页面查看Distcp作业。 • 只有创建集群时选择了Spark、Hive和Flink组件，并且集群处于运行中，才能新增Spark、Hive和Flink类型的作业。
状态	显示作业的状态。 • 已提交 • 已接受 • 运行中 • 已完成 • 已终止 • 异常
执行结果	显示作业执行完成的结果。 • 未定：正在执行的作业。 • 成功：执行成功的作业。 • 终止：执行中被手动终止的作业。 • 失败：执行失败的作业。 **说明** 作业执行成功或失败后都不能再次执行，只能新增作业，配置作业参数后重新提交作业。

参数	参数说明
队列名称	提交作业用户绑定的队列的名称。
作业提交时间	记录作业提交的开始时间。
作业结束时间	记录作业执行完成或手工停止的时间。
操作	• 查看日志：单击“查看日志”，查看运行中的作业执行的实时日志信息。操作方法，请参见**查看作业配置信息和日志**。 • 查看详情：单击“查看详情”，查看作业的详细配置信息。操作方法，请参见**查看作业配置信息和日志**。 • 更多 – 停止：单击“停止”，停止正在运行的作业。操作方法，请参见**停止作业**。 – 删除：单击“删除”，删除一个作业。操作方法，请参见**删除作业**。 – 结果：单击“结果”，查看SparkSql和SparkScript类型的“状态”为“已完成”且“执行结果”为“成功”的作业执行结果。 **说明** • Spark SQL作业不支持停止。 • 作业删除后不可恢复，请谨慎操作。 • 当选择保留作业日志到OBS或HDFS时，系统在作业执行结束后，将日志压缩并存储到对应路径。因此，此类作业运行结束后，作业状态仍然为“运行中”，需等日志存储成功后，状态变更为“已完成”。日志存储花费时间依赖于日志大小，需要数分钟以上。

表 5-23 按钮说明

按钮	说明
	选择提交作业的时间区间，筛选在对应时间区间内提交的作业。
	在下拉框中选择作业执行结果，筛选作业。 • 全部：表示筛选所有的作业。 • 成功：表示筛选执行成功的作业。 • 未定：表示筛选正在执行的作业。 • 终止：表示筛选被手动终止的作业。 • 失败：表示筛选执行失败的作业。

按钮	说明
全部作业类型	在下拉框中选择作业类型，筛选作业。 • 全部作业类型 • MapReduce • HiveScript • Distcp • SparkScript • Spark SQL • Hive SQL • SparkSubmit • Flink
作业名称	在搜索框中根据搜索条件输入对应内容，单击，搜索作业。 • 作业名称 • 作业ID • 用户名称 • 队列名称
	单击，手动刷新作业列表。

作业执行权限说明

对于开启Kerberos认证的安全集群，用户在MRS界面提交作业时，要先执行IAM用户同步操作，同步完成后会在MRS系统中产生同IAM用户名的用户。IAM同步用户是否有提交作业权限，取决于IAM同步时，用户所绑定的IAM策略，提交作业策略请参考**IAM用户同步MRS说明**章节中**表3-3**。

用户提交作业，如果涉及到具体组件的资源使用，如HDFS的目录访问、Hive表的访问等相关组件的权限时，需由admin（Manager管理员）用户进行授权，给提交作业用户赋予相关组件权限。具体操作如下：

步骤1 使用admin用户登录Manager。

步骤2 参考**创建角色**内容，增加用户具体需要的组件权限的角色。

步骤3 参考**相关任务**修改提交作业用户所属的用户组，将新增的组件角色加入到该用户组中。

说明

用户所在用户组绑定的组件角色修改后，权限生效需要一定时间，请耐心等待。

----结束

5.5.2 运行 MapReduce 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节指导您在MRS集群页面如何提交一个新的MapReduce作业。MapReduce作业用于提交jar程序快速并行处理大量数据，是一种分布式数据处理模式和执行环境。

若在集群详情页面不支持“作业管理”和“文件管理”功能，请通过后台功能来提交作业。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

如果用户将程序包和数据文件上传至OBS中，需要先创建OBS委托并与MRS集群绑定，具体请参考**配置存算分离集群（委托方式）**。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

说明

IAM用户名存在空格时（如admin 01），不支持添加作业。

步骤6 “作业类型”选择“MapReduce”，并配置其他作业信息。

- 若集群为MRS 2.0.1及之后版本，请参考**表5-24**配置MapReduce作业信息。
- 若集群为MRS 2.0.1之前版本，请参考**表5-26**配置MapReduce作业信息。

表 5-24 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 • SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。
执行程序参数	可选参数，程序执行的关键参数。多个参数间使用空格隔开。 配置方法：*程序类名 数据输入路径 数据输出路径* • 程序类名：由用户程序内的函数指定，MRS只负责参数的传入。 • 数据输入路径：通过单击“HDFS”或者“OBS”选择或者直接手动输入正确路径。 • 数据输出路径：输出路径请手动输入一个不存在的目录。最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-25**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-25 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

表 5-26 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。示例：s3a://wordcount/program/xxx.jar – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript需要以“.sql”结尾，MapReduce和Spark需要以“.jar”结尾。sql、jar不区分大小写。
执行程序参数	程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 配置方法：包名.类名。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **说明** 用户输入带有敏感信息（如登录密码）的参数时，可通过在参数名前添加“@”的方式，为该参数值加密，以防止敏感信息被明文形式持久化。在MRS管理控制台查看作业信息时，敏感信息显示为“*”。 例如：username=admin @password=admin_123
数据输入路径	数据输入地址。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

参数	参数说明
输出路径	数据输出地址。 **说明** • 配置此参数时，单击“OBS”或“HDFS”，并选择文件目录，或者手动输入文件目录，然后单击“确定”。 • 若添加hadoop-mapreduce-examples-x.x.x.jar样例程序或和hadoop-mapreduce-examples-x.x.x.jar类似的程序，请手动输入一个不存在的目录。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 • OBS：以“s3a://”开头。（仅MRS 1.8.10及之前版本支持。） • HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。
日志路径	作业日志存储地址，该日志信息记录作业运行状态。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 • OBS：以“s3a://”开头。 • HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录任意一个Master节点，具体请参考**登录集群节点**。

步骤2 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤3 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证，则无需执行该步骤。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 执行如下命令拷贝OBS文件系统中的程序到集群的Master节点。

hadoop fs -Dfs.obs.access.key=AK -Dfs.obs.secret.key=SK -copyToLocal source_path.jar target_path.jar

例如：**hadoop fs -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX -copyToLocal "obs://mrs-word/program/hadoop-mapreduce-examples-XXX.jar" "/home/omm/hadoop-mapreduce-examples-XXX.jar"**

AK/SK可登录OBS控制台，请在集群控制台页面右上角的用户名下拉框中选择“我的凭证 > 访问密钥”页面获取。

步骤5 执行如下命令提交wordcount作业，如需从OBS读取或向OBS输出数据，需要增加AK/SK参数。

source /opt/Bigdata/client/bigdata_env;hadoop jar execute_jar wordcount input_path output_path

例如：**source /opt/Bigdata/client/bigdata_env;hadoop jar /home/omm/hadoop-mapreduce-examples-XXX.jar wordcount -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX "obs://mrs-word/input/*" "obs://mrs-word/output/"**

input_path为OBS上存放作业输入文件的路径。output_path为OBS上存放作业输出文件地址，请设置为一个不存在的目录。

----结束

5.5.3 运行 SparkSubmit 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在MRS集群页面如何提交一个新的Spark作业。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 配置作业信息。

- 若集群为MRS 2.0.1及之后版本，“作业类型”选择“SparkSubmit”，请参考**表5-27**配置SparkSubmit作业其他参数信息。

- 若集群为MRS 2.0.1之前版本，“作业类型”选择“Spark”，请参考表5-30配置Spark作业信息。

表 5-27 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考导入数据。 • SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行程序参数如表5-28。
执行程序参数	可选参数，程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考配置服务参数页面进行修改。 如需添加多个参数，请单击右侧增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如表5-29。 **说明** 如需运行长时作业如SparkStreaming等，且需要访问OBS, 需要通过“服务配置参数”传入访问OBS的AK/SK。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-28 运行程序参数

参数	参数说明	取值样例
--conf	添加任务配置项。	spark.executor.memory=2G
--driver-memory	设置driver的运行内存。	2G
--num-executors	设置executor启动数量。	5
--executor-cores	设置executor核数。	2
--class	设置任务的主类。	org.apache.spark.examples.SparkPi
--files	上传文件给任务，可以是自己定义的配置文件或者某些数据文件。来源可以是OBS或者HDFS。	-
--jars	上传任务额外依赖包，用于给任务添加任务的外部依赖包。	-
--executor-memory	设置executor内存。	2G
--conf spark-yarn.maxAppAttempts	控制AM的重试次数。	设置为0时，不允许重试；设置为1时，允许重试一次。

表 5-29 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

表 5-30 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。

参数	参数说明
执行程序路径	待执行程序包地址，需要满足如下要求： ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。示例：s3a://wordcount/program/xxx.jar – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● Spark Script需要以“.sql”结尾，MapReduce和Spark需要以“.jar”结尾。sql、jar不区分大小写。
执行程序参数	程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 配置方法：包名.类名 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **说明** 用户输入带有敏感信息（如登录密码）的参数时，可通过在参数名前添加“@”的方式，为该参数值加密，以防止敏感信息被明文形式持久化。在MRS管理控制台查看作业信息时，敏感信息显示为“*”。 例如：username=admin @password=admin_123
数据输入路径	数据输入地址。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。
输出路径	数据输出地址。 **说明** ● 配置此参数时，单击“OBS”或“HDFS”，并选择文件目录，或者手动输入文件目录，然后单击“确定”。 ● 若添加hadoop-mapreduce-examples-x.x.x.jar样例程序或和hadoop-mapreduce-examples-x.x.x.jar类似的程序，请手动输入一个不存在的目录。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

参数	参数说明
日志路径	作业日志存储地址，该日志信息记录作业运行状态。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 参考**创建用户**页面，创建一个用于提交作业的用户。

本示例创建一个用户开发场景使用的机机用户，并分配了正确的用户组（hadoop、supergroup）、主组（supergroup）和角色权限（System_administrator、default）。

步骤2 下载认证凭据。

- 对于MRS 3.x及之后版本集群，请登录FusionInsight Manager页面选择“系统 > 权限 > 用户”，在新增用户的操作列单击“更多 > 下载认证凭据”。
- 对于MRS 3.x之前版本集群，请登录MRS Manager页面选择“系统设置 > 用户管理”，在新增用户的操作列单击“更多 > 下载认证凭据”。

步骤3 将与作业相关的jar包上传到集群中，本示例使用Spark自带的样例jar包，位置在“$SPARK_HOME/examples/jars”下。

步骤4 上传**步骤2**创建的用户认证凭据到集群的“/opt”目录下，并执行如下命令解压

tar –xvf MRSTest _*xxxxxx*_keytab.tar

您将会得到user.keytab和krb5.conf两个文件。

步骤5 在对集群操作之前首先需要执行：

source /opt/Bigdata/client/bigdata_env

cd $SPARK_HOME

步骤6 提交spark作业，使用的命令如下：

./bin/spark-submit --master yarn --deploy-mode client --conf spark.yarn.principal=MRSTest --conf spark.yarn.keytab=/opt/user.keytab --class org.apache.spark.examples.SparkPi examples/jars/spark-examples_2.11-2.3.2-mrs-2.0.jar 10

参数解释：

1. yarn的计算能力，指定使用client模式提交该作业。
2. Spark作业的配置项，这里是传入了认证文件和用户名。
3. spark.yarn.principal 第一步创建的用户
4. spark.yarn.keytab 认证使用的keytab文件
5. xx.jar 作业的使用的jar。

----结束

5.5.4 运行 HiveSql 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在MRS集群页面如何提交一个新的HiveSql作业。HiveSql作业用于提交SQL语句和SQL脚本文件查询和分析数据，包括SQL语句和Script脚本两种形式，如果SQL语句涉及敏感信息，请使用Script提交。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 配置作业信息。

- 若集群为MRS 2.0.1及之后版本，“作业类型”选择“HiveSQL”，请参考**表5-31**配置HiveSQL作业信息。
- 若集群为MRS 2.0.1之前版本，“作业类型”选择“Hive Script”，请参考**表5-34**配置Hive Script作业信息。

表 5-31 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
SQL类型	SQL查询语句提交类型。 ● SQL ● Script
SQL语句	“SQL类型”参数为“SQL”时参数有效，请输入待运行的SQL语句，然后单击“检查”来检查SQL语句的正确性，确保输入语句正确。如果同时需要提交多条语句并执行，使用“;”分隔不同语句。
SQL文件	“SQL类型”参数为“Script”时参数有效，待执行SQL文件的路径，需要满足以下要求。 ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。 **说明** MRS 1.9.2及之后版本的集群，存储在OBS上的文件路径支持以“obs://”开头格式。如需使用该格式提交作业，访问OBS需要配置对应权限。 ● 创建集群时开启“OBS权限控制”功能时，可直接使用“obs://”路径，无需单独配置。 ● 创建集群时未开启或不支持“OBS权限控制”功能时，请参考**访问OBS**页面进行配置。
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行参数如**表5-32**。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-33**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-32 运行程序参数

参数	参数说明	取值样例
--hiveconf	设置Hive服务相关配置，例如设置执行引擎为MR。	设置执行引擎为MR：--hiveconf "hive.execution.engine=mr"
--hivevar	设置用户自定义变量，例如设置变量id。	设置变量id：--hivevar id="123" select * from test where id = ${hivevar:id};

表 5-33 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-
hive.execution.engine	选择执行作业的引擎。	• mr • tez

表 5-34 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。示例：s3a://wordcount/program/xxx.jar – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 • SparkScript需要以“.sql”结尾，MapReduce和Spark需要以“.jar”结尾。sql、jar不区分大小写。

参数	参数说明
执行程序参数	程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 配置方法：包名.类名 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **说明** 用户输入带有敏感信息（如登录密码）的参数时，可通过在参数名前添加“@”的方式，为该参数值加密，以防止敏感信息被明文形式持久化。在MRS管理控制台查看作业信息时，敏感信息显示为“*”。 例如：username=admin @password=admin_123
数据输入路径	数据输入地址。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。
输出路径	数据输出地址。 **说明** ● 配置此参数时，单击“OBS”或“HDFS”，并选择文件目录，或者手动输入文件目录，然后单击“确定”。 ● 若添加hadoop-mapreduce-examples-x.x.x.jar样例程序或和hadoop-mapreduce-examples-x.x.x.jar类似的程序，请手动输入一个不存在的目录。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。
日志路径	作业日志存储地址，该日志信息记录作业运行状态。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ● OBS：以“s3a://”开头。 ● HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

步骤1 登录任意一个Master节点，具体请参考**登录集群节点**。

步骤2 执行如下命令初始化环境变量。

source /opt/BigData/client/bigdata_env

说明

- MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。
- 若安装了Hive多实例，在使用客户端连接具体Hive实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载Hive2实例变量：

  source /opt/BigData/client/Hive2/component_env

步骤3 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证(普通模式），则无需执行该步骤。

kinit *MRS集群用户* (用户需要有hive组）

步骤4 执行beeline连接hiveserver，运行任务。

beeline

普通模式，则执行以下命令，如果不指定组件业务用户，则会以当前操作系统用户连接hiveserver。

beeline -n*组件业务用户*

beeline -f sql文件（执行文件里的sql）

----结束

5.5.5 运行 SparkSql 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在MRS集群页面如何提交一个新的SparkSql作业。SparkSQL作业用于查询和分析数据，包括SQL语句和Script脚本两种形式，如果SQL语句涉及敏感信息，请使用Spark Script提交。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 若集群为MRS 2.0.1及之后版本，请单击“添加”，进入添加作业页面，“作业类型”选择“SparkSql”，作业参考**表5-35**配置SparkSql作业信息。

表 5-35 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
SQL类型	SQL查询语句提交类型。 ● SQL ● Script
SQL语句	“SQL类型”参数为“SQL”时参数有效，请输入待运行的SQL语句，然后单击“检查”来检查SQL语句的正确性，确保输入语句正确。如果同时需要提交多条语句并执行，使用“;”分隔不同语句。
SQL文件	“SQL类型”参数为“Script”时参数有效，待执行SQL文件的路径，需要满足以下要求。 ● 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 ● 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 ● SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。 **说明** MRS 2.0.1及之后版本的集群，存储在OBS上的文件路径支持以“obs://”开头格式。如需使用该格式提交作业，访问OBS需要配置对应权限。 ● 创建集群时开启“OBS权限控制”功能时，可直接使用“obs://”路径，无需单独配置。 ● 创建集群时未开启或不支持“OBS权限控制”功能时，请参考**访问OBS**页面进行配置。

参数	参数说明
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行程序参数如**表5-36**。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧 增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-37**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-36 运行程序参数

参数	参数说明	取值样例
--conf	添加任务配置项	spark.executor.memory=2G
--driver-memory	设置driver的运行内存	2G
--num-executors	设置executor启动数量	5
--executor-cores	设置executor核数	2
--jars	上传任务额外依赖包，用于给任务添加任务的外部依赖包	-
--executor-memory	设置executor内存	2G

表 5-37 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤6 若集群为MRS 2.0.1及之前版本，按照如下方法添加SparkScript作业和Spark SQL作业。

- SparkScript作业：在作业页签单击“添加”进入添加作业页面，“作业类型”选择“SparkScript”，请参考**表5-38**配置作业信息。
- Spark SQL作业：选择Spark SQL页签，添加SQL语句，然后检查并提交SQL语句。

表 5-38 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： – 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 – 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 ▪ OBS：以“s3a://”开头。示例：s3a://wordcount/program/xxx.jar ▪ HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 – SparkScript需要以“.sql”结尾，MapReduce和Spark需要以“.jar”结尾。sql、jar不区分大小写。
执行程序参数	程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 配置方法：包名.类名。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **说明** 用户输入带有敏感信息（如登录密码）的参数时，可通过在参数名前添加“@”的方式，为该参数值加密，以防止敏感信息被明文形式持久化。在MRS管理控制台查看作业信息时，敏感信息显示为“*”。 例如：username=admin @password=admin_123
数据输入路径	数据输入地址。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

参数	参数说明
输出路径	数据输出地址。 **说明** – 配置此参数时，单击“OBS”或“HDFS”，并选择文件目录，或者手动输入文件目录，然后单击“确定”。 – 若添加hadoop-mapreduce-examples-x.x.x.jar样例程序或和hadoop-mapreduce-examples-x.x.x.jar类似的程序，请手动输入一个不存在的目录。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。 – HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。
日志路径	作业日志存储地址，该日志信息记录作业运行状态。 数据可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“s3a://”开头。 – HDFS：以“/user”开头。 最多为1023字符，不能包含;|&>,<'$特殊字符，可为空。

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 参考**创建用户**页面，创建一个用于提交作业的用户。

本示例创建一个用户开发场景使用的机机用户，并分配了正确的用户组（hadoop、supergroup）、主组（supergroup）和角色权限（System_administrator、default）。

步骤2 下载认证凭据。

- 对于MRS 3.x及之后版本集群，请登录FusionInsight Manager页面选择“系统 > 权限 > 用户”，在新增用户的操作列单击“更多 > 下载认证凭据”。
- 对于MRS 3.x之前版本集群，请登录MRS Manager页面选择“系统设置 > 用户管理”，在新增用户的操作列单击“更多 > 下载认证凭据”。

步骤3 登录Spark客户端所在节点，上传2创建的用户认证凭据到集群的“/opt”目录下，并执行如下命令解压：

tar –xvf MRSTest *_xxxxxx_***keytab.tar**

得到user.keytab和krb5.conf两个文件。

步骤4 在对集群操作之前首先需要执行：

source /opt/Bigdata/client/bigdata_env

cd $SPARK_HOME

步骤5 打开spark-sql命令行，进入spark-sql命令行后可执行SQL语句，执行命令如下：

./bin/spark-sql --conf spark.yarn.principal=MRSTest --conf spark.yarn.keytab=/opt/user.keytab

若需要执行SQL文件，需要上传SQL文件（如上传到“/opt/”目录），上传文件后执行命令如下：

./bin/spark-sql --conf spark.yarn.principal=MRSTest --conf spark.yarn.keytab=/opt/user.keytab -f /opt/script.sql

----结束

5.5.6 运行 Flink 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节指导用户在MRS集群页面如何提交一个新的Flink作业。Flink作业用于提交jar程序处理流式数据。

前提条件

用户已经将运行作业所需的程序包和数据文件上传至OBS系统或HDFS中。

通过界面提交作业

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，具体介绍请参考**IAM用户同步MRS说明**。

说明

- 当IAM用户的用户组的所属策略从MRS ReadOnlyAccess向MRS CommonOperations、MRS FullAccess、MRS Administrator变化时，由于集群节点的SSSD（System Security Services Daemon）缓存刷新需要时间，因此同步完成后，请等待5分钟，等待新修改策略生效之后，再进行提交作业。否则，会出现提交作业失败的情况。
- 当IAM用户的用户组的所属策略从MRS CommonOperations、MRS FullAccess、MRS Administrator向MRS ReadOnlyAccess变化时，由于集群节点的SSSD缓存刷新需要时间，因此同步完成后，请等待5分钟，新修改策略才能生效。

步骤4 单击“作业管理”，进入“作业管理”页签。

步骤5 单击“添加”，进入“添加作业”页面。

步骤6 “作业类型”选择“Flink”，参考**表 1**配置Flink作业信息。

表 5-39 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： • 最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。 • 执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。 – OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。 – HDFS：以“/user”开头。数据导入HDFS请参考**导入数据**。
运行程序参数	可选参数，为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 常用运行程序参数如**表5-40**。
执行程序参数	可选参数，程序执行的关键参数，该参数由用户程序内的函数指定，MRS只负责参数的传入。多个参数间使用空格隔开。 最多为150000字符，不能包含;|&><'$特殊字符，可为空。 **注意** 若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。
服务配置参数	可选参数，用于为本次执行的作业修改服务配置参数。该参数的修改仅适用于本次执行的作业，如需对集群永久生效，请参考**配置服务参数**页面进行修改。 如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。 常用服务配置参数如**表5-41**。
命令参考	用于展示提交作业时提交到后台执行的命令。

表 5-40 运行程序参数

参数	参数说明	取值样例
-ytm	设置每个TaskManager容器的内存（单位可选, 默认单位：MB）。	1024
-yjm	设置JobManager容器内存（单位可选，默认单位：MB）。	1024
-yn	设置分配给应用程序的Yarn容器的数量，该值与TaskManager数量相同。	2

参数	参数说明	取值样例
-ys	设置TaskManager的核数。	2
-ynm	自定义Yarn上应用程序名称。	test
-c	设置程序入口点的类（如“main”或“getPlan()”方法）。该参数仅在JAR文件未指定其清单的类时需要。	com.bigdata.mrs.test

说明

针对MRS 3.x及之后版本，运行程序参数不支持“-yn”。

表 5-41 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

步骤7 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

通过后台提交作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录MRS客户端。

步骤2 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤3 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 使用新创建的用户登录Manager页面。
   - MRS 3.x之前版本，登录集群的Manager界面，选择“系统设置 > 用户管理”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
   - MRS 3.x及之后版本，登录集群的Manager界面，选择“系统 > 权限 > 用户”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/Bigdata/client/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。

4. MRS 3.x及之后版本，安全模式下需要将客户端安装节点的业务IP以及Manager的浮动ip追加到“/opt/Bigdata/client/Flink/flink/conf/flink-conf.yaml”文件中的“jobmanager.web.allow-access-address”配置项中，ip之间使用英文逗号分隔。
5. 配置安全认证，在“/opt/Bigdata/client/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/Bigdata/client/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test
6. 在Flink的客户端bin目录下，执行如下命令进行安全加固，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/Bigdata/client/Flink/flink/conf/flink-conf.yaml”中关于SSL的值，针对MRS 3.x之前版本，安全集群默认没有开启外部SSL，用户如果需要启用外部SSL，进行配置后再次运行该脚本即可，配置参数在MRS的Flink默认配置中不存在，用户如果开启外部连接SSL，则需要添加**表5-42**中参数。

表 5-42 参数描述

参数	参数值示例	描述
security.ssl.rest.enabled	true	打开外部SSL开关。
security.ssl.rest.keystore	${path}/flink.keystore	keystore的存放路径。
security.ssl.rest.keystore-password	123456	keystore的password，“123456”表示需要用户输入自定义设置的密码值。
security.ssl.rest.key-password	123456	ssl key的password，“123456”表示需要用户输入自定义设置的密码值。
security.ssl.rest.truststore	${path}/flink.truststore	truststore存放路径。
security.ssl.rest.truststore-password	123456	truststore的password，“123456”表示需要用户输入自定义设置的密码值。

说明

- 针对MRS 3.x之前版本，generate_keystore.sh脚本无需手动生成。
- **认证和加密**会将生成的flink.keystore、flink.truststore、security.cookie自动填充到“flink-conf.yaml”对应配置项中。
- 针对MRS 3.x及之后版本，“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值需要使用Manager明文加密API进行获取：

  curl -k -i -u <*user name*>:<*password*> -X POST -HContent-type:application/json -d '{"plainText":"<*password*>"}' 'https://*x.x.x.x*:28443/web/api/v2/tools/encrypt'；其中<*password*>要与签发证书时使用的密码一致，*x.x.x.x*为集群Manager的浮动IP。

7. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/Bigdata/client/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/Bigdata/client/Flink/flink/conf/”目录下新建目录，例如ssl。
     ii. 移动flink.keystore和flink.truststore文件到“/opt/Bigdata/client/Flink/flink/conf/ssl/”中。
     iii. 针对MRS 3.x及之后版本，修改flink-conf.yaml文件中如下两个参数为相对路径。
     ```
     security.ssl.keystore: ssl/flink.keystore
     security.ssl.truststore: ssl/flink.truststore
     ```
     iv. 针对MRS 3.x之前版本，修改flink-conf.yaml文件中如下两个参数为相对路径。
     ```
     security.ssl.internal.keystore: ssl/flink.keystore
     security.ssl.internal.truststore: ssl/flink.truststore
     ```
8. 如果客户端安装在集群外节点，请在配置文件（如：**“/opt/Bigdata/client/Flink/fink/conf/flink-conf.yaml”**）中增加如下配置值，其中**xx.xx.xxx.xxx**请替换为客户端所在节点的IP。
   ```
   web.access-control-allow-origin: xx.xx.xxx.xxx
   jobmanager.web.allow-access-address: xx.xx.xxx.xxx
   ```

步骤4 运行wordcount作业。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。
    ```
    yarn-session.sh -nm "session-name"
    flink run /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```
  - 执行如下命令在Yarn上提交单个作业。
    ```
    flink run -m yarn-cluster /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```
- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。
      ```
      yarn-session.sh -nm "session-name"
      flink run /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
      ```
    - 执行如下命令在Yarn上提交单个作业。
      ```
      flink run -m yarn-cluster /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
      ```

- flink.keystore和flink.truststore文件路径为相对路径时：
  - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/Bigdata/client/Flink/flink/conf/”，则在“opt/Bigdata/client/Flink/flink/conf/”目录下执行命令。
    ```
    yarn-session.sh -t ssl/ -nm "session-name"
    flink run /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```
  - 执行如下命令在Yarn上提交单个作业。
    ```
    flink run -m yarn-cluster -yt ssl/ /opt/Bigdata/client/Flink/flink/examples/streaming/WordCount.jar
    ```

----结束

5.5.7 运行 Kafka 作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。本章节教您在Kafka主题中产生和消费消息。

暂不支持通过界面提交Kafka作业，请通过后台功能来提交作业。

通过后台提交作业

先查询ZooKeeper和Kafka的实例地址，再运行Kafka作业。

查询实例地址（3.x版本）

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 请参考**访问FusionInsight Manager（MRS 3.x及之后版本）**，跳转至FusionInsight Manager页面。然后选择“服务 > ZooKeeper > 实例”，查看ZooKeeper角色实例的IP地址。记录ZooKeeper角色实例中任意一个的IP地址即可。

步骤4 选择“服务 > Kafka > 实例”，查看Kafka角色实例的IP地址。记录Kafka角色实例中任意一个的IP地址即可。

----结束

查询实例地址（3.x之前版本）

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

步骤3 在MRS集群详情页面，选择“组件管理 > ZooKeeper > 实例”，查看ZooKeeper角色实例的IP地址。记录ZooKeeper角色实例中任意一个的IP地址即可。

步骤4 选择“组件管理 > Kafka > 实例”，查看Kafka角色实例的IP地址。记录Kafka角色实例中任意一个的IP地址即可。

----结束

运行Kafka作业

MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。

步骤1 登录Master2节点，具体请参考**登录集群节点**。

步骤2 执行如下命令初始化环境变量。

source /opt/Bigdata/client/bigdata_env

步骤3 如果当前集群已开启Kerberos认证，执行以下命令认证当前用户。如果当前集群未开启Kerberos认证，则无需执行该步骤。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 执行如下命令，创建kafka topic。

kafka-topics.sh --create --zookeeper <ZooKeeper角色实例IP:2181/kafka> --partitions 2 --replication-factor 2 --topic <Topic名称>

步骤5 在topic test中产生消息。

首先执行命令**kafka-console-producer.sh --broker-list <Kafka角色实例IP:9092> --topic <Topic名称> --producer.config /opt/Bigdata/client/Kafka/kafka/config/producer.properties**。

然后输入指定的内容作为生产者产生的消息，输入完成后按回车发送消息。如果需要结束产生消息，使用“Ctrl + C”退出任务。

步骤6 消费topic test中的消息。

kafka-console-consumer.sh --topic <Topic名称> --bootstrap-server <Kafka角色实例IP:9092> --consumer.config /opt/Bigdata/client/Kafka/kafka/config/consumer.properties

说明

如果集群开启Kerberos认证，则执行如上两个命令时请修改端口号9092为21007，详见**开源组件端口列表**。

----结束

5.5.8 查看作业配置信息和日志

本章节介绍如何查看作业的配置信息和运行日志信息。

背景信息

- 支持查看所有作业的配置信息。
- 只有运行中的作业才能查看运行日志信息。

  由于Spark SQL和Distcp作业在后台无日志，运行中的Spark SQL和Distcp作业不能查看运行日志信息。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名，进入集群基本信息页面。

步骤3 选择“作业管理”。

步骤4 在需要查看作业对应的“操作”列中，单击“查看详情”。

弹出“查看详情”窗口，显示该作业的配置信息。

步骤5 选择一个运行中的作业，在作业对应的“操作”列中，单击“查看日志”。

弹出一个新页面，显示作业执行的实时日志信息。

每个租户并发提交作业和查看日志的个数均为10。

----结束

5.5.9 停止作业

本章节介绍如何手动停止正在运行的MRS作业。

背景信息

Spark SQL作业不支持停止。作业停止后状态更新为“已终止”，并且该作业不可重新执行。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名。

进入集群基本信息页面。

步骤3 选择“作业管理”。

步骤4 选择一个运行中的作业，在作业对应的“操作”列中，选择“更多 > 停止”。

作业状态由“运行中”更新为“已终止”。

----结束

5.5.10 删除作业

本章节介绍如何删除MRS作业，作业执行完成后，若不需要再查看使用其相关信息，可以选择删除作业。

背景信息

支持删除单个作业和批量删除作业。作业删除后不可恢复，请谨慎操作。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名。

进入集群基本信息页面。

步骤3 选择“作业管理”。

步骤4 在需要删除作业对应的“操作”列中，选择“更多 > 删除”。

此处只能删除单个作业。

步骤5 勾选多个作业，单击作业列表左上方的“删除”。

可以删除一个、多个或者全部作业。

----结束

5.5.11 使用 OBS 加密数据运行作业

MRS 1.9.xMRS 2.1.x版本支持使用OBS文件系统中加密后的数据来运行作业，同时支持将加密后的作业运行结果存储在OBS文件系统中。目前仅支持通过OBS协议访问数据。

OBS支持使用KMS密钥的加解密方式对数据进行加解密，所有的加解密操作都在OBS完成，同时密钥管理在DEW服务。

如需在MRS中使用OBS加密功能，用户需要有“KMS Administrator”权限，且需要在相应组件进行如下配置。

说明

如果集群同时开启“**OBS权限控制**”功能，此时会使用ECS配置的默认委托“MRS_ECS_DEFAULT_AGENCY”或者用户设置的自定义委托的AK/SK访问OBS服务，同时OBS服务会使用接收到的AK/SK访问密码安全中心获取KMS密钥状态，因此需要在使用的委托上绑定“KMS Administrator”策略，否则在处理加密数据时OBS会返回“403 Forbidden”的错误信息。目前MRS服务会在默认委托“MRS_ECS_DEFAULT_AGENCY”绑定“KMS Administrator”策略，用户使用的自定义委托则需要用户自己绑定。

前提条件

如需使用OBS加密功能，请先配置MRS访问OBS功能，具体请参考**配置存算分离集群（委托方式）**。

Hive 配置

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > Hive > 服务配置”。

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-43 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。

参数	取值	说明
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，根据界面提示保存修改的参数。

----结束

Hadoop 配置

方式一：通过界面配置。

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > HDFS > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-44 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

步骤5 以root用户登录Master节点，密码为用户创建集群时设置的root密码（若集群存在多个Master节点，请分别登录每个Master节点进行**步骤5**~**步骤7**的操作）。

步骤6 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤7 执行以下命令更新客户端配置，并输入用户名和密码，用户名为admin，密码为用户创建集群时设置的admin密码。

./ autoRefreshConfig.sh

----结束

方式二：通过客户端配置文件配置。

在Master节点上的客户端配置文件（例如“/opt/Bigdata/client/HDFS/hadoop/etc/hadoop/core-site.xml”）中的增加如下参数配置（若集群存在多个Master节点，请分别登录每个Master节点进行该操作）。

表 5-45 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 • true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 • false：关闭安全连接。

HBase 配置

方式一：通过界面配置。

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > HBase > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-46 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 • true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 • false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

步骤5 以root用户登录Master节点，密码为用户创建集群时设置的root密码（若集群存在多个Master节点，请分别登录每个Master节点进行**步骤5**~**步骤7**的操作）。

步骤6 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤7 执行以下命令更新客户端配置，并输入用户名和密码，用户名为admin，密码为用户创建集群时设置的admin密码。

./ autoRefreshConfig.sh

----结束

方式二：通过客户端配置文件配置。

在Master节点上的客户端配置文件（例如“/opt/Bigdata/client/HBase/hbase/conf/core-site.xml”）中的增加如下参数配置（若集群存在多个Master节点，请分别登录每个Master节点进行该操作）。

表 5-47 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	• SSE-KMS：表示使用KMS密钥的加解密方式。 • NONE：表示关闭加密功能。

参数	取值	说明
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

Spark 配置

方式一：通过界面配置。

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > Spark > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-48 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

步骤5 以root用户登录Master节点，密码为用户创建集群时设置的root密码（若集群存在多个Master节点，请分别登录每个Master节点进行步骤5~步骤7的操作）。

步骤6 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤7 执行以下命令更新客户端配置，并输入用户名和密码，用户名为admin，密码为用户创建集群时设置的admin密码。

./autoRefreshConfig.sh

----结束

方式二：通过客户端配置文件配置。

在Master节点上的客户端配置文件（例如“/opt/Bigdata/client/Spark/spark/conf/core-site.xml”）中的增加如下参数配置（若集群存在多个Master节点，请分别登录每个Master节点进行该操作）。

表 5-49 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

Presto 配置

步骤1 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

步骤2 选择“组件管理 > Presto > 服务配置”

步骤3 将“基础配置”切换为“全部配置”，搜索并配置如下参数：

表 5-50 数据加密参数

参数	取值	说明
fs.obs.server-side-encryption-type	SSE-KMS	● SSE-KMS：表示使用KMS密钥的加解密方式。 ● NONE：表示关闭加密功能。

参数	取值	说明
fs.obs.server-side-encryption-key	-	表示用来加密的KMS密钥ID。该参数可不配置。 当参数“fs.obs.server-side-encryption-type”配置为“SSE-KMS”且该参数未配置时，OBS会使用OBS服务的默认KMS密钥完成加密。
fs.obs.connection.ssl.enabled	true	标识是否与OBS建立安全连接。 ● true：开启安全连接，当需要使用OBS加解密功能时该参数必须配置为“true”。 ● false：关闭安全连接。

步骤4 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。

----结束

5.5.12 配置作业消息通知

MRS联合消息通知服务（SMN），采用主题订阅模型，提供一对多的消息订阅以及通知功能，能够实现一站式集成多种推送通知方式（短信和邮件通知）。通过配置作业消息通知可以实现您在作业执行成功或作业执行失败时能立即接收到通知。

操作步骤

步骤1 登录管理控制台。

步骤2 单击“服务列表”选择“管理与监管 > 消息通知服务”，进入消息通知服务页面。

步骤3 创建主题并向主题中添加订阅，具体请参考**配置消息通知**。

步骤4 进入MRS管理控制台，单击集群名称进入集群详情页面。

步骤5 选择“告警管理 > 消息订阅规则 > 添加消息订阅规则”。

步骤6 配置向订阅者发送作业执行结果消息的规则。

表 5-51 消息订阅规则参数说明

参数	说明
规则名称	用户自定义发送订阅消息的规则名称，只能包含数字、英文字符、中划线和下划线。
提醒通知	选择开启，将向订阅者发送对应订阅消息。
主题名称	选择已创建的主题，也可以单击“创建主题”重新创建。
消息类型	选择“事件”。

参数	说明
订阅规则	1. 单击“提示”前的。 2. 单击“Manager”前的。 3. 勾选“作业执行成功”和“作业执行失败”。

----结束

5.6 组件管理

5.6.1 对象管理简介

MRS集群包含了各类不同的基本对象，不同对象的描述介绍如表5-52所示：

表 5-52 MRS 基本对象概览

对象	描述	举例
服务	可以完成具体业务的一类功能集合。	例如KrbServer服务和LdapServer服务。
服务实例	服务的具体实例，一般情况下可使用服务表示。	例如KrbServer服务。
服务角色	组成一个完整服务的一类功能实体，一般情况下可使用角色表示。	例如KrbServer由KerberosAdmin角色和KerberosServer角色组成。
角色实例	服务角色在主机节点上运行的具体实例。	例如运行在Host2上的KerberosAdmin，运行在Host3上的KerberosServer。
主机	一个弹性云服务器，可以运行Linux系统。	例如Host1～Host5。
机架	一组包含使用相同交换机的多个主机集合的物理实体。	例如Rack1，包含Host1～Host5。
集群	由多台主机组成的可以提供多种服务的逻辑实体。	例如名为Cluster1的集群由（Host1～Host5）5个主机组成，提供了KrbServer和LdapServer等服务。

5.6.2 查看配置

用户可以在MRS上查看服务（含角色）和角色实例的配置。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

- 查看服务的配置。
  a. 在集群详情页，单击“组件管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“服务配置”。
  d. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
  e. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

     在服务节点下的参数属于服务配置参数，在角色节点下的参数是角色配置参数。
  f. 在“——请选择——”选项中选择“非默认”，界面上显示参数值为非默认值的参数（MRS 3.x之前版本存在此选项）。
- 查看角色实例的配置。
  a. 在集群详情页，单击“组件管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“实例”页签。
  d. 单击角色实例列表中指定的角色实例名称。
  e. 单击“实例配置”。
  f. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
  g. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
  h. 在“——请选择——”选项中选择“非默认”，界面上显示参数值为非默认值的参数（MRS 3.x之前版本存在此选项）。

5.6.3 管理服务操作

用户可以在MRS：

- 启动操作状态为“已停止”、“停止失败”或“启动失败”服务，以使用该服务。
- 停止不再使用或异常服务。
- 重启异常或配置过期的服务，以恢复或生效服务功能。

前提条件

- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

对系统影响

- 在Task节点组无法添加stateful的组件。

添加服务

步骤1 在集群详情页，选择“组件管理 > 添加服务”。

步骤2 在服务列表中，勾选需要添加的服务，单击“下一步”。

说明

- 添加服务时，该服务依赖的下层服务也会自动勾选，可以同时添加多个服务。
- 只能在状态正常的节点上添加服务。
- 未安装Hadoop的集群，在添加Hadoop后，需要刷新MRS控制台的集群详情页面，并进行IAM用户同步后，才能在作业管理中成功提交作业。
- 集群不支持添加Hadoop服务的单个组件，只能添加Hadoop服务，Hadoop服务包含MapReduce、Yarn和HDFS。
- 添加Spark2x组件后，如需在Hue WEBUI操作SparkSQL，请先重启Hue服务。

步骤3 在“拓扑调整”页面勾选服务部署的节点，部署方案可参考表4-8。

步骤4 单击“确认”。服务添加完成后，在“组件管理”页面可以查看所添加的服务。

说明

在Console上添加服务会自动同步到Manager页面。

----结束

删除服务

步骤1 在集群详情页，单击“组件管理”。

步骤2 在指定服务所在行，单击“删除”操作。

说明

- 如果需要删除的服务存在上层依赖，则不可删除，每次只能删除一个服务。
- 不支持删除Hadoop（HDFS、Yarn、MapReduce）、Ranger、DBService、KrbServer、LdapServer和meta服务，其他已安装服务可删除。

步骤3 在弹出的对话框中单击“是”，删除该服务。

注意

- 在Console上删除服务会自动同步到Manager页面。
- 删除服务前，请先备份该服务的数据，避免数据丢失。

----结束

启动、停止和重启服务

步骤1 在集群详情页，单击“组件管理”。

步骤2 在指定服务所在行，单击“启动”、“停止”和“重启”执行启动、停止和重启操作。

服务之间存在依赖关系。对某服务执行启动、停止和重启操作时，与该服务存在依赖关系的服务将受到影响。

具体影响如下：

- 启动某服务，该服务依赖的下层服务需先启动，服务功能才可生效。
- 停止某服务，依赖该服务的上层服务将无法提供功能。
- 重启某服务，依赖该服务且启动的上层服务需重启后才可生效。

----结束

5.6.4 配置服务参数

用户可以根据实际业务场景，在MRS中快速查看和修改服务默认的配置，及导出或导入配置。

对系统的影响

- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。
- 集群中只剩下一个DBService角色实例时，不支持修改DBService服务的参数。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

修改服务参数

1. 在集群详情页，单击“组件管理”。
2. 单击服务列表中指定的服务名称。
3. 单击“服务配置”。
4. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
5. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

   修改某个参数的值后需要取消修改，可以单击恢复。
6. 单击“保存配置”，根据界面提示保存参数并重启服务。

   说明

   MRS 3.x之前版本更新YARN服务队列的配置且不重启服务时，在服务状态页签选择“更多 > 刷新队列”更新队列使配置生效。

5.6.5 配置服务自定义参数

MRS各个组件支持开源的所有参数，在MRS支持修改部分关键使用场景的参数，且部分组件的客户端可能不包含开源特性的所有参数。如果需要修改其他MRS未直接支持的组件参数，用户可以在MRS通过自定义配置项功能为组件添加新参数。添加的新参数最终将保存在组件的配置文件中并在重启后生效。

对系统的影响

- 配置服务属性后，需要重启此服务，重启期间无法访问服务。
- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

前提条件

- 用户已充分了解需要新添加的参数意义、生效的配置文件以及对组件的影响。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“服务配置”。

步骤4 将页面右侧“基础配置”切换为“全部配置”。

步骤5 在左侧导航栏选择“自定义”，MRS将显示当前组件的自定义参数。

“参数文件”显示保存用户新添加的自定义参数的配置文件。每个配置文件中可能支持相同名称的开源参数，设置不同参数值后生效结果由组件加载配置文件的顺序决定。自定义参数支持服务级别与角色级别，请根据业务实际需要选择。不支持单个角色实例添加自定义参数。

步骤6 根据配置文件与参数作用，在对应参数项所在行“参数”列输入组件支持的参数名，在“值”列输入此参数的参数值。

- 支持单击 和 增加或删除一条自定义参数。第一次单击添加自定义参数后才支持删除操作。
- 修改某个参数的值后需要取消修改，可以单击 恢复。

步骤7 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

任务示例

配置Hive自定义参数

Hive依赖于HDFS，默认情况下Hive访问HDFS时是HDFS的客户端，生效的配置参数统一由HDFS控制。例如HDFS参数“ipc.client.rpc.timeout”影响所有客户端连接HDFS服务端的RPC超时时间，如果用户需要单独修改Hive连接HDFS的超时时间，可以使用自定义配置项功能进行设置。在Hive的“core-site.xml”文件增加此参数可被Hive服务识别并代替HDFS的设置。

步骤1 在集群详情页，单击“组件管理”。

步骤2 选择 “Hive > 服务配置”。

步骤3 将页面右侧“基础配置”切换为“全部配置”。

步骤4 在左侧导航栏选择Hive服务级别“自定义”，MRS将显示Hive支持的服务级别自定义参数。

步骤5 在“core-site.xml”对应参数“core.site.customized.configs”的“参数”输入“ipc.client.rpc.timeout”，“值”输入新的参数值，例如“150000”。单位为毫秒。

步骤6 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

5.6.6 同步服务配置

操作场景

当用户发现部分服务的“配置状态”为“配置超期”或“配置失败”时，您可以尝试使用同步配置功能，以恢复配置状态。或者集群中所有服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

对系统的影响

同步服务配置后，需要重启配置过期的服务。重启时对应的服务不可用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 在服务列表中，单击指定服务名称。

步骤3 在服务状态页签，选择“更多 > 同步配置”。

步骤4 在弹出窗口勾选“重启配置过期的服务”，并单击“是”重启配置过期的服务。

----结束

5.6.7 管理角色实例操作

操作场景

用户可以在MRS启动操作状态为“停止”、“停止失败”或“启动失败”角色实例，以使用该角色实例，也可以停止不再使用或异常的角色实例，或者重启异常的角色实例，以恢复角色实例功能。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“实例”页签。

步骤4 勾选待操作角色实例前的复选框。

步骤5 选择“更多 > 启动实例”、“停止实例”、“重启实例”或“滚动重启实例”等，执行相应操作。

----结束

5.6.8 配置角色实例参数

操作场景

用户可以根据实际业务场景，在MRS中快速查看及修改角色实例默认的配置。支持导出或导入配置。

对系统的影响

配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

修改角色实例参数

1. 在集群详情页，单击“组件管理”。
2. 单击服务列表中指定的服务名称。
3. 单击“实例”页签。
4. 单击角色实例列表中指定的角色实例名称。
5. 单击“实例配置”页签。
6. 将页面右侧“基础配置”切换为“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
7. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

   修改某个参数的值后需要取消修改，可以单击恢复。
8. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”，重启角色实例。

5.6.9 同步角色实例配置

操作场景

当用户发现角色实例的“配置状态”为“配置超期”或“配置失败”时，可以在MRS尝试使用同步配置功能，同步角色实例的配置数据与后台配置数据，以恢复配置状态。

对系统的影响

同步配置角色实例后需要重启配置过期的角色实例。重启时对应的角色实例不可用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 选择服务名称。

步骤3 单击“实例”页签。

步骤4 在角色实例列表中，单击指定角色实例名称。

步骤5 在角色实例状态及指标信息上方，选择“更多 > 同步配置”。

步骤6 在弹出窗口勾选“重启配置过期的服务”，并单击“是”重启角色实例。

----结束

5.6.10 退服和入服角色实例

操作场景

某个Core或Task节点出现问题时，可能导致整个集群状态显示为“异常”。MRS集群支持将数据存储在不同Core节点，用户可以在MRS指定角色实例退服，使退服的角色实例不再提供服务。在排除故障后，可以将已退服的角色实例入服。

支持退服、入服的角色实例包括：

- HDFS的DataNode角色实例
- Yarn的NodeManager角色实例
- HBase的RegionServer角色实例
- ClickHouse的ClickHouseServer角色实例
- Kafka的Broker角色实例

限制：

- 当DataNode数量少于或等于HDFS的副本数时，不能执行退服操作。例如HDFS副本数为3时，则系统中少于4个DataNode，将无法执行退服，MRS在执行退服操作时会等待30分钟后报错并退出执行。

- Kafka Broker数量少于或等于副本数时，不能执行退服。例如Kafka副本数为2时，则系统中少于3个节点，将无法执行退服，MRS执行退服操作时会失败并退出执行。
- 已经退服的角色实例，必须执行入服操作启动该实例，才能重新使用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“组件管理”。

步骤2 单击服务列表中相应服务。

步骤3 单击“实例”页签。

步骤4 勾选指定角色实例名称前的复选框。

步骤5 选择“更多 > 退服”或“入服”执行相应的操作。

说明

实例退服操作未完成时在其他浏览器窗口重启集群中相应服务，可能导致MRS提示停止退服，实例的“操作状态”显示为“已启动”。实际上后台已将该实例退服，请重新执行退服操作同步状态。

----结束

5.6.11 启动及停止集群

集群是包含着服务组件的集合。用户可以启动或者停止集群中所有服务。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

启动及停止集群

在集群详情页，单击页面右上角“管理操作 > 启动所有组件”或“停止所有组件”执行相应的操作。

5.6.12 同步集群配置

操作场景

当MRS显示全部服务或部分服务的“配置状态”为“过期”或“失败”时，用户可以尝试使用同步配置功能，以恢复配置状态。

- 若集群中所有服务的配置状态为“失败”时，同步集群的配置数据与后台配置数据。
- 若集群中某些服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

> **说明**
>
> **MRS 3.x**版本暂不支持在管理控制台执行本章节操作。

对系统的影响

同步集群配置后，需要重启配置过期的服务。重启时对应的服务不可用。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击页面右上角“配置 > 同步配置”。

步骤2 在弹出窗口勾选“重启配置过期的服务或实例。”，并单击“确定”，重启配置过期的服务。

界面提示“操作成功”，单击“完成”，集群成功启动。

----结束

5.6.13 导出集群的配置数据

操作场景

为了满足实际业务的需求，用户可以在MRS中将集群所有配置数据导出，导出文件用于快速更新服务配置。

> **说明**
>
> **MRS 3.x**版本暂不支持在管理控制台执行本章节操作。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

在集群详情页，单击页面右上角“配置 > 导出集群配置”。

导出文件用于更新服务配置，请参见**配置服务参数**中**导入服务配置参数**。

5.6.14 支持滚动重启

在修改了大数据组件的配置项后，需要重启对应的服务来使得配置生效，使用普通重启方式会并发重启所有服务或实例，可能引起业务断服。为了确保服务重启过程中，尽量减少或者不影响业务运行，可以通过滚动重启来按批次重启服务或实例（对于有主备状态的实例，会先重启备实例，再重启主实例）。滚动重启方式的重启时间比普通重启时间久。

当前MRS集群中，服务和实例是否支持滚动重启如表5-53所示。

表 5-53 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	

服务	实例	是否支持滚动重启
	Supervisor	
	Logviewer	

使用限制

- 请在低业务负载时间段进行滚动重启操作。
  - 例如：在滚动重启kafka服务时候， 如果kafka服务业务吞吐量很高（100M/s以上的情况下），会出现kafka服务滚动重启失败的情况。
  - 例如：在滚动重启HBase服务时候，如果原生界面上每个RegionServer上每秒的请求数超过1W，需要增大handle数来预防重启过程中负载过大导致的RegionServer重启失败。
- 重启前需要观察当前hbase的负载请求数（原生界面上每个rs的请求数如果超过1W，需要增大handle数来预防到时候负载不过来）
- 在集群Core节点个数小于6个的情况下，可能会出现业务短时间受影响的情况。
- 请优先使用滚动重启操作来重启实例或服务，并勾选“仅重启配置过期的实例”。

滚动重启服务

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“组件管理”，选择需要滚动重启的服务，进入服务页面。

步骤3 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤4 弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启服务。

步骤5 滚动重启任务完成后，单击“完成”。

----结束

滚动重启实例

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“组件管理”，选择需要滚动重启的服务，进入服务页面。

步骤3 在“实例”页签，勾选要重启的实例，单击“更多”，选择“滚动重启实例”。

步骤4 弹出“滚动重启实例”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启实例。

步骤5 滚动重启任务完成后，单击“完成”。

----结束

滚动重启集群

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 在页面右上角选择“管理操作 > 滚动重启集群”。

步骤3 弹出“滚动重启集群”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启集群。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启参数说明

滚动重启参数说明如表5-54所示。

表 5-54 滚动重启参数说明

参数名称	描述
仅重启配置过期的实例	是否只重启集群内修改过配置的实例。
启用机架策略	是否启用机架并发滚动重启策略，只对满足机架策略滚动重启的角色（角色支持机架感知功能，且角色下的实例归属于2个或2个以上的机架）生效。 **说明** 该参数仅在MRS 3.x及之后版本滚动重启HDFS、Yarn时可设置。
数据节点滚动重启并发数	采用分批并发滚动重启策略的数据节点实例每一个批次重启的实例数，默认为1，取值范围为1～20。只对数据节点有效。
批次时间间隔	滚动重启实例批次之间的间隔时间，默认为0，取值范围为0～2147483647，单位为秒。 说明：设置批次时间间隔参数可以增加滚动重启期间大数据组件进程的稳定性。建议设置该参数为非默认值，例如10。
退服超时时间	角色实例在滚动重启过程中的退服等待时间。
批次容错阈值	滚动重启实例批次执行失败容错次数，默认为0，即表示任意一个批次的实例重启失败后，滚动重启任务终止。取值范围为0～2147483647。

典型场景操作步骤

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“组件管理”，选择HBase，进入HBase服务页面。

步骤3 单击“服务配置”页签，修改HBase某个参数并保存配置，根据界面提示进行保存。

说明

MRS 3.x之前版本存在“重新启动受影响的服务或实例”参数，请勿勾选，该处重启是普通重启方式，会并发重启所有服务或实例，引起业务断服。

步骤4 保存配置完成后，单击“完成”。

步骤5 选择“服务状态”页签。

步骤6 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤7 弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确定，开始滚动重启。

步骤8 滚动重启任务完成后，单击“完成”。

----结束

5.7 告警管理

5.7.1 查看告警列表

告警列表显示了MRS集群中的所有告警信息，MRS界面显示需要用户及时处理的“告警”和标志事情发生的“事件。

MRS管理控制台“告警管理”只能查询MRS中未清除告警的基本信息，查看详细信息或管理告警具体请参见**查看与手动清除告警**。

告警列表默认按时间顺序排列，时间最近的告警显示在最前端。

告警信息中的各字段说明如**表5-55**所示。

表 5-55 告警说明

参数	参数说明
告警ID	告警的ID。
告警名	告警的名称。

参数	参数说明
级别	告警级别。 MRS 3.x之前版本集群告警级别为： ● 致命 指集群服务不可用，节点故障、GaussDB主备数据不同步、LdapServer数据同步异常等影响集群正常运行的告警，需要根据告警及时检查集群情况并恢复。 ● 严重 指集群部分功能不可用的告警，包括进程故障、周期备份任务失败、关键文件权限异常等，需要根据告警及时检查报告告警的对象并恢复。 ● 一般 指不影响当前集群主要功能的告警，包括证书文件即将过期、审计日志转储失败、License文件即将过期等告警。 ● 提示 指级别最低的一种告警，起到信息展示或信息提示的作用，标识这件事情的发生，一般包括：停止服务、删除服务、停止实例、删除实例、删除节点、重启服务、重启实例、Manager主备倒换、缩容主机、实例恢复、实例故障、作业执行成功、作业执行失败等。 MRS 3.x及之后版本集群告警级别为： ● 紧急 指集群服务不可用，节点故障、GaussDB主备数据不同步、LdapServer数据同步异常等影响集群正常运行的告警，需要根据告警及时检查集群情况并恢复。 ● 重要 指集群部分功能不可用的告警，包括进程故障、周期备份任务失败、关键文件权限异常等，需要根据告警及时检查报告告警的对象并恢复。 ● 次要 指不影响当前集群主要功能的告警，包括证书文件即将过期、审计日志转储失败、License文件即将过期等告警。 ● 提示 指级别最低的一种告警，起到信息展示或信息提示的作用，标识这件事情的发生，一般包括：停止服务、删除服务、停止实例、删除实例、删除节点、重启服务、重启实例、Manager主备倒换、缩容主机、实例恢复、实例故障、作业执行成功、作业执行失败等。
生成时间	产生告警的时间。
定位信息	告警的详细信息。
操作	当告警可手动清除时，单击“清除告警”进行处理。

表 5-56 按钮说明

按钮	说明
	在下拉框中选择刷新告警列表的周期。 ● 每30s刷新一次 ● 每60s刷新一次 ● 停止刷新
	在下拉框中选择告警级别，筛选告警。 MRS 3.x之前版本集群可筛选告警包括：全部、致命、严重、一般、提示。 MRS 3.x及之后版本集群可筛选告警包括：全部、紧急、重要、次要、提示。
	单击 ，手动刷新告警列表。
高级搜索	单击“高级搜索”显示告警搜索区域，设置查询条件后，单击“搜索”，查看指定的告警信息。单击“重置”清除输入的搜索条件。

5.7.2 查看事件列表

事件列表显示了集群中的所有事件信息，如重启服务、停止服务等。

事件列表默认按时间顺序排列，时间最近的告警显示在最前端。

事件信息中的各字段说明如**表1 事件说明**所示。

表 5-57 事件说明

参数	参数说明
事件ID	事件的ID。
事件级别	事件级别。 MRS 3.x之前版本集群事件级别为： ● 致命 ● 严重 ● 一般 ● 提示 MRS 3.x及之后版本集群事件级别为： ● 紧急 ● 重要 ● 次要 ● 提示

参数	参数说明
事件名称	产生事件的名称。
生成时间	产生事件的时间。
定位信息	定位事件的详细信息。

表 5-58 按钮说明

按钮	说明
	在下拉框中选择刷新事件列表的周期。 ● 每30s刷新一次 ● 每60s刷新一次 ● 停止
	单击 ，手动刷新事件列表。
高级搜索	单击“高级搜索”显示事件搜索区域，设置查询条件后，单击“搜索”，查看指定的事件信息。单击“重置”清除输入的搜索条件。

导出事件

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“告警管理 > 事件”。

步骤3 单击“全部导出”。

步骤4 在弹框内选择保存类型，单击“确定”。

----结束

常见事件列表

表 5-59 常见事件列表

事件ID	事件名称
12019	停止服务
12020	删除服务
12021	停止实例
12022	删除实例
12023	删除节点

事件ID	事件名称
12024	重启服务
12025	重启实例
12026	Manager主备倒换
12065	进程重新启动
12070	作业执行成功
12071	作业执行失败
12072	作业被终止
12086	Agent进程重启
14005	NameNode主备倒换
14028	HDFS磁盘均衡任务
14029	主NameNode进入安全模式并生产新的Fsimage
17001	Oozie工作流执行失败
17002	Oozie定时任务执行失败
18001	ResourceManager主备倒换
18004	JobHistoryServer主备倒换
19001	HMaster主备倒换
20003	Hue发生主备切换
24002	Flume Channel溢出
25001	LdapServer主备倒换
27000	DBServer主备倒换
38003	Topic数据保存周期配置调整
43014	Spark2x数据倾斜
43015	Spark2x SQL超大查询结果
43016	Spark2x SQL执行超时
43024	启动JDBCServer
43025	停止JDBCServer
43026	ZooKeeper连接成功
43027	ZooKeeper连接异常

5.7.3 查看与手动清除告警

操作场景

用户可以在MRS上查看、清除告警。

一般情况下，告警处理后，系统自动清除该条告警记录。当告警不具备自动清除功能且用户已确认该告警对系统无影响时，可手动清除告警。

在MRS界面可查看最近十万条告警（包括未清除的、手动清除的和自动清除的告警）。如果已清除告警超过十万条达到十一万条，系统自动将最早的一万条已清除告警转存，转存路径为：

3.x以前版本，主管理节点的“${BIGDATA_HOME}/OMSV100R001C00x8664/workspace/data”。

3.x及后续版本，主管理节点的“${BIGDATA_HOME}/om-server/OMS/workspace/data”。

第一次转存告警时自动生成目录。

说明

用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止”：停止刷新。

操作步骤

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“告警管理”，在告警列表查看告警信息。

- 告警列表每页默认显示最近的十条告警。
- 默认以“生成时间”列按降序排列。针对MRS 3.x之前版本集群，单击“告警ID”、“级别”、“生成时间”可修改排列方式；针对MRS 3.x及以后版本集群，单击“级别”、“生成时间”可修改排列方式。
- 支持在告警“级别”筛选相同级别的全部告警。结果包含已清除和未清除的告警。
- 针对MRS 3.x之前版本集群分别单击页面右上角 、 、 或 可以快速筛选级别为“致命”、“严重”、“一般”或“提示”的未清除告警。
- 针对MRS 3.x及之后版本集群分别单击页面右上角 、 、 或 可以快速筛选级别为“紧急”、“重要”、“次要”或“提示”的未清除告警。

步骤3 单击“高级搜索”显示告警搜索区域，设置查询条件后，单击“搜索”，查看指定的告警信息。单击“重置”清除输入的搜索条件。

说明

“起止时间”表示时间范围的开始时间和结束时间，可以搜索此时间段内产生的告警。

查看“告警参考”章节告警帮助，按照帮助指导处理告警。如果某些场景中告警由于MRS依赖的其他云服务产生，可能需要联系对应云服务运维人员处理。

步骤4 处理完告警后，若需手动清除，单击“清除告警”，手动清除告警。

说明

如果有多个告警已完成处理，可选中一个或多个待清除的告警，单击“清除告警”，批量清除告警。每次最多批量清除300条告警。

----结束

导出告警

步骤1 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面。

步骤2 单击“告警管理 > 告警”。

步骤3 单击“全部导出”。

步骤4 在弹框内选择“保存类型”，单击“确定”。

----结束

5.8 补丁管理

5.8.1 MRS 1.7.0 前版本补丁操作指导

当您通过如下途径获知集群版本补丁信息，请根据您的实际需求进行补丁升级操作。

- 通过消息中心服务推送的消息获知MapReduce服务发布了补丁信息。
- 进入现有集群，查看“补丁信息”页面，呈现补丁信息。

安装补丁前准备

- 请参见**执行健康检查**检查集群状态，确认集群健康状态正常后再安装补丁。
- 已经将对应版本的集群补丁包上传至服务器，详见**上传补丁包**。
- 您根据“补丁内容”中的补丁信息描述，确认将要安装的目标补丁。

上传补丁包

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 单击“上传补丁”，分别设置如下参数

- 补丁文件路径:存储补丁包的OBS文件系统下创建的文件夹，例如MRS_1.6.2/MRS_1_6_2_11.tar.gz。
- 并行文件系统名:存储补丁包的OBS文件系统名称，例如mrs_patch。

说明

文件系统名和补丁文件路径可从管理面的“补丁信息”页签中获取，“补丁信息”页签中的“补丁路径”字段组成方式如下：【文件系统名】/【补丁文件路径】。

- AK：请参见“我的凭证 > 访问密钥”。
- SK：请参见“我的凭证 > 访问密钥”。

步骤4 单击“确定”，完成补丁的上传。

----结束

安装补丁

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“安装”。

步骤4 系统弹出提示框，单击“确定”，执行补丁的安装。

步骤5 补丁安装完成后，可在进度条中查看其安装状态，若安装失败，请联系MRS集群管理员处理。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“卸载”。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

5.8.2 MRS 1.7.0 至 2.0.1 版本补丁操作指导

当您通过如下途径获知集群版本补丁信息，请根据您的实际需求进行补丁升级操作。

- 通过消息中心服务推送的消息获知MapReduce服务发布了补丁信息。
- 进入现有集群，查看“补丁信息”页面，呈现补丁信息。

安装补丁前准备

- 请参见**执行健康检查**检查集群状态，确认集群健康状态正常后再安装补丁。
- 您根据“补丁内容”中的补丁信息描述，确认将要安装的目标补丁。

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“安装”，安装目标补丁。

说明

- 对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“卸载”，卸载目标补丁。

说明

- 对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

5.8.3 滚动补丁

滚动补丁是指在补丁安装/卸载时，采用滚动重启服务（按批次重启服务或实例）的方式，在不中断或尽可能短地中断集群各个服务业务的前提下完成对集群中单个或多个服务的补丁安装/卸载操作。集群中的服务根据对滚动补丁的支持程度，分为三种：

- 支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的全部业务或部分业务（因服务而异，不同服务存在差别）不中断。
- 不支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的业务会中断。
- 部分角色支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的部分业务不中断。

说明

MRS 3.x版本暂不支持在管理控制台执行本章节操作。

当前MRS集群中，服务和实例是否支持滚动重启如**表5-60**所示。

表 5-60 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	

服务	实例	是否支持滚动重启
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	
	Logviewer	

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“安装”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动安装补丁功能开启：补丁安装前不会停止服务，补丁安装后滚动重启服务来完成补丁安装，可以最大程度减少对集群业务的影响，但相比普通方式安装耗时更久。
- 滚动安装补丁功能关闭：补丁安装前会停止服务，补丁安装后再重新启动服务来完成补丁安装，会造成集群和服务暂时中断，但相比滚动方式安装补丁耗时更短。
- 少于2个Master节点和少于3个Core节点的集群不支持滚动方式安装补丁。

步骤5 单击“是”，安装目标补丁。

步骤6 查看补丁安装进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁安装进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁管理”页面，在操作列表中单击“卸载”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动卸载补丁功能开启：补丁卸载前不会停止服务，补丁卸载后滚动重启服务来完成补丁卸载，可以最大程度减少对集群业务的影响，但相比普通方式卸载耗时更久。
- 滚动卸载补丁功能关闭：补丁卸载前会停止所有服务，补丁卸载后再重新启动所有服务来完成补丁卸载，会造成集群和服务暂时中断，但相比滚动方式卸载补丁耗时更短。
- 仅通过滚动方式安装的补丁支持滚动方式卸载补丁。

步骤5 单击“是”，卸载目标补丁。

步骤6 查看补丁卸载进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁卸载进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

5.8.4 修复隔离主机补丁

若集群中存在主机被隔离的情况，集群补丁安装完成后，请参见本节操作对隔离主机进行补丁修复。修复完成后，被隔离的主机节点版本将与其他未被隔离的主机节点一致。

说明

MRS 3.x版本暂不支持在管理控制台执行本章节操作。

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“详情”。

步骤4 在补丁详情界面，选中“Status”是“Isolated”的主机节点。

步骤5 单击“Select and Restore”，修复被隔离的主机节点。

----结束

5.9 租户管理

5.9.1 使用前须知

本章节指导用户在MRS控制台执行租户管理操作。

在控制台界面执行租户管理操作仅适用于**MRS 3.x之前版本**集群。

在Manager界面执行租户管理操作适用于所有版本。

MRS 3.x之前版本请参考**租户简介**。

5.9.2 租户简介

定义

MRS集群拥有的不同资源和服务支持多个组织、部门或应用共享使用。集群提供了一个逻辑实体来统一使用不同资源和服务，这个逻辑实体就是租户。多个不同的租户统称多租户。当前仅分析集群支持租户。

原理

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

计算资源指租户Yarn任务队列资源，可以修改任务队列的配额，并查看任务队列的使用状态和使用统计。

存储资源目前支持HDFS存储，可以添加删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额。

租户可以在界面上根据业务需要，在集群中创建租户、管理租户。

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。默认情况下，新的计算资源和存储资源的全部权限将分配给租户的角色。
- 默认情况下，查看当前租户的资源，在当前租户中添加子租户并管理子租户资源的权限将分配给租户的角色。
- 修改租户的计算资源或存储资源，对应的角色关联权限将自动更新。

MRS中最多支持512个租户。系统默认创建的租户包含“default”。和默认租户同处于最上层的租户，可以统称为一级租户。

资源池

YARN任务队列支持一种调度策略，称为标签调度（Label Based Scheduling）。通过此策略，YARN任务队列可以关联带有特定节点标签（Node Label）的NodeManager，使YARN任务在指定的节点运行，实现任务的调度与使用特定硬件资源的需求。例如，需要使用大量内存的YARN任务，可以通过标签关联具有大量内存的节点上运行，避免性能不足影响业务。

在MRS集群中，租户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个资源池。YARN任务队列通过配置队列容量策略，与指定的资源池进行关联，可以更有效地使用资源池中的资源，且互不影响。

MRS中最多支持50个资源池。系统默认包含一个“default”资源池。

5.9.3 添加租户

操作场景

当租户需要根据业务需求指定资源使用情况时，可以在MRS创建租户。

前提条件

- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的完整存储路径在HDFS目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“添加租户”，打开添加租户的配置页面，参见以下表格内容为租户配置属性。

表 5-61 租户参数一览表

参数名	描述
名称	指定当前租户的名称，长度为3到50，可包含数字、字母和下划线。
租户类型	可选参数值为“叶子租户”和“非叶子租户”。当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。

参数名	描述
动态资源	为当前租户选择动态计算资源。系统将自动在Yarn中以租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量 (%)	配置当前租户在“default”资源池中使用的计算资源百分比。
默认资源池最大容量 (%)	配置当前租户在“default”资源池中使用的最大计算资源百分比。
储存资源	为当前租户选择存储资源。系统将自动在“/tenant”目录中以租户名称创建文件夹。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。
存储空间配额 (MB)	配置当前租户使用的HDFS存储空间配额。取值范围为“1”到“8796093022208”。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间表示所有副本文件在HDFS中占用的磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
存储路径	配置租户在HDFS中的存储目录。系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“tenant/ta1”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
服务	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
描述	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配Manager_tenant角色以及租户对应的角色。具体操作请参见**创建用户**。

----结束

相关任务

查看已添加的租户

步骤1 在集群详情页，单击“租户管理”。

步骤2 在左侧租户列表，单击已添加租户的名称。

默认在右侧显示“概述”页签。

步骤3 查看当前租户的“基本信息”、“资源配额”和“统计”。

如果HDFS处于“已停止”状态，“资源配额”中“Space”的“可用”和“已使用”会显示为“unknown”。

----结束

5.9.4 添加子租户

操作场景

当租户需要根据业务需求，将当前租户的资源进一步分配时，可以在MRS添加子租户。

前提条件

- 已添加“非叶子租户”。
- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果子租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的存储目录在父租户的存储目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 在左侧租户列表，将光标移动到需要添加子租户的租户节点上，单击“添加子租户”，打开添加子租户的配置页面，参见以下表格内容为租户配置属性。

表 5-62 子租户参数一览表

参数名	描述
父租户	显示上级父租户的名称。
名称	指定当前租户的名称，长度为3到20，可包含数字、字母和下划线。

参数名	描述
租户类型	可选参数值为“叶子租户”和“非叶子租户”，当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
动态资源	为当前租户选择动态计算资源。系统将自动在Yarn父租户队列中以子租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。如果父租户未选择动态资源，子租户也无法使用动态资源。
默认资源池容量 (%)	配置当前租户使用的资源百分比，基数为父租户的资源总量。
默认资源池最大容量 (%)	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。
储存资源	为当前租户选择存储资源。系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。如果父租户未选择存储资源，子租户也无法使用存储资源。
存储空间配额 (MB)	配置当前租户使用的HDFS存储空间配额。最小值值为“1”，最大值为父租户的全部存储配额。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。若此配额大于父租户的配额，实际存储量受父租户配额影响。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“父租户中分配资源”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
存储路径	配置租户在HDFS中的存储目录。系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“tenant/ta1”，系统默认自动配置此参数值为“tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。支持在父目录中自定义存储路径。存储路径的父目录必需是父租户的存储目录。
服务	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
描述	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成子租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配租户对应的角色。具体操作请参见**创建用户**。

----结束

5.9.5 删除租户

操作场景

当租户需要根据业务需求，将当前不再使用的租户删除时，可以在MRS完成操作。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。该任务对应取消角色与用户的绑定，请参见**修改用户信息**。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 在左侧租户列表，将光标移动到需要删除的租户节点上，单击“删除”。

界面显示删除租户对话框。根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户的数据”，否则将自动删除租户对应的存储空间。

步骤3 单击“是”，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

- 租户删除后，Yarn中对应的租户任务队列不会被删除。
- 删除父租户时选择不保留数据，如果存在子租户且子租户使用了存储资源，则子租户的数据也会被删除。

----结束

5.9.6 管理租户目录

操作场景

用户根据业务需求，可以在MRS对指定租户使用的HDFS存储目录，进行管理操作。支持用户对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

- 已添加关联了HDFS存储资源的租户。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

- 查看租户目录
  a. 在集群详情页，单击“租户管理”。
  b. 在左侧租户列表，单击目标的租户。
  c. 单击“资源”页签。
  d. 查看“HDFS 存储”表格。
     - 指定租户目录的“文件/目录数上限”列表示文件和目录数量配额。
     - 指定租户目录的“存储空间配额”列表示租户目录的存储空间大小。
- 添加租户目录
  a. 在集群详情页，单击“租户管理”。
  b. 在左侧租户列表，单击需要添加HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS 存储”表格，单击“添加目录”。
     - “路径”填写租户目录的路径。

       说明

       - 如果当前租户不是子租户，新路径将在HDFS的根目录下创建。
       - 如果当前租户是一个子租户，新路径将在指定的目录下创建。

       完整的HDFS存储目录最多包含1023个字符。HDFS目录名称包含数字、大小写字母、空格和下划线。空格只能在HDFS目录名称的中间使用。
     - “文件/目录数上限”填写文件和目录数量配额。
       “文件/目录数上限”为可选参数，取值范围从1到9223372036854775806。
     - “存储空间配额 ”填写租户目录的存储空间大小。
       “存储空间配额 ”的取值范围从1到8796093022208。

> **说明**
>
> 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

e. 单击“确定”完成租户目录添加，系统将在HDFS根目录下创建租户的目录。

- 修改租户目录
  a. 在集群详情页，单击“租户管理”。
  b. 在左侧租户列表，单击需要修改HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。
     - “文件/目录数上限”填写文件和目录数量配额。
       “文件/目录数上限”为可选参数，取值范围从1到9223372036854775806。
     - “存储空间配额”填写租户目录的存储空间大小。
       “存储空间配额”的取值范围从1到8796093022208。

       > **说明**
       >
       > 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

  e. 单击“确定”完成租户目录修改。
- 删除租户目录
  a. 在集群详情页，单击“租户管理”。
  b. 在左侧租户列表，单击需要删除HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS 存储”表格，指定租户目录的“操作”列，单击“删除”。
     创建租户时设置的默认HDFS存储目录不支持删除，仅支持删除新添加的HDFS存储目录。
  e. 单击“确认”完成租户目录删除。

5.9.7 恢复租户数据

操作场景

租户的数据默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，需要手动恢复。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“基本信息”左侧的圆圈，绿色表示租户可用，灰色表示租户不可用。
2. 单击“资源”，查看“Yarn”或者“HDFS 存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击“恢复租户数据”。

步骤5 在“恢复租户数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

5.9.8 添加资源池

操作场景

在MRS集群中，用户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个YARN资源池。每个NodeManager只能属于一个资源池。系统中默认包含了一个名为“default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

该任务指导用户通过MRS添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 单击“添加资源池”。

步骤4 在“添加资源池”设置资源池的属性。

- “名称”：填写资源池的名称。不支持创建名称为“default”的资源池。

  资源池的名称，长度为1到20个字节，可包含数字、字母和下划线，且不能以下划线开头。
- “可用主机”：在界面左边主机列表，勾选指定的主机名称加入资源池。只支持选择本集群中的主机。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

步骤6 完成资源池创建后，用户可以在资源池的列表中查看资源池的“名称”、“成员”、“类型”、“虚拟核数”与“内存”。已加入自定义资源池的主机，不再是“default”资源池的成员。

----结束

5.9.9 修改资源池

操作场景

该任务指导用户通过MRS修改已有资源池中的成员。

前提条件

已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“修改”。

步骤4 在“修改资源池”修改“已添加主机”。

- 增加主机：在界面左边主机列表，勾选指定的主机名称加入资源池。
- 删除主机：在界面右边主机列表，单击指定主机后的×将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

5.9.10 删除资源池

操作场景

该任务指导用户通过MRS删除已有资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列配置**。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

在弹出窗口中单击“确定”。

----结束

5.9.11 配置队列

操作场景

用户根据业务需求，可以在MRS修改指定租户的队列配置。

前提条件

- 已添加关联Yarn并分配了动态资源的租户。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“队列配置”页签。

步骤3 在租户队列表格，指定租户队列的“操作”列，单击“修改”。

说明

- 在“租户管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击 也能打开修改队列配置页面。
- 一个队列只能绑定一个非default资源池。

MRS 3.x之前版本：

表 5-63 队列配置参数

参数名	描述
最大应用数量	表示最大应用程序数量。取值范围从“1”到“2147483647”。
AM最大资源百分比	表示集群中可用于运行application master的最大资源占比。取值范围从“0”到“1”。
用户资源最小上限百分比 (%)	表示用户使用的最小资源上限百分比。取值范围从“0”到“100”。
用户资源上限因子	表示用户使用的最大资源限制因子，与当前租户在集群中实际资源百分比相乘，可计算出用户使用的最大资源百分比。最小值为“0”。

参数名	描述
状态	表示资源计划当前的状态，“运行”为运行状态，“停止”为停止状态。
默认资源池	表示队列使用的资源池。默认为“default”，如果需要修改为其他资源，需要先配置队列容量，请参见**配置资源池的队列容量策略**。

MRS 3.x及之后版本：

表 5-64 队列配置参数

参数名	描述
AM最多占有资源（%）	表示当前队列内所有Application Master所占的最大资源百分比。
每个YARN容器最多分配核数	表示当前队列内单个YARN容器可分配的最多核数，默认为-1，表示取值范围内不限制。
每个YARN容器最大分配内存（MB）	表示当前队列内单个YARN容器可分配的最大内存，默认为-1，表示取值范围内不限制。
最多运行任务数	表示当前队列最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
每个用户最多运行任务数	表示每个用户在当前队列中最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
最多挂起任务数	表示当前队列最多同时可挂起任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可挂起任务。取值范围为-1～2147483647。
资源分配规则	表示单个用户任务间的资源分配规则，包括FIFO和FAIR。 一个用户若在当前队列上提交了多个任务，FIFO规则代表一个任务完成后再执行其他任务，按顺序执行。FAIR规则代表各个任务同时获取到资源并平均分配资源。
默认资源标签	表示在指定资源标签（Label）的节点上执行任务。 **说明** 如果需要使用新的资源池，需要修改默认标签为新的资源池标签。
Active状态	• ACTIVE表示当前队列可接受并执行任务。 • INACTIVE表示当前队列可接受但不执行任务，若提交任务，任务将处于挂起状态。
Open状态	• OPEN表示当前队列处于打开状态。 • CLOSED表示当前队列处于关闭状态，若提交任务，任务直接会被拒绝。

----结束

5.9.12 配置资源池的队列容量策略

操作场景

添加资源池后，需要为YARN任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。每个队列只能配置一个资源池的队列容量策略。用户可以在任何一个资源池中查看队列并配置队列容量策略。配置队列策略后，YARN任务队列与资源池形成关联关系。

该任务指导用户通过MRS配置队列策略。

前提条件

- 已添加资源池。
- 任务队列与其他资源池无关联关系。默认情况下，所有队列与“default”资源池存在关联关系。
- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“资源分布策略”页签。

步骤3 在“资源池”选择指定的资源池。

“可用资源配额”：表示每个资源池默认所有资源都可分配给队列。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“修改”。

步骤5 在“修改资源分配”窗口设置任务队列在此资源池中的资源容量策略。

- “资源容量 (%)”：表示当前租户计算资源使用的资源百分比。
- “最大资源容量 (%)”：表示当前租户计算资源使用的最大资源百分比。

步骤6 单击“确定”保存配置。

----结束

5.9.13 清除队列配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，用户可以在MRS清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

- 如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

- 已完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

操作步骤

步骤1 在集群详情页，单击“租户管理”。

步骤2 单击“资源分布策略”页签。

步骤3 在“资源池”选择指定的资源池。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“清除”。

在“清除队列设置”中单击“是”，清除队列在当前资源池的配置。

说明

如果用户未配置队列的资源容量策略，则清除功能默认不可用。

----结束

5.10 引导操作

5.10.1 引导操作简介

引导操作是指启动集群组件前（或后）在指定的节点上执行脚本。您可以通过引导操作来完成安装其他第三方软件，修改集群运行环境等自定义操作。

如果集群扩容，选择执行引导操作，则引导操作也会以相同方式在新增节点上执行。如果集群开启弹性伸缩功能，可以在配置资源计划的同时添加自动化脚本，则自动化脚本会在弹性伸缩的节点上执行，实现用户自定义操作。

MRS 3.x之前版本会使用root用户执行您指定的脚本，脚本内部您可以通过**su - XXX**命令切换用户。

MRS 3.x及之后版本会使用omm用户执行您指定的脚本，脚本内部您可以通过**su - XXX**命令切换用户。

说明

MRS 3.x之前版本：引导操作脚本以root身份执行，使用不当可能会对集群可用性造成影响，请谨慎操作。

MRS 3.x及之后版本：引导操作脚本以omm身份执行，使用不当可能会对集群可用性造成影响，请谨慎操作。

MRS通过引导操作脚本返回码来判断结果，如果返回零，则代表脚本执行成功，非零代表执行失败。一个节点上执行某个引导脚本失败，则会导致相应引导脚本失败，您可以通过“失败后操作”来选择是否继续执行后续脚本。举例1：创建集群指定所有脚本的“失败后操作”都选择“继续”，则不管这些脚本实际执行成功或失败，都会全部执行，并完成启动流程。举例2：如果一个脚本执行失败，且“失败后操作”选择“终止”，则不会执行后续脚本，集群创建或扩容也随之失败。

您最多可以添加18个引导操作，它们会按照您指定的顺序在集群组件启动前（或后）执行。组件启动前（或后）执行的引导操作，必须在60分钟内完成，否则会引起集群创建或扩容失败。

5.10.2 准备引导操作脚本

引导操作目前仅支持linux shell脚本，脚本文件需以.sh结尾。

上传所需安装包等文件至 OBS 文件系统

正式编写脚本前，您需要将所需安装包、配置包的所有相关文件都上传到同region的OBS文件系统中。因为不同region间有网络隔离，MRS虚拟机无法下载其他region上的OBS文件。

脚本中如何从 OBS 文件系统下载文件

您可以在脚本中指定从OBS下载需要的文件。如果将文件上传到私有文件系统，需要用**hadoop fs**下载，下面的例子会将 obs://yourbucket/myfile.tar.gz 这个文件下载到本地，并解压到 /your-dir 目录下：

```
#!/bin/bash
source /opt/Bigdata/client/bigdata_env;hadoop fs -D fs.obs.endpoint=<obs-endpoint> -D
fs.obs.access.key=<your-ak> -D fs.obs.secret.key=<your-sk> -copyToLocal obs://yourbucket/
myfile.tar.gz ./
mkdir -p /<your-dir>
tar -zxvf myfile.tar.gz -C /<your-dir>
```

说明

- MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。
- Hadoop客户端已预安装在MRS节点上，**hadoop fs**命令可对OBS做下载、上传等操作。
- 各region下obs-endpoint请参见**终端节点及区域说明**。

上传脚本至 OBS 文件系统

脚本完成后上传到同region的OBS文件系统中。在您选定的时机，集群各节点会从OBS将脚本下载下来并以root用户执行。

5.10.3 查看执行记录

您可以在集群详情页选择“引导操作” 页签查看引导操作的执行结果。

查看执行结果

1. 登录MRS管理控制台。
2. 在“集群列表 > 现有集群”中单击需要查询的集群名称。

   系统跳转至该集群详情页面。
3. 在集群详情页面选择“引导操作”页签。系统显示创建集群时所添加的引导操作信息。

   说明

   - 可以通过选择右上角的“组件首次启动前”或者“组件首次启动后”查询相关的引导操作信息。
   - 这里列出的是上次执行结果。对于新创建的集群，则列出的是创建时执行引导操作的记录；如果集群被扩容了，则列出的是上次扩容对新增节点执行引导操作的记录。

查看执行日志

如果需要查看引导操作的执行日志，请在添加引导操作时将“失败操作”配置为“继续”，然后登录到各个节点上查看运行日志，运行日志在/var/log/Bootstrap目录下。如果您对组件启动前后都添加了引导操作，可通过时间戳前后关系来区分两个阶段引导操作的日志。

建议您在脚本中尽量详细地打印日志，以方便查看运行结果。MRS将脚本的标准输出和错误输出都重定向到了引导操作日志目录下。

5.10.4 添加引导操作

本章节介绍如何添加引导操作。

该操作适用于MRS 3.x之前版本集群。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 在集群基本信息页，单击“引导操作”。

步骤4 单击“添加”，根据界面信息填写对应参数。

表 5-65 参数描述

参数	说明
名称	引导操作脚本的名称。 只能由数字、英文字符、空格、中划线和下划线组成，且不能以空格开头。 可输入的字符串长度为1～64个字符。 **说明** 同一集群内，不允许配置相同的名称。不同集群之间，可以配置相同的名称。
脚本路径	脚本的路径。路径可以是OBS文件系统的路径或虚拟机本地的路径。 • OBS文件系统的路径，必须以s3a://开头，以.sh结尾。例如：s3a://mrs-samples/xxx.sh • 虚拟机本地的路径，脚本所在的路径必须以‘/’开头，以.sh结尾。 **说明** 同一集群内，不允许配置相同的路径。不同集群之间，可以配置相同的路径。
参数	引导操作脚本参数。
执行节点	选择引导操作脚本所执行的节点类型。

参数	说明
执行时机	选择引导操作脚本执行的时间。 ● 组件首次启动前 ● 组件首次启动后
失败操作	该脚本执行失败后，是否继续执行后续脚本和创建集群。 **说明** 建议您设置为“继续”，无论此引导操作是否执行成功，则集群都能继续创建。

步骤5 单击“确定”保存添加的引导操作。

步骤6 单击“是”即可添加成功。

----结束

5.10.5 修改引导操作

操作场景

该任务指导用户通过MRS修改已有引导操作。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 在集群基本信息页，单击“引导操作”。

步骤4 在列表中勾选需要修改的列表项，单击“编辑”。

步骤5 根据界面信息修改对应参数。

步骤6 单击“确定”保存修改的参数。

步骤7 单击“是”，即可修改成功。

----结束

5.10.6 删除引导操作

操作场景

该任务指导用户通过MRS删除已有引导操作。

操作步骤

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 在集群基本信息页，单击“引导操作”。

步骤4 在列表中勾选需要修改的列表项，单击“删除”。

步骤5 单击“确定”。

----结束

6 使用 MRS 客户端

6.1 安装客户端

6.1.1 安装客户端（3.x 及之后版本）

操作场景

该操作指导安装工程师安装MRS集群所有服务（不包含Flume）的客户端。Flume客户端安装请参见“组件操作指南 > 使用Flume > 安装Flume客户端”。

客户端可以安装集群内节点，也可以安装在集群外节点，本章节以安装目录“/opt/client”为例进行介绍，请以实际集群版本为准。

在集群外节点安装客户端前提条件

- 已准备一个Linux弹性云服务器，主机操作系统及版本建议参见表6-1。

  表 6-1 参考列表

CPU架构	操作系统	支持的版本号
x86计算	Euler	EulerOS 2.5
	SUSE	SUSE Linux Enterprise Server 12 SP4（SUSE 12.4）
	Red Hat	Red Hat-7.5-x86_64（Red Hat 7.5）
	CentOS	CentOS-7.6版本（CentOS 7.6）
鲲鹏计算(ARM)	Euler	EulerOS 2.8
	CentOS	CentOS-7.6版本（CentOS 7.6）

  同时为弹性云服务分配足够的磁盘空间，例如“40GB”。

- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。
- 弹性云服务器操作系统已安装NTP服务，且NTP服务运行正常。

  若未安装，在配置了**yum**源的情况下，可执行**yum install ntp -y**命令自行安装。
- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式）。

集群内节点安装客户端

1. 获取软件包。

   访问FusionInsight Manager（MRS 3.x及之后版本），在“集群”下拉列表中单击需要操作的集群名称。

   选择“更多 > 下载客户端”，弹出“下载集群客户端”信息提示框。

   说明

   在只安装单个服务的客户端的场景中，选择“集群 > 服务 > *服务名称* > 更多 > 下载客户端”，弹出“下载客户端”信息提示框。
2. “选择客户端类型”中选择“完整客户端”。

   “仅配置文件”下载的客户端配置文件，适用于应用开发任务中，完整客户端已下载并安装后，通过Manager界面修改了服务端配置，开发人员需要更新客户端配置文件的场景。

   平台类型包括x86_64和aarch64两种：

   - x86_64：可以部署在X86平台的客户端软件包。
   - aarch64：可以部署在TaiShan服务器的客户端软件包。

   说明

   集群支持下载x86_64和aarch64两种类型客户端，但是客户端类型必须与待安装节点的架构匹配，否则客户端会安装失败。
3. 勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件。

   文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。单击“确定”，等待下载完成后，使用**omm**用户或**root**用户将获取的软件包复制到将要安装客户端的服务器文件目录。

   客户端软件包名称格式为：“FusionInsight_Cluster_*<集群ID>*_Services_Client.tar”。本章节仅以*集群ID*为**1**进行介绍，请以实际*集群ID*为准。

   后续步骤及章节以FusionInsight_Cluster_1_Services_Client.tar进行举例。

   说明

   当用户无法获取**root**用户权限，需要用**omm**用户操作。

   如需安装客户端至集群内其他节点，则执行以下命令复制客户端到待安装客户端的节点：

   scp -p /*tmp/FusionInsight-Client*/**FusionInsight_Cluster_1_Services_Client.tar** *待安装客户端节点的IP地址:/opt/Bigdata/client*
4. 以**user_client**用户登录将要安装客户端的服务器。
5. 解压软件包。

   进入安装包所在目录，例如“/tmp/FusionInsight-Client”。执行如下命令解压安装包到本地目录。

   tar -xvf FusionInsight_Cluster_1_Services_Client.tar

6. 校验软件包。

   执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

   sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256

   ```
   FusionInsight_Cluster_1_Services_ClientConfig.tar: OK
   ```

7. 解压获取的安装文件。

   tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

8. 进入安装包所在目录，执行如下命令安装客户端到指定目录（绝对路径），例如安装到“/opt/client”目录。

   cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig

   执行**./install.sh /opt/client**命令，等待客户端安装完成（以下只显示部分屏显结果）。

   ```
   The component client is installed successfully
   ```

   说明

   - 如果已经安装的全部服务或某个服务的客户端使用了“/opt/client”目录，再安装其他服务的客户端时，需要使用不同的目录。
   - 卸载客户端请删除客户端安装目录。
   - 如果要求安装后的客户端仅能被该安装用户（如“user_client”）使用，请在安装时加“-o”参数，即执行**./install.sh /opt/client -o**命令安装客户端。
   - 由于HBase使用的Ruby语法限制，如果安装的客户端中包含了HBase客户端，建议客户端安装目录路径只包含大写字母、小写字母、数字以及_-?.@+=字符。

使用客户端

1. 在已安装客户端的节点，执行**sudo su - omm**命令切换用户。执行以下命令切换到客户端目录：

   cd /opt/client

2. 执行以下命令配置环境变量：

   source bigdata_env

3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit admin**。

   说明

   启用Kerberos认证的MRS集群默认创建“admin”用户账号，用于集群管理员维护集群。

4. 直接执行组件的客户端命令。

   例如：使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

集群外节点安装客户端

1. 创建一个满足要求的弹性云服务器，要求如下：
   - 已准备一个Linux弹性云服务器，主机操作系统及版本建议参见**表6-2**。

表 6-2 参考列表

CPU架构	操作系统	支持的版本号
x86计算	Euler	EulerOS 2.5
	SUSE	SUSE Linux Enterprise Server 12 SP4（SUSE 12.4）
	Red Hat	Red Hat-7.5-x86_64（Red Hat 7.5）
	CentOS	CentOS-7.6版本（CentOS 7.6）
鲲鹏计算(ARM)	Euler	EulerOS 2.8
	CentOS	CentOS-7.6版本（CentOS 7.6）

同时为弹性云服务分配足够的磁盘空间，例如“40GB”。

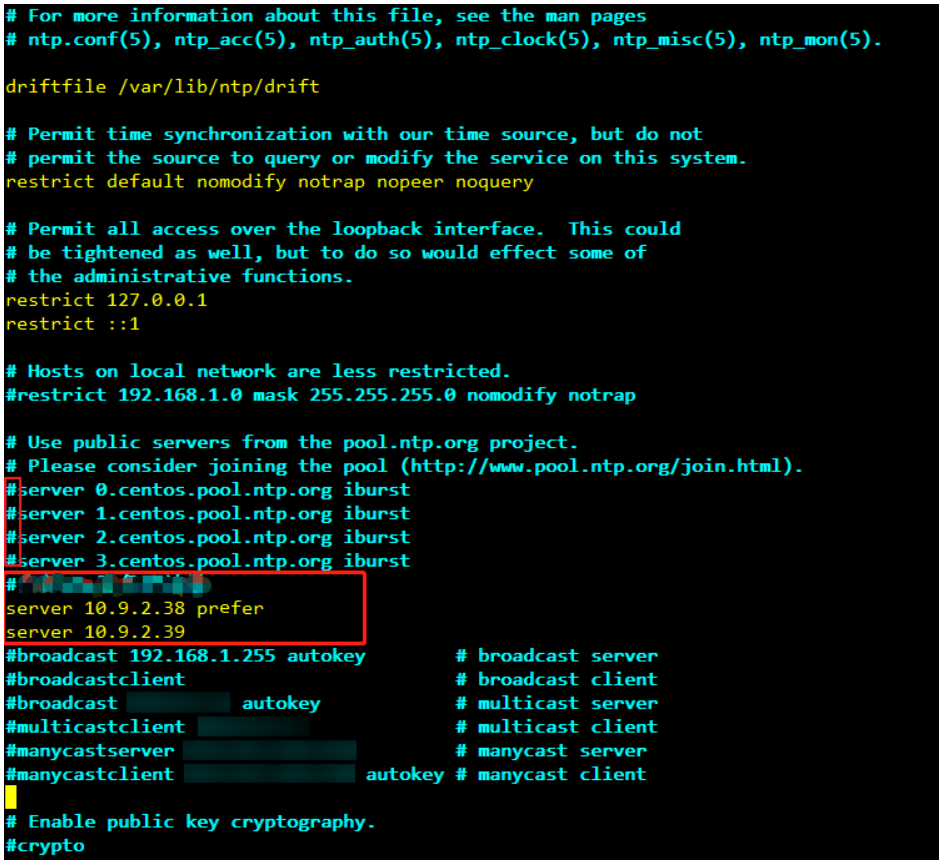
- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。
- 弹性云服务器操作系统已安装NTP服务，且NTP服务运行正常。

  若未安装，在配置了**yum**源的情况下，可执行**yum install ntp -y**命令自行安装。
- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式）。
- MRS集群安全组入方向将所有端口对客户端节点放开，具体操作请参考**添加安全组规则**。

2. 执行ntp时间同步，使集群外节点的时间与MRS集群时间同步。
   a. 执行**vi /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

```
server master1_ip prefer
server master2_ip
```

图 6-1 增加 Master 节点的 IP



b. 执行**service ntpd stop**命令关闭NTP服务。

c. 执行如下命令，手动同步一次时间：

/usr/sbin/ntpdate *192.168.10.8*

说明

192.168.10.8为主Master节点的IP地址。

d. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。

e. 执行**ntpstat**命令查看时间同步结果。

3. 参考以下步骤，从FusionInsight Manager下载集群客户端软件包并复制到ECS节点后安装客户端。

a. **访问FusionInsight Manager（MRS 3.x及之后版本）**，参考**集群内节点安装客户端**下载集群客户端到主管理节点的指定目录。

b. 使用**root**用户登录主管理节点，执行以下命令复制客户端安装包到待安装客户端的节点：

scp -p /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar *待安装客户端节点的IP地址*:**/tmp**

c. 使用待安装客户端的用户登录待安装客户端节点。

执行以下命令安装客户端，如果当前用户无客户端软件包以及客户端安装目录的操作权限，需使用**root**用户进行赋权：

cd /tmp

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

cd FusionInsight_Cluster_1_Services_ClientConfig

./install.sh /opt/client

d. 执行以下命令，切换到客户端目录并配置环境变量：

cd /opt/client

source bigdata_env

e. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如，**kinit admin**。

f. 直接执行组件的客户端命令。

例如使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

6.1.2 安装客户端（3.x 之前版本）

操作场景

用户需要使用MRS客户端。MRS集群客户端可以安装在集群内的Master节点或者Core节点，也可以安装在集群外节点上。

MRS 3.x之前版本集群在集群创建后，在主Master节点默认安装有客户端，可以直接使用，安装目录为“/opt/client”。

MRS 3.x及之后版本客户端的安装请参考**安装客户端（3.x及之后版本）**。

说明

如果集群外的节点已安装客户端且只需要更新客户端，请使用安装客户端的用户例如“root”。

在集群外节点安装客户端前提条件

- 已准备一个弹性云服务器，主机操作系统及版本请参见**表6-3**。

  表 6-3 参考列表

操作系统	支持的版本号
Euler	• 可用：Euler OS 2.2 • 可用：Euler OS 2.3 • 可用：Euler OS 2.5

  例如，用户可以选择企业镜像“Enterprise_SLES11_SP4_latest(4GB)”或标准镜像“Standard_CentOS_7.2_latest(4GB)”为弹性云服务器准备操作系统。

  同时为弹性云服务分配足够的磁盘空间，例如“40GB”。

- 弹性云服务器的VPC需要与MRS集群在同一个VPC中。
- 弹性云服务器的安全组需要和MRS集群Master节点的安全组相同。

如果不同，请修改弹性云服务器安全组或配置弹性云服务器安全组的出入规则允许MRS集群所有安全组的访问。

- 需要允许用户使用密码方式登录Linux弹性云服务器（SSH方式），请参见弹性云服务器《用户指南》中“实例> 登录Linux弹性云服务器 >SSH密码方式登录”。

在 Core 节点安装客户端

1. 登录MRS Manager页面，选择“服务管理 > 下载客户端”下载客户端安装包至主管理节点。

   说明

   如仅需更新客户端配置文件，请参考**更新客户端（3.x之前版本）**页面的方法二操作。
2. 使用IP地址搜索主管理节点并使用VNC登录主管理节点。
3. 在主管理节点，执行以下命令切换用户。

   sudo su - omm
4. 在MRS管理控制台，查看指定集群“节点管理”页面的“IP”地址。

   记录需使用客户端的Core节点IP地址。
5. 在主管理节点，执行以下命令，将客户端安装包从主管理节点文件拷贝到当前Core节点：

   scp -p /tmp/MRS-client/MRS_Services_Client.tar *Core节点的IP地址*:**/opt/client**
6. 使用“root”登录Core节点。
7. 执行以下命令，安装客户端：

   cd /opt/client

   tar -xvf MRS_Services_Client.tar

   tar -xvf MRS_Services_ClientConfig.tar

   cd /opt/client/MRS_Services_ClientConfig

   ./install.sh *客户端安装目录*

   例如，执行命令：

   ./install.sh /opt/client
8. 客户端的使用请参见**使用MRS客户端**。

使用 MRS 客户端

1. 在已安装客户端的节点，执行**sudo su - omm**命令切换用户。执行以下命令切换到客户端目录：

   cd /opt/client
2. 执行以下命令配置环境变量：

   source bigdata_env
3. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

   kinit *MRS集群用户*

   例如，**kinit admin**。

   说明

   启用Kerberos认证的MRS集群默认创建“admin”用户账号，用于集群管理员维护集群。

4. 直接执行组件的客户端命令。
   例如：使用HDFS客户端命令查看HDFS根目录文件，执行**hdfs dfs -ls /**。

在集群外节点上安装客户端

步骤1 根据前提条件，创建一个满足要求的弹性云服务器。

步骤2 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**，然后选择“服务管理”。

步骤3 单击“下载客户端”。

步骤4 在“客户端类型”选择“完整客户端”。

步骤5 在“下载路径”选择“远端主机”。

步骤6 将“主机IP”设置为ECS的IP地址，设置“主机端口”为“22”，并将“存放路径”设置为“/tmp”。

- 如果使用SSH登录ECS的默认端口“22”被修改，请将“主机端口”设置为新端口。
- “存放路径”最多可以包含256个字符。

步骤7 “登录用户”设置为“root”。

如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。

步骤8 在“SSH私钥”选择并上传创建B集群时使用的密钥文件。

步骤9 单击“确定”开始生成客户端文件。

若界面显示以下提示信息表示客户端包已经成功保存。单击“关闭”。客户端文件请到下载客户端时设置的远端主机的“存放路径”中获取。

```
下载客户端文件到远端主机成功。
```

若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从**步骤2**执行重新开始下载客户端。

```
连接到服务器失败，请检查网络连接或参数设置。
```

说明

生成客户端会占用大量的磁盘IO，不建议在集群处于安装中、启动中、打补丁中等非稳态场景下载客户端。

步骤10 使用VNC方式，登录弹性云服务器。参见弹性云服务器《用户指南》的**远程登录（VNC方式）**章节（“实例 > 登录Linux弹性云服务器 > 远程登录（VNC方式）”）。

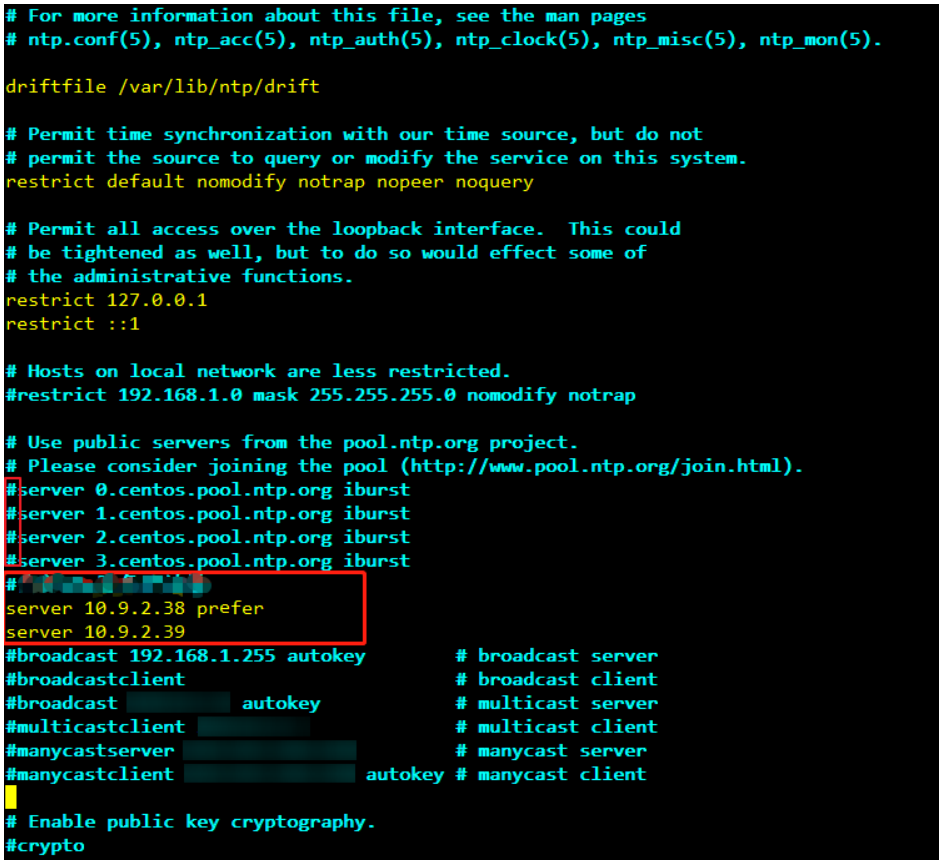
请先参考**SSH密钥方式登录**登录弹性云服务器，设置云服务器密码，然后再采用VNC方式登录。

步骤11 执行ntp时间同步，使集群外节点的时间与MRS集群时间同步。

1. 检查安装NTP服务有没有安装，未安装请执行**yum install ntp -y**命令自行安装。
2. 执行**vim /etc/ntp.conf**命令编辑NTP客户端配置文件，并增加MRS集群中Master节点的IP并注释掉其他**server**的地址。

```
server master1_ip prefer
server master2_ip
```

图 6-2 增加 Master 节点的 IP



3. 执行**service ntpd stop**命令关闭NTP服务。
4. 执行如下命令，手动同步一次时间：

   /usr/sbin/ntpdate *192.168.10.8*

   说明

   192.168.10.8为主Master节点的IP地址。
5. 执行**service ntpd start**或**systemctl restart ntpd**命令启动NTP服务。
6. 执行**ntpstat**命令查看时间同步结果。

步骤12 在弹性云服务器，切换到**root**用户，并将**步骤6**中“存放路径”中的安装包复制到目录“/opt”，例如“存放路径”设置为“/tmp”时命令如下。

sudo su - root

cp /tmp/MRS_Services_Client.tar /opt

步骤13 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Services_Client.tar

步骤14 执行以下命令，校验文件包。

sha256sum -c MRS_Services_ClientConfig.tar.sha256

界面显示如下：

```
MRS_Services_ClientConfig.tar: OK
```

步骤15 执行以下命令，解压“MRS_Services_ClientConfig.tar”。

tar -xvf MRS_Services_ClientConfig.tar

步骤16 执行以下命令，安装客户端到新的目录，例如“/opt/Bigdata/client”。安装时自动生成目录。

sh /opt/MRS_Services_ClientConfig/install.sh /opt/Bigdata/client

查看安装输出信息，如有以下结果表示客户端安装成功：

```
Components client installation is complete.
```

步骤17 验证弹性云服务器节点是否与集群Master节点的IP是否连通？

例如，执行以下命令：**ping** *Master节点IP地址*

- 是，执行**步骤18**。
- 否，检查VPC、安全组是否正确，是否与MRS集群在相同VPC和安全组，然后执行**步骤18**。

步骤18 执行以下命令配置环境变量：

source /opt/Bigdata/client/bigdata_env

步骤19 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit*MRS集群用户*

例如, **kinit admin**

步骤20 执行组件的客户端命令。

例如，执行以下命令查看HDFS目录：

hdfs dfs -ls /

----结束

6.2 更新客户端

6.2.1 更新客户端（3.x 及之后版本）

集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户如果在Manager修改了服务配置参数并重启了服务，已安装的客户端需要重新下载并安装，或者使用配置文件更新客户端。

更新客户端配置

方法一：

步骤1 **访问FusionInsight Manager（MRS 3.x及之后版本）**，在“集群”下拉列表中单击需要操作的集群名称。

步骤2 选择“更多 > 下载客户端 > 仅配置文件”。

此时生成的压缩文件包含所有服务的配置文件。

步骤3 是否在集群的节点中生成配置文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。然后执行**步骤4**。
- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，然后执行**步骤4**。

步骤4 使用WinSCP工具，以客户端安装用户将压缩文件保存到客户端安装的目录，例如“/opt/hadoopclient”。

步骤5 解压软件包。

例如下载的客户端文件为“FusionInsight_Cluster_1_Services_Client.tar”执行如下命令进入客户端所在目录，解压文件到本地目录。

cd /opt/hadoopclient

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤6 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如:

sha256sum -c
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar: OK
```

步骤7 解压获取配置文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar

步骤8 在客户端安装目录下执行如下命令，使用配置文件更新客户端。

sh refreshConfig.sh *客户端安装目录 配置文件所在目录*

例如，执行以下命令:

sh refreshConfig.sh /opt/hadoopclient /opt/hadoopclient/
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles

界面显示以下信息表示配置刷新更新成功:

```
Succeed to refresh components client config.
```

----结束

方法二:

步骤1 以**root**用户登录客户端安装节点。

步骤2 进入客户端安装的目录，例如“/opt/hadoopclient”，执行以下命令更新配置文件:

cd /opt/hadoopclient

sh autoRefreshConfig.sh

步骤3 按照提示输入FusionInsight Manager管理员用户名，密码以及FusionInsight Manager界面浮动IP。

步骤4 输入需要更新配置的组件名，组件名之间使用“,”分隔。如需更新所有组件配置，可直接单击回车键。

界面显示以下信息表示配置刷新更新成功：

```
Succeed to refresh components client config.
```

----结束

6.2.2 更新客户端（3.x 之前版本）

说明

本章节适用于MRS 3.x之前版本的集群。MRS 3.x及之后版本，请参考**更新客户端（3.x及之后版本）**。

更新客户端配置文件

操作场景

MRS集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户使用MRS的客户端时，如果在MRS Manager修改了服务配置参数并重启了服务或者重启了服务，需要先下载并更新客户端配置文件。

用户创建集群时，默认在集群所有节点的“/opt/client”目录安装保存了原始客户端。集群创建完成后，仅Master节点的客户端可以直接使用，Core节点客户端在使用前需要更新客户端配置文件。

操作步骤

方法一：

步骤1 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**，然后选择“服务管理”。

步骤2 单击“下载客户端”。

“客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤3 查询并登录主Master节点。

步骤4 若在集群内使用客户端，执行以下命令切换到omm用户，若在集群外使用客户端，请切换到root用户：

sudo su - omm

步骤5 执行以下命令切换客户端目录，例如“**/opt/Bigdata/client**”：

cd /opt/Bigdata/client

步骤6 执行以下命令，更新客户端配置：

sh refreshConfig.sh *客户端安装目录客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/Bigdata/client /tmp/MRS-client/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 集群安装完成之后，执行以下命令切换到omm用户，若在集群外使用客户端，请切换到root用户。

sudo su - omm

步骤2 执行以下命令切换客户端目录，例如“**/opt/Bigdata/client**”。

cd /opt/Bigdata/client

步骤3 执行以下命令并按照提示输入MRS Manager有下载权限的用户名和密码（例如，用户名为admin，密码为创建集群时设置的密码），更新客户端配置。

sh autoRefreshConfig.sh

步骤4 命令执行后显示如下信息，其中XXX表示集群安装的组件名称，如需更新全部组件配置，单击“Enter”键，如需更新部分组件配置，请输入需要更新的组件名称，多个组件名称以逗号相隔。

```
Components "xxx" have been installed in the cluster. Please input the comma-separated names of the components for which you want to update client configurations. If you press Enter without inputting any component name, the client configurations of all components will be updated:
```

界面显示以下信息表示配置更新成功：

```
Succeed to refresh components client config.
```

界面显示以下信息表示用户名或者密码错误：

```
login manager failed,Incorrect username or password.
```

说明

- 该脚本会自动连接到集群并调用refreshConfig.sh脚本下载并刷新客户端配置文件。
- 客户端默认使用安装目录下文件Version中的“wsom=xxx”所配置的浮动IP刷新客户端配置，如需刷新为其他集群的配置文件，请执行本步骤前修改Version文件中“wsom=xxx”的值为对应集群的浮动IP地址。

----结束

全量更新主 Master 节点的原始客户端

场景描述

用户创建集群时，默认在集群所有节点的“/opt/client”目录安装保存了原始客户端。以下操作以“/opt/Bigdata/client”为例进行说明。

- MRS普通集群，在console页面提交作业时，会使用master节点上预置安装的客户端进行作业提交。
- 用户也可使用master节点上预置安装的客户端来连接服务端、查看任务结果或管理数据等

对集群安装补丁后，用户需要重新更新master节点上的客户端，才能保证继续使用内置客户端功能。

操作步骤

步骤1 登录MRS Manager页面，具体请参见**访问MRS Manager（MRS 2.x及之前版本）**，然后选择“服务管理”。

步骤2 单击“下载客户端”。

“客户端类型”选择“完整客户端”，“下载路径”选择“服务器端”，单击“确定”开始生成客户端配置文件，文件生成后默认保存在主管理节点“/tmp/MRS-client”。文件保存路径支持自定义。

步骤3 查询并登录主Master节点。

步骤4 在弹性云服务器，切换到**root**用户，并将安装包复制到目录“/opt”。

sudo su - root

cp /tmp/MRS-client/MRS_Services_Client.tar /opt

步骤5 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Services_Client.tar

步骤6 执行以下命令，校验文件包。

sha256sum -c MRS_Services_ClientConfig.tar.sha256

界面显示如下：

```
MRS_Services_ClientConfig.tar: OK
```

步骤7 执行以下命令，解压“MRS_Services_ClientConfig.tar”。

tar -xvf MRS_Services_ClientConfig.tar

步骤8 执行以下命令，移走原来老的客户端到/opt/Bigdata/client_bak目录下

mv /opt/Bigdata/client /opt/Bigdata/client_bak

步骤9 执行以下命令，安装客户端到新的目录，客户端路径必须为“/opt/Bigdata/client”。

sh /opt/MRS_Services_ClientConfig/install.sh /opt/Bigdata/client

查看安装输出信息，如有以下结果表示客户端安装成功：

```
Components client installation is complete.
```

步骤10 执行以下命令，修改/opt/Bigdata/client目录的所属用户和用户组。

chown omm:wheel /opt/Bigdata/client -R

步骤11 执行以下命令配置环境变量：

source /opt/Bigdata/client/bigdata_env

步骤12 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit*MRS集群用户*

例如, **kinit admin**

步骤13 执行组件的客户端命令。

例如，执行以下命令查看HDFS目录：

hdfs dfs -ls /

----结束

全量更新备 Master 节点的原始客户端

步骤1 参见步骤1~步骤3登录备Master节点，执行如下命令切换到omm用户。

sudo su - omm

步骤2 在备master节点上执行如下命令，从主master节点拷贝下载的客户端包。

scp omm@*master1节点IP地址*:/tmp/MRS-client/MRS_Services_Client.tar /tmp/MRS-client/

说明

- 该命令以master1节点为主master节点为例。
- 目的路径以备master节点的/tmp/MRS-client/目录为例，请根据实际路径修改。

步骤3 参见步骤4~步骤13，更新备Master节点的客户端。

----结束

6.3 各组件客户端使用实践

6.3.1 使用 ClickHouse 客户端

ClickHouse是面向联机分析处理的列式数据库，支持SQL查询，且查询性能好，特别是基于大宽表的聚合分析查询性能非常优异，比其他分析型数据库速度快一个数量级。

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。在使用客户端前，需要先下载并更新客户端配置文件，确认Manager的主管理节点后才能使用客户端。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建ClickHouse表的权限，具体请参见“组件操作指南 > 使用ClickHouse > 用户管理及认证 > ClickHouse用户及权限管理”章节，为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行本步骤。

如果是MRS 3.1.0版本集群，则需要先执行：**export CLICKHOUSE_SECURITY_ENABLED=true**

kinit *组件业务用户*

例如，**kinit** clickhouseuser。

步骤5 执行ClickHouse组件的客户端命令。

执行**clickhouse -h**，查看ClickHouse组件命令帮助。

回显信息如下：

```
Use one of the following commands:
clickhouse local [args]
clickhouse client [args]
clickhouse benchmark [args]
clickhouse server [args]
clickhouse performance-test [args]
clickhouse extract-from-config [args]
clickhouse compressor [args]
clickhouse format [args]
clickhouse copier [args]
clickhouse obfuscator [args]
...
```

MRS 3.1.0及之后版本，使用**clickhouse client**命令连接ClickHouse服务端：

- 例如，当前集群未启用Kerberos认证，使用非ssl方式登录：

  clickhouse client --host *ClickHouse的实例IP* **--user** *用户名* **--password --port** 9000

  输入用户密码

- 例如，当前集群已启用Kerberos认证，使用ssl安全方式登录。

  Kerberos集群场景下没有默认用户，必须在Manager上创建用户，详细参考“组件操作指南 > 使用ClickHouse > 用户管理及认证 > ClickHouse用户及权限管理”。

  使用kinit认证成功后，客户端登录时可以不携带--user和--password参数，即使用kinit认证的用户登录。

  clickhouse client --host *ClickHouse的实例IP* **--port** 9440 **--secure**

  相关参数使用说明如下表：

表 6-4 clickhouse client 命令行参数说明

参数名	参数说明
--host	服务端的host名称，默认是localhost。您可以选择使用ClickHouse实例所在节点主机名或者IP地址。 **说明** ClickHouse的实例IP地址可登录集群FusionInsight Manager，然后选择“集群 > 服务 > ClickHouse > 实例”，获取ClickHouseServer实例对应的业务IP地址。

参数名	参数说明
--port	连接的端口。 ● 如果使用ssl安全连接则默认端口为9440，并且需要携带参数--secure。具体的端口值可通过ClickHouseServer实例配置搜索“tcp_port_secure”参数获取。 ● 如果使用非ssl安全连接则默认端口为9000，不需要携带参数--secure。具体的端口值可通过ClickHouseServer实例配置搜索“tcp_port”参数获取。
--user	用户名。 可以在Manager上创建该用户名并绑定对应的角色权限，具体可以参考“组件操作指南 > 使用ClickHouse > 用户管理及认证 > ClickHouse用户及权限管理”。 ● 如果当前集群已启用Kerberos认证（集群为安全模式），使用kinit认证成功后，客户端登录时可以不携带--user和--password参数，即使用kinit认证的用户登录。Kerberos集群场景下没有默认用户，必须在Manager上创建该用户名。 ● 如果当前集群未启用Kerberos认证（集群为普通模式），客户端登录时如果需要指定用户名和密码，不能使用FusionInsight Manager页面创建的ClickHouse用户，需要使用客户端命令行执行**create user SQL**语句或修改users.xml文件中用户信息的方式来创建ClickHouse用户。客户端登录时如果不需要指定用户名和密码参数时，默认使用default用户登录。
--password	密码。 默认值：空字符串。该参数和--user参数配套使用，可以在Manager上创建用户名时设置该密码。
--query	使用非交互模式查询。
--database	默认当前操作的数据库。默认值：服务端默认的配置（默认是default）。
--multiline	如果指定，允许多行语句查询（Enter仅代表换行，不代表查询语句完结）。
--multiquery	如果指定，允许处理用;号分隔的多个查询，只在非交互模式下生效。
--format	使用指定的默认格式输出结果。
--vertical	如果指定，默认情况下使用垂直格式输出结果。在这种格式中，每个值都在单独的行上打印，适用显示宽表的场景。
--time	如果指定，非交互模式下会打印查询执行的时间到stderr中。
--stacktrace	如果指定，如果出现异常，会打印堆栈跟踪信息。
--config-file	配置文件的名称。
--secure	如果指定，将通过ssl安全模式连接到服务器。
--history_file	存放命令历史的文件的路径。

参数名	参数说明
--param_<name>	带有参数的查询，并将值从客户端传递给服务器。具体用法详见 https://clickhouse.tech/docs/zh/interfaces/cli/#cli-queries-with-parameters。

----结束

6.3.2 使用 Flink 客户端

本节提供使用Flink运行wordcount作业的操作指导。

前提条件

- MRS集群中已安装Flink组件。
- 集群正常运行，已安装集群客户端，例如安装目录为“/opt/hadoopclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

使用 Flink 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行如下命令初始化环境变量。

source /opt/hadoopclient/bigdata_env

步骤4 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 登录Manager，下载认证凭据。

   登录集群的Manager界面，具体请参见访问MRS Manager（MRS 2.x及之前版本），选择“系统设置 > 用户管理”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/hadoopclient/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。
4. 配置安全认证，在“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/hadoopclient/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test

5. 在Flink的客户端bin目录下，执行如下命令进行安全加固，请参考“组件操作指南 > 使用Flink > 安全加固 > 认证和加密”，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中关于SSL的值，针对MRS2.x及之前版本，安全集群默认没有开启外部SSL，用户如果需要启用外部SSL，请参考“组件操作指南 > 使用Flink > 安全加固”进行配置后再次运行该脚本即可。

   说明

   - generate_keystore.sh脚本无需手动生成。
   - 执行认证和加密后会将生成的flink.keystore、flink.truststore、security.cookie自动填充到“flink-conf.yaml”对应配置项中。

6. 客户端访问flink.keystore和flink.truststore文件的路径配置。
   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/hadoopclient/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。
   - 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
     i. 在“/opt/hadoopclient/Flink/flink/conf/”目录下新建目录，例如ssl。

        cd /opt/hadoopclient/Flink/flink/conf/

        mkdir ssl

     ii. 移动flink.keystore和flink.truststore文件到“/opt/hadoopclient/Flink/flink/conf/ssl/”中。

        mv flink.keystore ssl/

        mv flink.truststore ssl/

     iii. 修改flink-conf.yaml文件中如下两个参数为相对路径。

```
security.ssl.internal.keystore: ssl/flink.keystore
security.ssl.internal.truststore: ssl/flink.truststore
```

步骤5 运行wordcount作业。

须知

用户在Flink提交作业或者运行作业时，应具有如下权限：

- 如果启用Ranger鉴权，当前用户必须属于hadoop组或者已在Ranger中为该用户添加“/flink”的读写权限。
- 如果停用Ranger鉴权，当前用户必须属于hadoop组。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。

    yarn-session.sh -nm "*session-name***"**

    flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - 执行如下命令在Yarn上提交单个作业。

flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。

      yarn-session.sh -nm "*session-name***"**

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - flink.keystore和flink.truststore文件路径为相对路径时：
    - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/hadoopclient/Flink/flink/conf/”，则在“opt/hadoopclient/Flink/flink/conf/”目录下执行命令。

      yarn-session.sh -t ssl/ -nm "*session-name***"**

      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

    - 执行如下命令在Yarn上提交单个作业。

      flink run -m yarn-cluster -yt ssl/ /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

步骤6 作业提交成功后，客户端界面显示如下。

图 6-3 在 Yarn 上提交作业成功

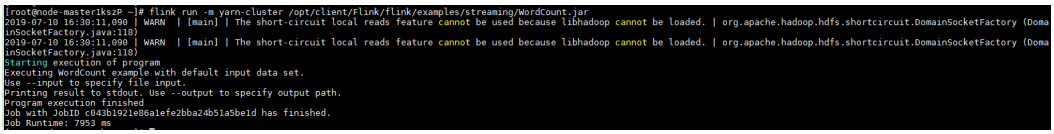


图 6-4 启动 session 成功

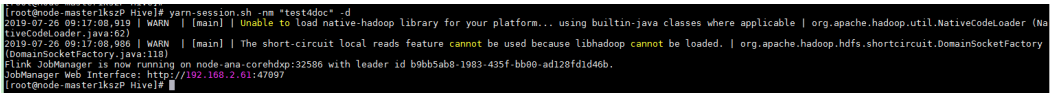
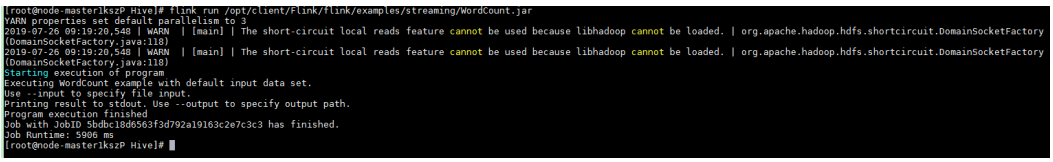


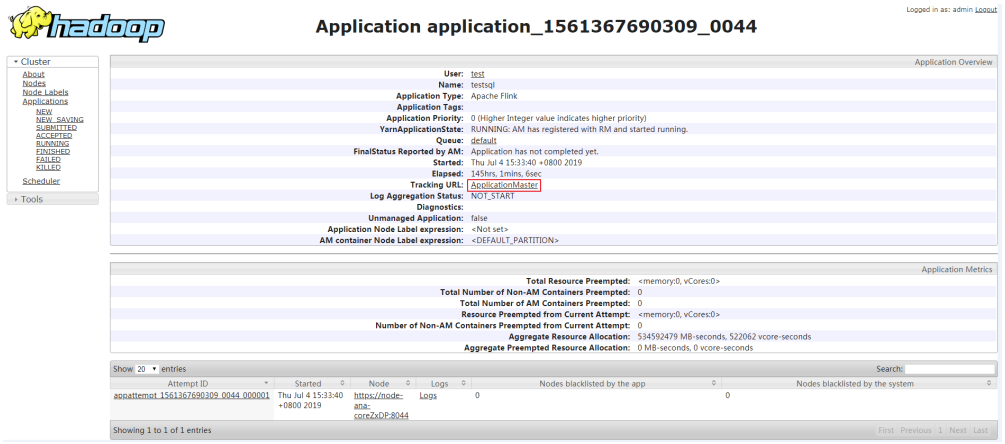
图 6-5 在 session 中提交作业成功



步骤7 使用运行用户进入Yarn服务的原生页面，具体操作参考“组件操作指南 > 使用Flink > 查看Flink作业”，找到对应作业的application，单击application名称，进入到作业详情页面。

- 若作业尚未结束，可单击“Tracking URL”链接进入到Flink的原生页面，查看作业的运行信息。
- 若作业已运行结束，对于在session中提交的作业，可以单击“Tracking URL”链接登录Flink原生页面查看作业信息。

图 6-6 application



----结束

使用 Flink 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行如下命令初始化环境变量。

source /opt/hadoopclient/bigdata_env

步骤4 若集群开启Kerberos认证，需要执行以下步骤，若集群未开启Kerberos认证请跳过该步骤。

1. 准备一个提交Flink作业的用户。
2. 登录Manager，下载认证凭据。

   登录集群的Manager界面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“系统 > 权限 > 用户”，在已增加用户所在行的“操作”列，选择“更多 > 下载认证凭据”。
3. 将下载的认证凭据压缩包解压缩，并将得到的user.keytab文件拷贝到客户端节点中，例如客户端节点的“/opt/hadoopclient/Flink/flink/conf”目录下。如果是在集群外节点安装的客户端，需要将得到的krb5.conf文件拷贝到该节点的“/etc/”目录下。
4. 将客户端安装节点的业务IP和Master节点IP添加到配置文件“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中的“jobmanager.web.access-control-allow-origin”和“jobmanager.web.allow-access-address”配置项中，IP地址之间使用英文逗号分隔。

   ```
   jobmanager.web.access-control-allow-origin: xx.xx.xxx.xxx,xx.xx.xxx.xxx,xx.xx.xxx.xxx
   jobmanager.web.allow-access-address: xx.xx.xxx.xxx,xx.xx.xxx.xxx,xx.xx.xxx.xxx
   ```

说明

客户端安装节点的业务IP获取方法：

- 集群内节点：

  登录MapReduce服务管理控制台，选择“集群列表 > 现有集群”，选中当前的集群并单击集群名，进入集群信息页面。

  在“节点管理”中查看安装客户端所在的节点IP。
- 集群外节点：安装客户端所在的弹性云服务器的IP。

5. 配置安全认证，在“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”配置文件中的对应配置添加keytab路径以及用户名。

   security.kerberos.login.keytab: <*user.keytab文件路径*>

   security.kerberos.login.principal: <*用户名*>

   例如：

   security.kerberos.login.keytab: /opt/hadoopclient/Flink/flink/conf/user.keytab

   security.kerberos.login.principal: test

6. 在Flink的客户端bin目录下，执行如下命令进行安全加固，请参考“组件操作指南 > 使用Flink > 安全加固 > 认证和加密”，password请重新设置为一个用于提交作业的密码。

   sh generate_keystore.sh <*password*>

   该脚本会自动替换“/opt/hadoopclient/Flink/flink/conf/flink-conf.yaml”中关于SSL的值。

   sh generate_keystore.sh <*password*>

   说明

   执行认证和加密后会在Flink客户端的“conf”目录下生成“flink.keystore”和“flink.truststore”文件，并且在客户端配置文件“flink-conf.yaml”中将以下配置项进行了默认赋值：

   - 将配置项“security.ssl.keystore”设置为“flink.keystore”文件所在绝对路径。
   - 将配置项“security.ssl.truststore”设置为“flink.truststore”文件所在的绝对路径。
   - 将配置项“security.cookie”设置为“generate_keystore.sh”脚本自动生成的一串随机规则密码。
   - 默认“flink-conf.yaml”中“security.ssl.encrypt.enabled: false”，“generate_keystore.sh”脚本将配置项“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值设置为调用“generate_keystore.sh”脚本时输入的密码。
   - MRS 3.x及之后版本，如果需要使用密文时，设置“flink-conf.yaml”中“security.ssl.encrypt.enabled: true”，“generate_keystore.sh”脚本不会配置“security.ssl.key-password”、“security.ssl.keystore-password”和“security.ssl.truststore-password”的值，需要使用Manager明文加密API进行获取，执行**curl -k -i -u** *user name*:*password* **-X POST -HContent-type:application/json -d '{"plainText":"***password***"}' 'https://***x.x.x.x***:28443/web/api/v2/tools/encrypt'**

     其中*user name*:password分别为当前系统登录用户名和密码；**"plainText"**的password为调用“generate_keystore.sh”脚本时的密码；x.x.x.x为集群Manager的浮动IP。

7. 客户端访问flink.keystore和flink.truststore文件的路径配置。

   - 绝对路径：执行该脚本后，在flink-conf.yaml文件中将flink.keystore和flink.truststore文件路径自动配置为绝对路径“/opt/hadoopclient/Flink/flink/conf/”，此时需要将conf目录中的flink.keystore和flink.truststore文件分别放置在Flink Client以及Yarn各个节点的该绝对路径上。

- 相对路径：请执行如下步骤配置flink.keystore和flink.truststore文件路径为相对路径，并确保Flink Client执行命令的目录可以直接访问该相对路径。
  i. 在“/opt/hadoopclient/Flink/flink/conf/”目录下新建目录，例如ssl。
     cd /opt/hadoopclient/Flink/flink/conf/
     mkdir ssl
  ii. 移动flink.keystore和flink.truststore文件到“/opt/hadoopclient/Flink/flink/conf/ssl/”中。
     mv flink.keystore ssl/
     mv flink.truststore ssl/
  iii. 修改flink-conf.yaml文件中如下两个参数为相对路径。
     ```
     security.ssl.keystore: ssl/flink.keystore
     security.ssl.truststore: ssl/flink.truststore
     ```

步骤5 运行wordcount作业。

须知

用户在Flink提交作业或者运行作业时，应具有如下权限：

- 如果启用Ranger鉴权，当前用户必须属于hadoop组或者已在Ranger中为该用户添加“/flink”的读写权限。
- 如果停用Ranger鉴权，当前用户必须属于hadoop组。

- 普通集群（未开启Kerberos认证）
  - 执行如下命令启动session，并在session中提交作业。
    yarn-session.sh -nm "*session-name***"**
    flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - 执行如下命令在Yarn上提交单个作业。
    flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
- 安全集群（开启Kerberos认证）
  - flink.keystore和flink.truststore文件路径为绝对路径时：
    - 执行如下命令启动session，并在session中提交作业。
      yarn-session.sh -nm "*session-name***"**
      flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
    - 执行如下命令在Yarn上提交单个作业。
      flink run -m yarn-cluster /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar
  - flink.keystore和flink.truststore文件路径为相对路径时：
    - 在“ssl”的同级目录下执行如下命令启动session，并在session中提交作业，其中“ssl”是相对路径，如“ssl”所在目录是“opt/hadoopclient/Flink/flink/conf/”，则在“opt/hadoopclient/Flink/flink/conf/”目录下执行命令。

yarn-session.sh -t ssl/ -nm "*session-name*"

flink run /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

- 执行如下命令在Yarn上提交单个作业。

  flink run -m yarn-cluster -yt ssl/ /opt/hadoopclient/Flink/flink/examples/streaming/WordCount.jar

步骤6 作业提交成功后，客户端界面显示如下。

图 6-7 在 Yarn 上提交作业成功

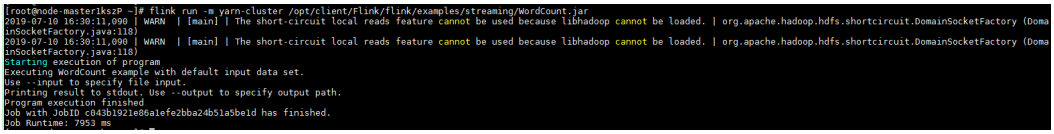
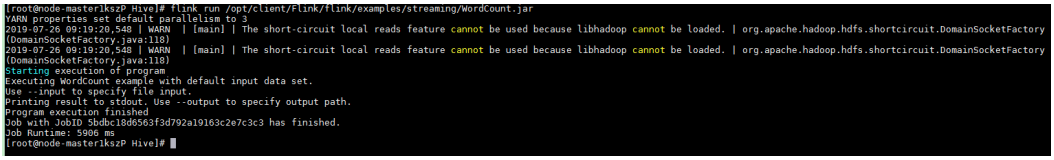


图 6-8 启动 session 成功

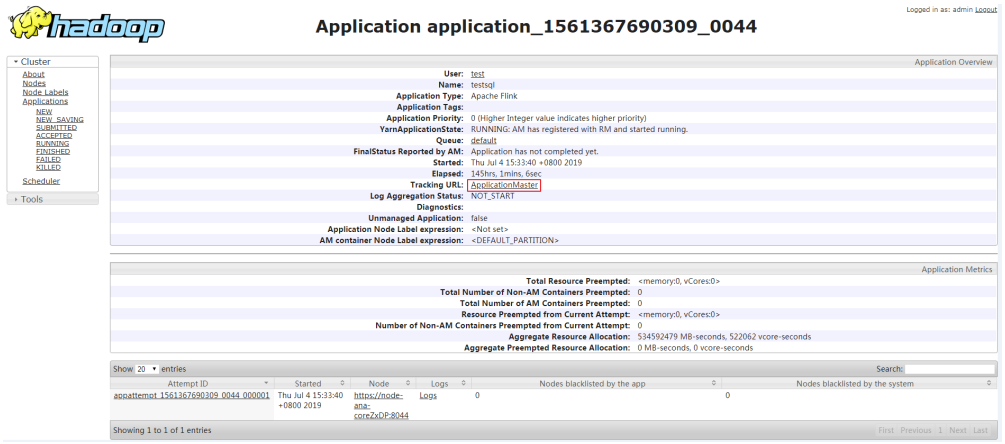
图 6-9 在 session 中提交作业成功



步骤7 使用运行用户进入Yarn服务的原生页面，具体操作参考“组件操作指南 > 使用Flink > 查看Flink作业”，找到对应作业的application，单击application名称，进入到作业详情页面

- 若作业尚未结束，可单击“Tracking URL”链接进入到Flink的原生页面，查看作业的运行信息。
- 若作业已运行结束，对于在session中提交的作业，可以单击“Tracking URL”链接登录Flink原生页面查看作业信息。

图 6-10 application



----结束

6.3.3 使用 Flume 客户端

操作场景

Flume支持将采集的日志信息导入到Kafka。

前提条件

- 已创建开启Kerberos认证的包含Flume、Kafka等组件的流式集群。
- 已配置网络，使日志生成节点与流集群互通。

使用 Flume 客户端（MRS 3.x 之前版本）

说明

普通集群不需要执行**步骤2-步骤6**。

步骤1 安装Flume客户端。

可参考**安装MRS 3.x之前版本Flume客户端**在日志生成节点安装Flume客户端，例如安装目录为“/opt/Flumeclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

步骤2 将Master1节点上的认证服务器配置文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为${BIGDATA_HOME}/MRS_Current/1_*X*_KerberosClient/etc/kdc.conf。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤3 查看任一部署Flume角色节点的“业务IP”：

- MRS 2.0.1之前版本，登录MRS Manager，选择“集群 > 服务管理 > Flume > 实例”，查看任一部署Flume角色节点的“业务IP”。
- MRS 2.0.1及至3.x之前版本，在MRS控制台单击集群名称，选择“集群 > 组件管理 > Flume > 实例”，查看任一部署Flume角色节点的“业务IP”。

步骤4 将此节点上的用户认证文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为${BIGDATA_HOME}/MRS_*XXX*/install/FusionInsight-Flume-*Flume组件版本号*/flume/conf/flume.keytab。

其中“XXX”为产品版本号，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤5 将此节点上的配置文件“jaas.conf”，复制到安装Flume客户端的节点，保存到Flume客户端中“conf”目录。

文件完整路径为${BIGDATA_HOME}/MRS_Current/1_*X*_Flume/etc/jaas.conf。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤6 登录安装Flume客户端节点，切换到客户端安装目录，执行以下命令修改文件：

vi conf/jaas.conf

修改参数“keyTab”定义的用户认证文件完整路径即**步骤4**中保存用户认证文件的目录：“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”，然后保存并退出。

步骤7 执行以下命令，修改Flume客户端配置文件“flume-env.sh”：

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号***/conf/flume-env.sh**

在“-XX:+UseCMSCompactAtFullCollection”后面，增加以下内容：

```
-Djava.security.krb5.conf=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/kdc.conf -Djava.security.auth.login.config=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/jaas.conf -Dzookeeper.request.timeout=120000
```

例如：**"-XX:+UseCMSCompactAtFullCollection -Djava.security.krb5.conf=/opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/conf/kdc.conf -Djava.security.auth.login.config=/opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/conf/jaas.conf -Dzookeeper.request.timeout=120000"**

请根据实际情况，修改“Flume客户端安装目录”，然后保存并退出。

步骤8 执行以下命令，重启Flume客户端：

cd *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号*/**bin**

./flume-manage.sh restart

例如：

cd /opt/FlumeClient/fusioninsight-flume-*Flume组件版本号*/**bin**

./flume-manage.sh restart

步骤9 执行以下命令，根据实际业务需求，在Flume客户端配置文件“properties.properties”中配置并保存作业。

vi *Flume客户端安装目录*/**fusioninsight-flume-***Flume组件版本号*/**conf/properties.properties**

以配置SpoolDir Source+File Channel+Kafka Sink为例：

```
################################################################################
################
client.sources = static_log_source
client.channels = static_log_channel
client.sinks = kafka_sink
################################################################################
################
#LOG_TO_HDFS_ONLINE_1

client.sources.static_log_source.type = spooldir
client.sources.static_log_source.spoolDir = 监控目录
client.sources.static_log_source.fileSuffix = .COMPLETED
client.sources.static_log_source.ignorePattern = ^$
client.sources.static_log_source.trackerDir = 传输过程中元数据存储路径
client.sources.static_log_source.maxBlobLength = 16384
client.sources.static_log_source.batchSize = 51200
client.sources.static_log_source.inputCharset = UTF-8
client.sources.static_log_source.deserializer = LINE
client.sources.static_log_source.selector.type = replicating
client.sources.static_log_source.fileHeaderKey = file
client.sources.static_log_source.fileHeader = false
client.sources.static_log_source.basenameHeader = true
client.sources.static_log_source.basenameHeaderKey = basename
client.sources.static_log_source.deletePolicy = never

client.channels.static_log_channel.type = file
client.channels.static_log_channel.dataDirs = 数据缓存路径，设置多个路径可提升性能，中间用逗号分开
client.channels.static_log_channel.checkpointDir = 检查点存放路径
client.channels.static_log_channel.maxFileSize = 2146435071
client.channels.static_log_channel.capacity = 1000000
client.channels.static_log_channel.transactionCapacity = 612000
client.channels.static_log_channel.minimumRequiredSpace = 524288000

client.sinks.kafka_sink.type = org.apache.flume.sink.kafka.KafkaSink
client.sinks.kafka_sink.kafka.topic = 数据写入的topic，如flume_test
client.sinks.kafka_sink.kafka.bootstrap.servers = XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口
号,XXX.XXX.XXX.XXX:kafka端口号
client.sinks.kafka_sink.flumeBatchSize = 1000
client.sinks.kafka_sink.kafka.producer.type = sync
client.sinks.kafka_sink.kafka.security.protocol = SASL_PLAINTEXT
client.sinks.kafka_sink.kafka.kerberos.domain.name = Kafka Domain名称，安全集群必填，如hadoop.xxx.com
client.sinks.kafka_sink.requiredAcks = 0

client.sources.static_log_source.channels = static_log_channel
client.sinks.kafka_sink.channel = static_log_channel
```

说明

- client.sinks.kafka_sink.kafka.topic：数据写入的topic。若kafka中该topic不存在，默认情况下会自动创建该topic。
- client.sinks.kafka_sink.kafka.bootstrap.servers：Kafkabrokers列表，多个用英文逗号分隔。默认情况下，安全集群端口21007，普通集群对应端口9092。
- client.sinks.kafka_sink.kafka.security.protocol：安全集群为SASL_PLAINTEXT，普通集群为PLAINTEXT。
- client.sinks.kafka_sink.kafka.kerberos.domain.name：

  普通集群无需配置此参数。安全集群对应此参数的值为Kafka集群中“kerberos.domain.name”对应的值。

  具体可到Broker实例所在节点上查看${BIGDATA_HOME}/MRS_Current/1_*X*_Broker/etc/server.properties。

  其中X为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤10 参数配置并保存后，Flume客户端将自动加载“properties.properties”中配置的内容。当spoolDir生成新的日志文件，文件内容将发送到Kafka生产者，并支持Kafka消费者消费。

----结束

使用 Flume 客户端（MRS 3.x 及之后版本）

说明

普通集群不需要执行**步骤2-步骤6**。

步骤1 安装Flume客户端。

可参考**安装MRS 3.x及之后版本Flume客户端**在日志生成节点安装Flume客户端，例如安装目录为“/opt/Flumeclient”。以下操作的客户端目录只是举例，请根据实际安装目录修改。

步骤2 将Master1节点上的认证服务器配置文件，复制到安装Flume客户端的节点，保存到Flume客户端中*Flume客户端安装目录*/fusioninsight-flume-*Flume组件版本号*/conf目录下。

文件完整路径为“${BIGDATA_HOME}/FusionInsight_BASE_*XXX*/1_*X*_KerberosClient/etc/kdc.conf”。其中“XXX”为产品版本号，“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤3 查看任一部署Flume角色节点的“业务IP”：

登录FusionInsight Manager页面，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**，选择“集群 > 服务 > Flume > 实例”。查看任一部署Flume角色节点的“业务IP”。

步骤4 将此节点上的用户认证文件，复制到安装Flume客户端的节点，保存到Flume客户端中“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”目录下。

文件完整路径为${BIGDATA_HOME}/FusionInsight_Porter_*XXX*/install/FusionInsight-Flume-*Flume组件版本号*/flume/conf/flume.keytab。

其中“XXX”为产品版本号，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤5 将此节点上的配置文件“jaas.conf”，复制到安装Flume客户端的节点，保存到Flume客户端中“conf”目录。

文件完整路径为${BIGDATA_HOME}/FusionInsight_Current/1_*X*_Flume/etc/jaas.conf。

其中“X”为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤6 登录安装Flume客户端节点，切换到客户端安装目录，执行以下命令修改文件：

vi conf/jaas.conf

修改参数“keyTab”定义的用户认证文件完整路径即**步骤4**中保存用户认证文件的目录：“Flume客户端安装目录/fusioninsight-flume-*Flume组件版本号*/conf”，然后保存并退出。

步骤7 执行以下命令，修改Flume客户端配置文件“flume-env.sh”：

vi *Flume客户端安装目录*/**fusioninsight-flume-**_Flume组件版本号_**/conf/flume-env.sh**

在“-XX:+UseCMSCompactAtFullCollection”后面，增加以下内容：

```
-Djava.security.krb5.conf=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/kdc.conf -Djava.security.auth.login.config=Flume客户端安装目录/fusioninsight-flume-1.9.0/conf/jaas.conf -Dzookeeper.request.timeout=120000
```

例如：**"-XX:+UseCMSCompactAtFullCollection -Djava.security.krb5.conf=/opt/FlumeClient/fusioninsight-flume-**_Flume组件版本号_**/conf/kdc.conf -Djava.security.auth.login.config=*/opt/FlumeClient*/fusioninsight-flume-**_Flume组件版本号_**/conf/jaas.conf -Dzookeeper.request.timeout=120000"**

请根据实际情况，修改“Flume客户端安装目录”，然后保存并退出。

步骤8 执行以下命令，重启Flume客户端：

cd *Flume客户端安装目录*/**fusioninsight-flume-**_Flume组件版本号_**/bin**

./flume-manage.sh restart

例如：

cd /opt/FlumeClient/fusioninsight-flume-_Flume组件版本号_**/bin**

./flume-manage.sh restart

步骤9 根据实际业务场景配置作业。

- MRS 3.x及之后版本部分参数可直接在Manager界面配置。
- 在“properties.properties”文件中配置，以配置SpoolDir Source+File Channel +Kafka Sink为例。

  在安装Flume客户端的节点执行以下命令，根据实际业务需求，在Flume客户端配置文件“properties.properties”中配置并保存作业。

  vi *Flume客户端安装目录*/**fusioninsight-flume-**_Flume组件版本号_**/conf/properties.properties**

  ```
  ################################################################################################################
  client.sources = static_log_source
  client.channels = static_log_channel
  client.sinks = kafka_sink
  ################################################################################################################
  #LOG_TO_HDFS_ONLINE_1

  client.sources.static_log_source.type = spooldir
  client.sources.static_log_source.spoolDir = 监控目录
  client.sources.static_log_source.fileSuffix = .COMPLETED
  client.sources.static_log_source.ignorePattern = ^$
  client.sources.static_log_source.trackerDir = 传输过程中元数据存储路径
  client.sources.static_log_source.maxBlobLength = 16384
  client.sources.static_log_source.batchSize = 51200
  client.sources.static_log_source.inputCharset = UTF-8
  client.sources.static_log_source.deserializer = LINE
  client.sources.static_log_source.selector.type = replicating
  client.sources.static_log_source.fileHeaderKey = file
  client.sources.static_log_source.fileHeader = false
  client.sources.static_log_source.basenameHeader = true
  client.sources.static_log_source.basenameHeaderKey = basename
  client.sources.static_log_source.deletePolicy = never
  ```

```
client.channels.static_log_channel.type = file
client.channels.static_log_channel.dataDirs = 数据缓存路径，设置多个路径可提升性能，中间用逗号分开
client.channels.static_log_channel.checkpointDir = 检查点存放路径
client.channels.static_log_channel.maxFileSize = 2146435071
client.channels.static_log_channel.capacity = 1000000
client.channels.static_log_channel.transactionCapacity = 612000
client.channels.static_log_channel.minimumRequiredSpace = 524288000

client.sinks.kafka_sink.type = org.apache.flume.sink.kafka.KafkaSink
client.sinks.kafka_sink.kafka.topic = 数据写入的topic，如flume_test
client.sinks.kafka_sink.kafka.bootstrap.servers = XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口号,XXX.XXX.XXX.XXX:kafka端口号
client.sinks.kafka_sink.flumeBatchSize = 1000
client.sinks.kafka_sink.kafka.producer.type = sync
client.sinks.kafka_sink.kafka.security.protocol = SASL_PLAINTEXT
client.sinks.kafka_sink.kafka.kerberos.domain.name = Kafka Domain名称，安全集群必填，如hadoop.xxx.com
client.sinks.kafka_sink.requiredAcks = 0

client.sources.static_log_source.channels = static_log_channel
client.sinks.kafka_sink.channel = static_log_channel
```

说明

- client.sinks.kafka_sink.kafka.topic：数据写入的topic。若kafka中该topic不存在，默认情况下会自动创建该topic。
- client.sinks.kafka_sink.kafka.bootstrap.servers：Kafkabrokers列表，多个用英文逗号分隔。默认情况下，安全集群端口21007，普通集群对应端口9092。
- client.sinks.kafka_sink.kafka.security.protocol：安全集群为SASL_PLAINTEXT，普通集群为PLAINTEXT。
- client.sinks.kafka_sink.kafka.kerberos.domain.name：

  普通集群无需配置此参数。安全集群对应此参数的值为Kafka集群中“kerberos.domain.name”对应的值。

  具体可到Broker实例所在节点上查看${BIGDATA_HOME}/MRS_Current/1_*X*_Broker/etc/server.properties。

  其中X为随机生成的数字，请根据实际情况修改。同时文件需要以Flume客户端安装用户身份保存，例如**root**用户。

步骤10 参数配置并保存后，Flume客户端将自动加载“properties.properties”中配置的内容。当spoolDir生成新的日志文件，文件内容将发送到Kafka生产者，并支持Kafka消费者消费。

----结束

6.3.4 使用 HBase 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用HBase客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。

  “机机”用户需要下载keytab文件，“人机”用户第一次登录时需修改密码。
- 非**root**用户使用HBase客户端，请确保该HBase客户端目录的属主为该用户，否则请参考如下命令修改属主。

chown user:group -R *客户端安装目录*/**HBase**

使用 Hbase 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令切换到客户端目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限，具体请参见“用户指南 > 管理现有集群 > MRS多用户权限管理 > 创建角色”配置拥有对应权限的角色，参考“用户指南 > 管理现有集群 > MRS多用户权限管理 > 创建用户”章节，为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

步骤5 直接执行HBase组件的客户端命令。

hbase shell

----结束

使用 HBase 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令切换到客户端目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若安装了HBase多实例，在使用客户端连接具体HBase实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载HBase2实例变量：

source HBase2/component_env

步骤5 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限，具体请参见“用户指南 > 管理现有集群 > MRS多用户权限管理 > 创建角色”配置拥有对应权限的角色，参考“用户指南 > 管理现有集群 > MRS多用户权限管理 > 创建用户”章节，为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *组件业务用户*

例如，**kinit hbaseuser**。

步骤6 直接执行HBase组件的客户端命令。

hbase shell

----结束

HBase 客户端常用命令

常用的HBase客户端命令如下表所示。更多命令可参考http://hbase.apache.org/2.2/book.html

表 6-5 HBase 客户端命令

命令	说明
create	创建一张表，例如**create 'test', 'f1', 'f2', 'f3'**。
disable	停止指定的表，例如**disable 'test'**。
enable	启动指定的表，例如**enable 'test'**。
alter	更改表结构。可以通过alter命令增加、修改、删除列族信息以及表相关的参数值，例如**alter 'test', {NAME => 'f3', METHOD => 'delete'}**。
describe	获取表的描述信息，例如**describe 'test'**。
drop	删除指定表。删除前表必须已经是停止状态，例如**drop 'test'**。
put	写入指定cell的value。Cell的定位由表、rowk、列组合起来唯一决定，例如**put 'test','r1','f1:c1','myvalue1'**。
get	获取行的值或者行的指定cell的值。例如**get 'test','r1'**。
scan	查询表数据。参数中指定表名和scanner，例如**scan 'test'**。

6.3.5 使用 HDFS 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用HDFS客户端。

前提条件

- 已安装客户端。

  例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由MRS集群管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）

使用 HDFS 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行HDFS Shell命令。例如：

hdfs dfs -ls /

----结束

HDFS 客户端常用命令

常用的HDFS客户端命令如下表所示。

更多命令可参考**https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/CommandsManual.html#User_Commands**

表 6-6 HDFS 客户端常用命令

命令	说明	样例
hdfs dfs -mkdir *文件夹名称*	创建文件夹	**hdfs dfs -mkdir /tmp/mydir**
hdfs dfs -ls *文件夹名称*	查看文件夹	**hdfs dfs -ls /tmp**
hdfs dfs -put *客户端节点上本地文件 HDFS指定路径*	上传本地文件到HDFS指定路径	**hdfs dfs -put /opt/test.txt /tmp** 上传客户端节点“/opt/test.txt”文件到HDFS的“/tmp”路径下
hdfs dfs -get *hdfs指定文件 客户端节点上指定路径*	下载HDFS文件到本地指定路径	**hdfs dfs -get /tmp/test.txt /opt/** 下载HDFS的“/tmp/test.txt”文件到客户端节点的“/opt”路径下
hdfs dfs -rm -r -f *hdfs指定文件夹*	删除文件夹	**hdfs dfs -rm -r -f /tmp/mydir**
hdfs dfs -chmod *权限参数 文件目录*	为用户设置HDFS目录权限	**hdfs dfs -chmod 700 /tmp/test**

客户端常见使用问题

1. 当执行HDFS客户端命令时，客户端程序异常退出，报“java.lang.OutOfMemoryError”的错误。

   这个问题是由于HDFS客户端运行时的所需的内存超过了HDFS客户端设置的内存上限（默认为128MB）。可以通过修改“*<客户端安装路径>*/HDFS/

component_env”中的“CLIENT_GC_OPTS”来修改HDFS客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

```
CLIENT_GC_OPTS="-Xmx1G"
```

在修改完后，使用如下命令刷新客户端配置，使之生效：

source <*客户端安装路径*>/**bigdata_env**

2. 如何设置HDFS客户端运行时的日志级别?

   HDFS客户端运行时的日志是默认输出到Console控制台的，其级别默认是INFO级别。有的时候为了定位问题，需要开启DEBUG级别日志，可以通过导出一个环境变量来设置，命令如下：

   export HADOOP_ROOT_LOGGER=DEBUG,console

   在执行完上面命令后，再执行HDFS Shell命令时，即可打印出DEBUG级别日志。

   如果想恢复INFO级别日志，可执行如下命令：

   export HADOOP_ROOT_LOGGER=INFO,console

3. 如何彻底删除HDFS文件?

   由于HDFS的回收站机制，一般删除HDFS文件后，文件会移动到HDFS的回收站中。如果确认文件不再需要并且需要立马释放存储空间，可以继续清理对应的回收站目录（例如：hdfs://hacluster/user/*xxx*/.Trash/Current/*xxx*）。

6.3.6 使用 Hive 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Hive客户端。

前提条件

- 已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。

使用 Hive 客户端（MRS 3.x 之前版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 根据集群认证模式，完成Hive客户端登录。

- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。

  kinit *组件业务用户*

  beeline

- 普通模式，则执行以下命令，登录Hive客户端，如果不指定组件业务用户，则会以当前操作系统用户登录。

  beeline -n *组件业务用户*

说明

进行beeline连接后，可以编写并提交HQL语句执行相关任务。如需执行Catalog客户端命令，需要先执行**!q**命令退出beeline环境。

步骤5 使用以下命令，执行HCatalog的客户端命令。

hcat -e *"cmd"*

其中*"cmd"*必须为Hive DDL语句，如**hcat -e "show tables"**。

说明

- 若要使用HCatalog客户端，必须从“组件管理”页面单击“下载客户端”，下载全部服务的客户端。Beeline客户端不受此限制。
- 由于权限模型不兼容，使用HCatalog客户端创建的表，在HiveServer客户端中不能访问，但可以使用WebHCat客户端访问。
- 在普通模式下使用HCatalog客户端，系统将以当前登录操作系统用户来执行DDL命令。
- 退出beeline客户端时请使用**!q**命令，不要使用“Ctrl + c”。否则会导致连接生成的临时文件无法删除，长期会累积产生大量的垃圾文件。
- 在使用beeline客户端时，如果需要在一行中输入多条语句，语句之间以“;”分隔，需要将“entireLineAsCommand”的值设置为“false”。

  设置方法：如果未启动beeline，则执行**beeline --entireLineAsCommand=false**命令；如果已启动beeline，则在beeline中执行**!set entireLineAsCommand false**命令。

  设置完成后，如果语句中含有不是表示语句结束的“;”，需要进行转义，例如**select concat_ws('\;', collect_set(col1)) from tbl**。

----结束

使用 Hive 客户端（MRS 3.x 及之后版本）

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 MRS 3.X支持Hive多实例，若安装了Hive多实例，在使用客户端连接具体Hive实例时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，加载Hive2实例变量：

source Hive2/component_env

步骤5 根据集群认证模式，完成Hive客户端登录。

- 安全模式，则执行以下命令，完成用户认证并登录Hive客户端。

  kinit *组件业务用户*

  beeline

- 普通模式，则执行以下命令，登录Hive客户端，如果不指定组件业务用户，则会以当前操作系统用户登录。

  beeline -n *组件业务用户*

步骤6 使用以下命令，执行HCatalog的客户端命令。

hcat -e *"cmd"*

其中*"cmd"*必须为Hive DDL语句，如**hcat -e "show tables"**。

说明

- 若要使用HCatalog客户端，必须从服务页面选择“更多 > 下载客户端”，下载全部服务的客户端。Beeline客户端不受此限制。
- 由于权限模型不兼容，使用HCatalog客户端创建的表，在HiveServer客户端中不能访问，但可以使用WebHCat客户端访问。
- 在普通模式下使用HCatalog客户端，系统将以当前登录操作系统用户来执行DDL命令。
- 退出beeline客户端时请使用**!q**命令，不要使用“Ctrl + C”。否则会导致连接生成的临时文件无法删除，长期会累积产生大量的垃圾文件。
- 在使用beeline客户端时，如果需要在一行中输入多条语句，语句之间以“;”分隔，需要将“entireLineAsCommand”的值设置为“false”。

  设置方法：如果未启动beeline，则执行**beeline --entireLineAsCommand=false**命令；如果已启动beeline，则在beeline中执行**!set entireLineAsCommand false**命令。

  设置完成后，如果语句中含有不是表示语句结束的“;”，需要进行转义，例如**select concat_ws('\;', collect_set(col1)) from tbl**。

----结束

Hive 客户端常用命令

常用的Hive Beeline客户端命令如下表所示。

更多命令可参考**https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients#HiveServer2Clients-BeelineCommands**。

表 6-7 Hive Beeline 客户端常用命令

命令	说明
set <key>=<value>	设置特定配置变量（键）的值。 **说明** 若变量名拼错，Beeline不会显示错误。
set	打印由用户或Hive覆盖的配置变量列表。
set -v	打印Hadoop和Hive的所有配置变量。
add FILE[S] <filepath> <filepath>*add JAR[S] <filepath> <filepath>*add ARCHIVE[S] <filepath> <filepath>*	将一个或多个文件、JAR文件或ARCHIVE文件添加至分布式缓存的资源列表中。
add FILE[S] <ivyurl> <ivyurl>* add JAR[S] <ivyurl> <ivyurl>* add ARCHIVE[S] <ivyurl> <ivyurl>*	使用“ivy://goup:module:version?query_string”格式的Ivy URL，将一个或多个文件、JAR文件或ARCHIVE文件添加至分布式缓存的资源列表中。

命令	说明
list FILE[S]list JAR[S]list ARCHIVE[S]	列出已添加至分布式缓存中的资源。
list FILE[S] <filepath>*list JAR[S] <filepath>*list ARCHIVE[S] <filepath>*	检查给定的资源是否已添加至分布式缓存中。
delete FILE[S] <filepath>*delete JAR[S] <filepath>*delete ARCHIVE[S] <filepath>*	从分布式缓存中删除资源。
delete FILE[S] <ivyurl> <ivyurl>* delete JAR[S] <ivyurl> <ivyurl>* delete ARCHIVE[S] <ivyurl> <ivyurl>*	从分布式缓存中删除使用<ivyurl>添加的资源。
reload	使HiveServer2发现配置参数指定路径下JAR文件的变更“hive.reloadable.aux.jars.path”（无需重启HiveServer2）。更改操作包括添加、删除或更新JAR文件。
dfs <dfs command>	执行dfs命令。
<query string>	执行Hive查询，并将结果打印到标准输出。

6.3.7 使用 Impala 客户端

Impala是用于处理存储在Hadoop集群中的大量数据的MPP（大规模并行处理）SQL查询引擎。 它是一个用C++和Java编写的开源软件。 与其他Hadoop的SQL引擎相比，它拥有高性能和低延迟的特点。

背景信息

假定用户开发一个应用程序，用于管理企业中的使用A业务的用户信息，使用Impala客户端实现A业务操作流程如下：

普通表的操作：

- 创建用户信息表user_info。
- 在用户信息中新增用户的学历、职称信息。
- 根据用户编号查询用户姓名和地址。
- A业务结束后，删除用户信息表。

表 6-8 用户信息

编号	姓名	性别	年龄	地址
12005000201	A	男	19	A城市
12005000202	B	女	23	B城市
12005000203	C	男	26	C城市
12005000204	D	男	18	D城市
12005000205	E	女	21	E城市
12005000206	F	男	32	F城市
12005000207	G	女	29	G城市
12005000208	H	女	30	H城市
12005000209	I	男	26	I城市
12005000210	J	女	25	J城市

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 运行Impala客户端命令，实现A业务。

直接执行Impala组件的客户端命令：

impala-shell

说明

默认情况下，**impala-shell**尝试连接到localhost的21000端口上的Impala守护程序。如需连接到其他主机，请使用**-i <host:port>**选项，例如：impala-shell -i xxx.xxx.xxx.xxx:21000。要自动连接到特定的Impala数据库，请使用**-d <database>**选项。例如，如果您的所有Kudu表都位于数据库“impala_kudu”中，则**-d impala_kudu**可以使用此数据库。要退出Impala Shell，请使用**quit**命令。

内部表的操作：

1. 根据表6-8创建用户信息表user_info并添加相关数据。

```
create table user_info(id string,name string,gender string,age int,addr string);
insert into table user_info(id,name,gender,age,addr) values("12005000201","A","男",19,"A城市");
```

.....（其他语句相同）

2. 在用户信息表user_info中新增用户的学历、职称信息。

   以增加编号为12005000201的用户的学历、职称信息为例，其他用户类似。

   ```
   alter table user_info add columns(education string,technical string);
   ```

3. 根据用户编号查询用户姓名和地址。

   以查询编号为12005000201的用户姓名和地址为例，其他用户类似。

   ```
   select name,addr from user_info where id='12005000201';
   ```

4. 删除用户信息表。

   ```
   drop table user_info;
   ```

外部分区表的操作：

创建外部分区表并导入数据

1. 创建外部表数据存储路径。
   - 安全模式（集群开启了Kerberos认证）：

     cd /opt/hadoopclient

     source bigdata_env

     kinit *hive*

     说明

     用户hive需要具有Hive管理员权限。

     impala-shell

     hdfs dfs -mkdir /hive

     hdfs dfs -mkdir /hive/user_info
   - 普通模式（集群关闭了Kerberos认证）：

     su - omm

     cd /opt/hadoopclient

     source bigdata_env

     impala-shell

     hdfs dfs -mkdir /hive

     hdfs dfs -mkdir /hive/user_info
2. 建表。

   ```
   create external table user_info(id string,name string,gender string,age int,addr string) partitioned by(year string) row format delimited fields terminated by ' ' lines terminated by '\n' stored as textfile location '/hive/user_info';
   ```

   说明

   fields terminated指明分隔的字符,如按空格分隔，' '。

   lines terminated 指明分行的字符，如按换行分隔，'\n'。

   /hive/user_info为数据文件的路径。
3. 导入数据。
   a. 使用insert语句插入数据。

      ```
      insert into user_info partition(year="2018") values ("12005000201","A","男",19,"A城市");
      ```

   b. 使用load data命令导入文件数据。

i. 根据表6-8数据创建文件。如，文件名为txt.log，以空格拆分字段，以换行符作为行分隔符。

ii. 上传文件至hdfs。

hdfs dfs -put txt.log /tmp

iii. 加载数据到表中。

load data inpath '/tmp/txt.log' into table user_info partition (year='2018');

4. 查询导入数据。

```
select * from user_info;
```

5. 删除用户信息表。

```
drop table user_info;
```

----结束

6.3.8 使用 Kafka 客户端

操作场景

用户可以在集群客户端完成Topic的创建、查询、删除等基本操作。

前提条件

已安装客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

使用 Kafka 客户端（MRS 3.x 之前版本）

步骤1 进入ZooKeeper实例页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > ZooKeeper > 实例”。
- MRS 2.0.1及至3.x之前版本，在MRS控制台单击集群名称，选择“组件管理 > ZooKeeper > 实例”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient/Kafka/kafka/bin”。

cd /opt/hadoopclient/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/hadoopclient/bigdata_env

步骤6 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Kafka用户*

步骤7 创建一个Topic：

sh kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份个数* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

例如：**sh kafka-topics.sh --create --topic TopicTest --partitions 3 --replication-factor 3 --zookeeper 10.10.10.100:2181/kafka**

步骤8 执行以下命令，查询集群中的Topic信息：

sh kafka-topics.sh --list --zookeeper *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

例如：**sh kafka-topics.sh --list --zookeeper 10.10.10.100:2181/kafka**

步骤9 删除**步骤7**中创建的Topic：

sh kafka-topics.sh --delete --topic *主题名称* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

例如：**sh kafka-topics.sh --delete --topic TopicTest --zookeeper 10.10.10.100:2181/kafka**

输入 "y"，回车。

----结束

使用 Kafka 客户端（MRS 3.x 及之后版本）

步骤1 进入ZooKeeper实例页面：

登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 实例”。

步骤2 查看ZooKeeper角色实例的IP地址。

记录ZooKeeper角色实例其中任意一个的IP地址即可。

步骤3 登录安装客户端的节点。

步骤4 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient/Kafka/kafka/bin”。

cd /opt/hadoopclient/Kafka/kafka/bin

步骤5 执行以下命令，配置环境变量。

source /opt/hadoopclient/bigdata_env

步骤6 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Kafka用户*

步骤7 登录FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 配置 > 全部配置”，搜索参数“clientPort”,记录“clientPort”的参数值。

步骤8 创建一个Topic：

sh kafka-topics.sh --create --topic *主题名称* **--partitions** *主题占用的分区数* **--replication-factor** *主题的备份个数* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

例如：**sh kafka-topics.sh --create --topic TopicTest --partitions 3 --replication-factor 3 --zookeeper 10.10.10.100:2181/kafka**

步骤9 执行以下命令，查询集群中的Topic信息：

sh kafka-topics.sh --list --zookeeper *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

例如：**sh kafka-topics.sh --list --zookeeper 10.10.10.100:2181/kafka**

步骤10 删除步骤8中创建的Topic：

sh kafka-topics.sh --delete --topic *主题名称* **--zookeeper** *ZooKeeper角色实例所在节点IP地址:clientPort*/**kafka**

例如：**sh kafka-topics.sh --delete --topic TopicTest --zookeeper 10.10.10.100:2181/kafka**

----结束

6.3.9 使用 Kudu 客户端

Kudu是专为Apache Hadoop平台开发的列式存储管理器。Kudu具有Hadoop生态系统应用程序的共同技术特性：可水平扩展，并支持高可用性操作。

前提条件

已安装集群客户端，例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。

操作步骤

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 运行Kudu命令行工具。

直接执行Kudu组件的命令行工具，查看帮助。

kudu -h

回显信息如下：

```
Usage: kudu <command> [<args>]

<command> can be one of the following:
        cluster   Operate on a Kudu cluster
       diagnose   Diagnostic tools for Kudu servers and clusters
             fs   Operate on a local Kudu filesystem
           hms    Operate on remote Hive Metastores
```

```
 local_replica   Operate on local tablet replicas via the local filesystem
        master   Operate on a Kudu Master
           pbc   Operate on PBC (protobuf container) files
          perf   Measure the performance of a Kudu cluster
remote_replica   Operate on remote tablet replicas on a Kudu Tablet Server
         table   Operate on Kudu tables
        tablet   Operate on remote Kudu tablets
          test   Various test actions
       tserver   Operate on a Kudu Tablet Server
           wal   Operate on WAL (write-ahead log) files
```

说明

kudu命令行工具不提供DDL、DML等操作，但提供针对cluster、master、tserver、fs、table等的细化查询功能。

常用操作：

- 查看当前集群下有哪些表。

  kudu table list *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051*

- 查询Kudu服务KuduMaster实例的配置信息。

  kudu master get_flags *KuduMaster实例IP:7051*

- 查询表的schema。

  kudu table describe *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051 tablename*

- 删除表。

  kudu table delete *KuduMaster实例IP1:7051, KuduMaster实例IP2:7051, KuduMaster实例IP3:7051 tablename*

  说明

  KuduMaster实例IP获取方式：在集群详情页面，选择“组件管理 > Kudu > 实例”，获取角色KuduMaster的IP地址。

----结束

6.3.10 使用 Oozie 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Oozie客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/client”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由系统管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。

使用 Oozie 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录，该操作的客户端目录只是举例，请根据实际安装目录修改。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 判断集群认证模式。

- 安全模式，执行以下命令进行用户认证。*exampleUser*为提交任务的用户名。

  kinit *exampleUser*

- 普通模式，执行**步骤5**。

步骤5 配置Hue。

1. spark2x环境配置（如果不涉及spark2x任务，可以跳过此步骤）：

   hdfs dfs -put /opt/client/Spark2x/spark/jars/*.jar /user/oozie/share/lib/spark2x/

   当HDFS目录“/user/oozie/share”中的Jar包发生变化时，需要重启Oozie服务。

2. 上传Oozie配置文件以及Jar包至HDFS：

   hdfs dfs -mkdir /user/*exampleUser*

   hdfs dfs -put -f /opt/client/Oozie/oozie-client-*/examples /user/*exampleUser*/

   说明

   - *exampleUser*为提交任务的用户名。
   - 在提交任务的用户和非job.properties文件均无变更的前提下，客户端安装目录/Oozie/oozie-client-*/examples目录一经上传HDFS，后续可重复使用，无需多次提交。
   - 解决Spark和Yarn关于jetty的jar冲突。

     hdfs dfs -rm -f /user/oozie/share/lib/spark/jetty-all-9.2.22.v20170606.jar
   - 普通模式下，上传过程如果遇到“Permission denied”的问题，可执行以下命令进行处理。

     su - omm

     source /opt/client/bigdata_env

     hdfs dfs -chmod -R 777 /user/oozie

     exit

----结束

6.3.11 使用 Storm 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Storm客户端。

前提条件

- 已安装客户端。例如安装目录为“/opt/hadoopclient”。
- 各组件业务用户由MRS集群管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。（普通模式不涉及）

操作步骤

步骤1 根据业务情况，准备好客户端。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 若安装了Storm多实例，在使用Storm命令提交拓扑时，请执行以下命令加载具体实例的环境变量，否则请跳过此步骤。例如，Storm-2实例：

source Storm-2/component_env

步骤5 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

步骤6 执行命令进行客户端操作。

例如执行以下命令：

- **cql**
- **storm**

说明

同一个storm客户端不能同时连接安全和非安全的ZooKeeper。

----结束

6.3.12 使用 Yarn 客户端

操作场景

该任务指导用户在运维场景或业务场景中使用Yarn客户端。

前提条件

- 已安装客户端。
  例如安装目录为“/opt/hadoopclient”，以下操作的客户端目录只是举例，请根据实际安装目录修改。
- 各组件业务用户由MRS集群管理员根据业务需要创建。安全模式下，“机机”用户需要下载keytab文件。“人机”用户第一次登录时需修改密码。普通模式不需要下载keytab文件及修改密码操作。

使用 Yarn 客户端

步骤1 以客户端安装用户，登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/hadoopclient

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 直接执行Yarn命令。例如：

yarn application -list

----结束

客户端常见使用问题

1. 当执行Yarn客户端命令时，客户端程序异常退出，报“java.lang.OutOfMemoryError”的错误。

   这个问题是由于Yarn客户端运行时的所需的内存超过了Yarn客户端设置的内存上限（默认为128MB）。对于MRS 3.x后续版本集群，可以通过修改“<客户端安装路径>/HDFS/component_env”中的“CLIENT_GC_OPTS”来修改Yarn客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   export CLIENT_GC_OPTS="-Xmx1G"
   ```

   对于MRS 3.x之前版本集群，可以通过修改“<客户端安装路径>/HDFS/component_env”中的“GC_OPTS_YARN”来修改Yarn客户端的内存上限。例如，需要设置该内存上限为1GB，则设置：

   ```
   export GC_OPTS_YARN="-Xmx1G"
   ```

   在修改完后，使用如下命令刷新客户端配置，使之生效：

   source <*客户端安装路径*>/**bigdata_env**

2. 如何设置Yarn客户端运行时的日志级别？

   Yarn客户端运行时的日志是默认输出到Console控制台的，其级别默认是INFO级别。有的时候为了定位问题，需要开启DEBUG级别日志，可以通过导出一个环境变量来设置，命令如下：

   export YARN_ROOT_LOGGER=DEBUG,console

   在执行完上面命令后，再执行Yarn Shell命令时，即可打印出DEBUG级别日志。

   如果想恢复INFO级别日志，可执行如下命令：

   export YARN_ROOT_LOGGER=INFO,console

7 配置存算分离

7.1 存算分离简介

MRS支持在大数据存储容量大、计算资源需要弹性扩展的场景下，用户将数据存储在OBS服务中，使用MRS集群仅作数据计算处理的存算分离模式。

说明

大数据存算分离场景，请务必使用OBS并行文件系统，使用普通对象桶会对集群性能产生较大影响。

存算分离功能使用流程：

1. 配置存算分离集群。

   请选择如下其中一种配置即可（推荐使用委托方式）。

   - 通过为MRS集群绑定ECS委托方式访问OBS，避免了AK/SK直接暴露在配置文件中的风险，具体请参考**配置存算分离集群（委托方式）**。
   - 在MRS集群中配置AK/SK，AK/SK会明文暴露在配置文件中，请谨慎使用，具体请参考**配置存算分离集群（AKSK方式）**。

2. 使用存算分离集群。

   各个组件使用存算分离的具体操作请参考如下内容。

   - **Flink对接OBS文件系统**
   - **Flume对接OBS文件系统**
   - **HDFS客户端对接OBS文件系统**
   - **Hive对接OBS文件系统**
   - **MapReduce对接OBS文件系统**
   - **Spark2x对接OBS文件系统**
   - **Sqoop对接外部存储系统**

7.2 配置存算分离集群（委托方式）

MRS支持用户将数据存储在OBS服务中，使用MRS集群仅作数据计算处理的存算模式。MRS通过IAM服务的“委托”机制进行简单配置， 实现使用ECS自动获取的临时AK/SK访问OBS。避免了AK/SK直接暴露在配置文件中的风险。

通过绑定委托，ECS或BMS云服务将有权限来管理您的部分资源，请根据实际业务场景需求确认是否需要配置委托。

MRS提供如下访问OBS的配置方式，请选择其中一种配置即可（推荐使用委托方式）：

- 通过为MRS集群绑定ECS委托方式访问OBS，避免了AK/SK直接暴露在配置文件中的风险，具体请参考本章节。
- 在MRS集群中配置AK/SK，AK/SK会明文暴露在配置文件中，请谨慎使用，具体请参考**配置存算分离集群（AKSK方式）**。

集群的Hadoop、Hive、Spark、Presto、Flink组件支持该功能。

步骤一：创建具有访问 OBS 权限的 ECS 委托

说明

- MRS在IAM的委托列表中预置了**MRS_ECS_DEFAULT_AGENCY**委托，可在集群创建过程中可以选择该委托，该委托拥有对象存储服务的OBS OperateAccess权限和在集群所在区域拥有CES FullAccess（对开启细粒度策略的用户）、CES Administrator和KMS Administrator权限。同时请勿在IAM修改**MRS_ECS_DEFAULT_AGENCY**委托。
- 如需使用预置的委托，请跳过创建委托步骤。如需使用自定义委托，请参考如下步骤进行创建委托（创建或修改委托需要用户具有Security Administrator权限）。

1. 登录管理控制台。
2. 在服务列表中选择“管理与部署 > 统一身份认证服务”。
3. 选择“委托 > 创建委托”。
4. 设置“委托名称”。例如：mrs_ecs_obs。
5. “委托类型”选择“云服务”，在“云服务”中选择“弹性云服务器ECS 裸金属服务器BMS”，授权ECS或BMS调用OBS服务。
6. “持续时间”选择“永久”。
7. 单击“权限选择”区域的修改，在弹出授权页面的搜索框内，搜索“OBS OperateAccess”策略，勾选“OBS OperateAccess”策略。
8. 单击“确定”，委托成功创建。

步骤二：创建存算分离集群

配置存算分离支持在新建集群中配置委托实现，也可以通过为已有集群绑定委托实现。本示例以开启Kerberos认证的集群为例介绍。

新创建存算分离集群：

1. 登录MRS服务控制台。
2. 单击“创建集群”，进入“创建集群”页面。
3. 在创建集群页面，选择“自定义创建”页签。
4. 在“自定义创建”页签，填写“软件配置”参数。
   - 区域：请根据需要选择。
   - 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。
   - 集群版本：请选择集群版本。
   - 集群类型：选择“分析集群”或“混合集群”并勾选所有组件。

- 元数据：选择“本地元数据”。

5. 单击“下一步”，并配置硬件相关参数。
   - 可用区：默认即可。
   - 虚拟私有云：默认即可。
   - 子网：默认即可。
   - 安全组：默认即可。
   - 弹性公网IP：默认即可。
   - 集群节点：请根据自身需求选择节点规格和数量。
6. 单击“下一步”，并配置相关参数。
   - Kerberos认证：默认开启，请根据自身需要选择。
   - 用户名：默认为“admin”，用于登录集群管理页面。
   - 密码：设置admin用户密码。
   - 确认密码：再次输入设置的admin用户密码。
   - 密钥对：从下拉框中选择密钥对用于登录ECS机器，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件*SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。
7. 本例以配置委托为例介绍，其他参数暂不配置，如需配置请参考**高级配置（可选）**。

   委托：选择**步骤一：创建具有访问OBS权限的ECS委托**所创建的委托或MRS在IAM服务中预置的委托MRS_ECS_DEFAULT_AGENCY。
8. 通信安全授权请勾选“确认授权”，详细信息请参见**授权安全通信**。
9. 单击“立即申请”。等待集群创建成功。

   当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

为已有集群配置存算分离功能：

1. 登录MRS控制台，在导航栏选择“集群列表 > 现有集群”。
2. 单击集群名称，进入集群详情页面。
3. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。
4. 在集群详情页的“概览”页签，单击委托右侧的“管理委托”选择需要绑定的委托并单击“确定”进行绑定，或单击“新建委托”进入IAM控制台进行创建后再在此处进行绑定。

步骤三：创建 OBS 文件系统用于存放数据

说明

大数据存算分离场景，请务必使用OBS并行文件系统，使用普通对象桶会对集群性能产生较大影响。

1. 登录OBS控制台。
2. 单击“并行文件系统 > 创建并行文件系统”。
3. 填写文件系统名称，例如“mrs-word001”。

   其他参数请根据需要填写。

4. 单击“立即创建”。
5. 在OBS控制台并行文件系统列表中，单击文件系统名称进入详情页面。
6. 在左侧导航栏选择“文件”，新建program、input文件夹。
   - program：请上传程序包到该文件夹。
   - input：请上传输入数据到该文件夹。

步骤四：访问 OBS 文件系统

1. 用root用户登录集群Master节点，具体请参见**登录集群节点**。
2. 配置环境变量。

   MRS 3.x之前版本请执行：**source /opt/client/bigdata_env**

   MRS 3.x及之后版本请执行：**source /opt/Bigdata/client/bigdata_env**
3. 验证Hadoop访问OBS。
   a. 查看文件系统mrs-word001下面的文件列表。

      hadoop fs -ls obs://mrs-word001/

   b. 返回文件列表即表示访问OBS成功。

      图 7-1 Hadoop 验证返回文件列表

      ```
      Found 2 items
      drwxrwxrwx   - root root          0 2019-12-21 11:04 obs://mrs-word001/input
      drwxrwxrwx   - root root          0 2019-12-21 11:04 obs://mrs-word001/program
      ```
4. 验证Hive访问OBS。
   a. 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建Hive表的权限，具体请参见**创建角色**配置拥有对应权限的角色，参考**创建用户**创建用户并为用户绑定对应角色。如果当前集群未启用Kerberos认证，则无需执行此命令。

      kinit MRS集群用户

      例如，kinit hiveuser

   b. 执行Hive组件的客户端命令。

      beeline

   c. 在beeline中直接使用obs的目录进行访问。例如，执行如下命令创建Hive表并指定数据存储在mrs-word001文件系统的test_obs目录中。

      create table test_obs(a int, b string) row format delimited fields terminated by ',' stored as textfile location "obs://mrs-word001/test_obs";

   d. 执行如下命令查询所有表，返回结果中存在表test_obs，即表示访问OBS成功。

      show tables;

      图 7-2 Hive 验证返回已创建的表名

      ```
      +-----------+
      | tab_name  |
      +-----------+
      | test_obs  |
      +-----------+
      1 row selected (0.352 seconds)
      ```

e. 使用“Ctrl + C”退出hive beeline。

5. 验证Spark访问OBS。

a. 执行Spark组件的客户端命令。

spark-beeline

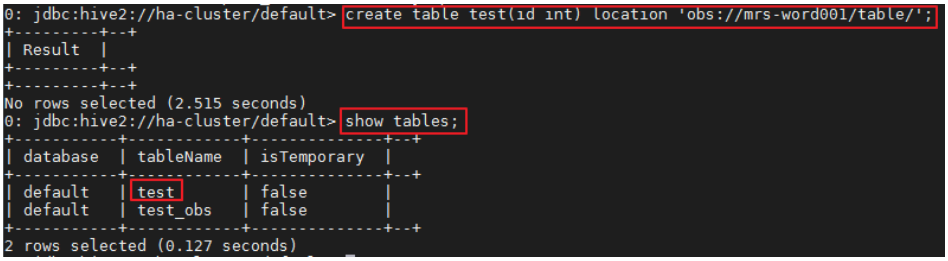
b. 在spark-beeline中访问OBS，例如在obs://mrs-word001/table/目录中创建表test。

create table test(id int) location 'obs://mrs-word001/table/';

c. 执行如下命令查询所有表，返回结果中存在表test，即表示访问OBS成功。

show tables;

图 7-3 Spark 验证返回已创建的表名



d. 使用“Ctrl + C”退出退出spark beeline。

6. 验证Presto访问OBS。

- 未开启Kerberos认证的普通集群

i. 执行如下命令连接客户端。

presto_cli.sh

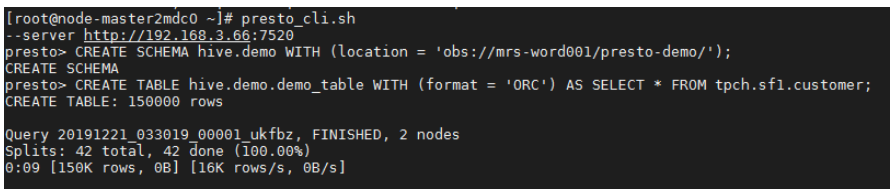
ii. 在Presto客户端中执行语句创建schema，指定location为OBS路径，例如：

CREATE SCHEMA hive.demo01 WITH (location = 'obs://mrs-word001/presto-demo002/');

iii. 在该schema中建表，该表的数据即存储在OBS文件系统内，例如：

CREATE TABLE hive.demo.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

图 7-4 普通集群 Presto 验证返回结果



iv. 执行**exit**退出客户端。

- 开启Kerberos认证的安全集群

i. 登录MRS Manager创建一个拥有“Hive Admin Privilege”权限的角色，例如prestorole，创建角色请参考**创建角色**。

ii. 创建一个属于“Presto”和“Hive”组的用户，同时为该用户绑定**6.i**中创建的角色，例如presto001，创建用户请参考**创建用户**。

iii. 认证当前用户。

kinit presto001

iv. 下载用户凭证。

1) 针对MRS 3.x之前版本集群，在MRS Manager页面，选择“系统设置 > 用户管理 ”，单击新增用户所在行的“更多 > 下载认证凭据”。

图 7-5 下载 Presto 用户认证凭据

2) 针对MRS 3.x及之后版本，在FusionInsight Manager页面，选择“系统 > 权限 > 用户”，单击新增用户所在行的“更多 > 下载认证凭据”。

v. 解压下载的用户凭证文件，得到“krb5.conf”和“user.keytab”两个文件并放入客户端目录，例如“/opt/Bigdata/client/Presto/”。

vi. 执行如下命令获取用户principal。

klist -kt /opt/Bigdata/client/Presto/user.keytab

vii. 启用Kerberos认证的集群，执行以下命令连接本集群的Presto Server。

presto_cli.sh --krb5-config-path {krb5.conf文件路径} --krb5-principal {用户principal} --krb5-keytab-path {user.keytab文件路径} --user {presto用户名}

- krb5.conf文件路径：请替换为6.v中设置的文件存放路径，例如“/opt/Bigdata/client/Presto/krb5.conf”
- user.keytab文件路径：请替换为6.v中设置的文件存放路径，例如“/opt/Bigdata/client/Presto/user.keytab”
- 用户principal：请替换为6.vi中返回的结果
- presto用户名：请替换为6.ii中创建的用户名，例如“presto001”

例如：presto_cli.sh --krb5-config-path /opt/Bigdata/client/Presto/krb5.conf --krb5-principal prest001@xxx_xxx_xxx_xxx.COM --krb5-keytab-path /opt/Bigdata/client/Presto/user.keytab --user presto001

viii. 在Presto客户端中执行语句创建schema，指定location为OBS路径，例如：

CREATE SCHEMA hive.demo01 WITH (location = 'obs://mrs-word001/presto-demo002/');

ix. 在该schema中建表，该表的数据即存储在OBS文件系统内，例如：

CREATE TABLE hive.demo01.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

图 7-6 安全集群 Presto 验证返回结果

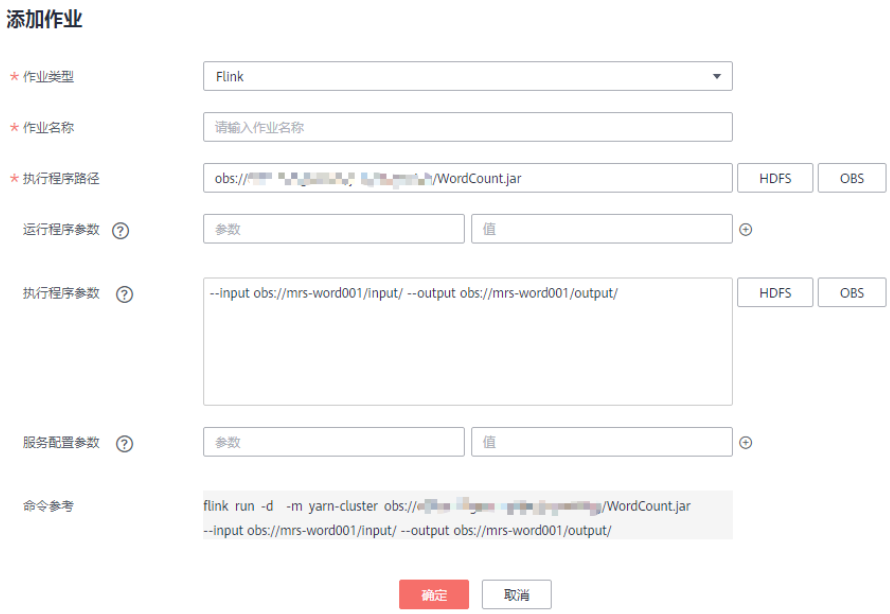
x. 执行**exit**退出客户端。

7. 验证Flink访问OBS。

a. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。

b. 用户同步完成后，在集群详情页选择“作业管理 > 添加”提交Flink作业，在“执行程序参数”中按照“--input <作业输入路径> --output <作业输出路径>”格式填写，其中作业输入路径选择OBS，输出路径请手动输入一个不存在的目录，例如obs://mrs-word001/output/，如图7-7所示。

图 7-7 添加 Flink 作业



c. 在OBS控制台，进入提交作业时选择的输出路径，即可查看到输出目录已经自动创建并存放着作业执行结果，表示访问OBS成功。

图 7-8 Flink 作业执行结果

相关参考

如需对访问OBS的权限进行控制，请参考**配置MRS多用户访问OBS细粒度权限**。

7.3 配置存算分离集群（AKSK 方式）

MRS 2.0.1及之后版本支持使用**obs://**的方式对接OBS服务，当前主要支持的组件为Hadoop、Hive、Spark、Presto、Flink。其中HBase组件使用**obs://**的方式对接OBS服务暂不支持。

MRS提供如下访问OBS的配置方式，请选择其中一种配置即可（推荐使用委托方式）：

- 通过为MRS集群绑定ECS委托方式访问OBS，避免了AK/SK直接暴露在配置文件中的风险，具体请参考**配置存算分离集群（委托方式）**。
- 在MRS集群中配置AK/SK，AK/SK会明文暴露在配置文件中，请谨慎使用，具体请参考本章节。

说明

- 为了提高数据写入性能，可以修改对应服务的配置参数fs.obs.buffer.dir的值为数据盘目录。
- 大数据存算分离场景，请务必使用OBS并行文件系统，使用普通对象桶会对集群性能产生较大影响。

Hadoop 访问 OBS

- 在MRS客户端的HDFS目录($client_home/ HDFS/hadoop/etc/hadoop)中修改core-site.xml文件，增加如下内容。

```
<property>
    <name>fs.obs.access.key</name>
    <value>ak</value>
</property>
<property>
    <name>fs.obs.secret.key</name>
    <value>sk</value>
</property>
<property>
    <name>fs.obs.endpoint</name>
    <value>obs endpoint</value>
</property>
```

须知

在文件中设置AK/SK会明文暴露在配置文件中，请谨慎使用。

添加配置后无需手动添加AK/SK、endpoint就可以直接访问OBS上的数据。例如执行如下命令查看文件系统obs-test下面的文件夹test_obs_orc的文件列表。

hadoop fs -ls "obs://obs-test/test_obs_orc"

- 每次在命令行中手动添加AK/SK、endpoint访问OBS上的数据。

  hadoop fs -Dfs.obs.endpoint=xxx -Dfs.obs.access.key=xx -Dfs.obs.secret.key=xx -ls "obs://obs-test/ test_obs_orc"

Hive 访问 OBS

步骤1 进入Hive服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 将“基础配置”切换为“全部配置”。

步骤3 搜索“fs.obs.access.key”和“fs.obs.secret.key”参数，并分别配置为OBS的AK和SK。

若当前集群中搜索不到如上两个参数，请在左侧导航选择“Hive > 自定义”，在自定义参数“core.site.customized.configs”中增加如上两个参数。

步骤4 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Hive服务。

步骤5 在beeline中直接使用obs的目录进行访问。例如，执行如下命令创建Hive表并指定数据存储在test-bucket文件系统的test_obs目录中。

create table test_obs(a int, b string) row format delimited fields terminated by "," stored as textfile location "obs://test-bucket/test_obs";

----结束

Spark 访问 OBS

说明

由于SparkSQL依赖Hive，所以在Spark上配置OBS时，需要同时修改**Hive访问OBS**的OBS配置。

- spark-beeline和spark-sql

  可以通过在shell中增加如下OBS的属性实现访问OBS。

  ```
  set fs.obs.endpoint=xxx
  set fs.obs.access.key=xxx
  set fs.obs.secret.key=xxx
  ```

- spark-beeline

  spark-beeline也可以通过在Manager中配置服务参数实现访问OBS。操作如下：

  a. 进入Spark服务配置页面：

  - MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Spark > 服务配置”。
  - MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Spark > 服务配置”。

    说明

    若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

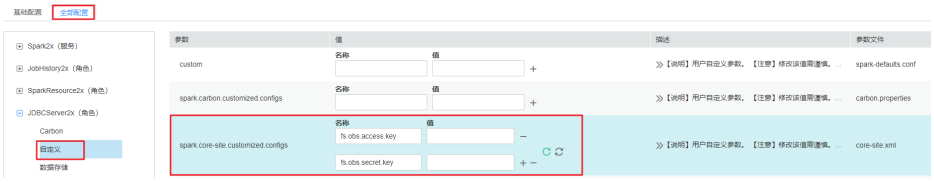
  - MRS 3.x及后续版本，登录FusionInsight Manager，具体请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。然后选择“集群 > *待操作的集群名*称 > 服务 > Spark2x > 配置”。

b. 将“基础配置”切换为“全部配置”。

c. 选择“JDBCServer > OBS”配置fs.obs.access.key、fs.obs.secret.key参数。

若当前集群中没有如上两个参数，请在左侧导航选择“JDBCServer > 自定义”，在自定义参数“spark.core-site.customized.configs”中增加如上两个参数。

图 7-9 自定义添加 OBS 参数



d. 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Spark服务。

e. 在spark-beeline中访问OBS，例如访问obs://obs-demo-input/table/目录：

create table test(id int) location 'obs://obs-demo-input/table/';

- spark-sql和spark-submit

  spark-sql也可以通过修改core-site.xml配置文件实现访问OBS。

  使用spark-sql和使用spark-submit提交任务访问OBS时，配置文件修改方法一致。

  修改MRS客户端中Spark配置文件夹（$client_home/Spark/spark/conf）中的core-site.xml，增加如下内容：

```
<property>
    <name>fs.obs.access.key</name>
    <value>ak</value>
</property>
<property>
    <name>fs.obs.secret.key</name>
    <value>sk</value>
</property>
<property>
    <name>fs.obs.endpoint</name>
    <value>obs endpoint</value>
</property>
```

Presto 访问 OBS

步骤1 登录集群详情页面，选择“组件管理 > Presto > 服务配置”。

步骤2 将“基础配置”切换为“全部配置”。

步骤3 搜索并配置如下参数。

- fs.obs.access.key配置为用户AK
- fs.obs.secret.key配置为用户SK

若当前集群中搜索不到如上两个参数，请在左侧导航选择“Presto > Hive”，在自定义参数“core.site.customized.configs”中增加如上两个参数。

步骤4 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Presto服务。

步骤5 选择“组件管理 > Hive > 服务配置”。

步骤6 将“基础配置”切换为“全部配置”。

步骤7 搜索并配置如下参数。

- fs.obs.access.key配置为用户AK
- fs.obs.secret.key配置为用户SK

步骤8 单击“保存配置”，并勾选“重新启动受影响的服务或实例。”重启Hive服务。

步骤9 在Presto客户端中执行语句创建schema，指定location为OBS路径，例如：

CREATE SCHEMA hive.demo WITH (location = 'obs://obs-demo/presto-demo/');

步骤10 在该schema中建表，该表的数据即存储在OBS文件系统内，例如：

CREATE TABLE hive.demo.demo_table WITH (format = 'ORC') AS SELECT * FROM tpch.sf1.customer;

----结束

Flink 访问 OBS

在MRS客户端的Flink配置文件“客户端安装路径/Flink/flink/conf/flink-conf.yaml”中，增加如下内容。

```
fs.obs.access.key：ak
fs.obs.secret.key: sk
fs.obs.endpoint: obs endpoint
```

须知

在文件中设置AK/SK会明文暴露在配置文件中，请谨慎使用。

添加配置后无需手动添加AK/SK、endpoint就可以直接访问OBS上的数据。

7.4 使用存算分离集群

7.4.1 Flink 对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 使用安装客户端的用户登录Flink客户端安装节点。

步骤2 执行如下命令初始化环境变量。

source ${client_home}/bigdata_env

步骤3 需要配置好Flink客户端。具体配置参考**安装客户端（3.x及之后版本）**。

步骤4 如果是安全集群，使用以下命令进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

步骤5 在Flink命令行显式添加要访问的OBS文件系统。

./bin/flink run --class com.xxx.bigdata.flink.examples.FlinkProcessingTimeAPIMain ./config/FlinkCheckpointJavaExample.jar --chkPath obs://*OBS并行文件系统名称*

----结束

说明

由于Flink作业是On Yarn运行，在配置Flink对接OBS文件系统之前需要确保Yarn对接OBS文件系统功能是正常的。

7.4.2 Flume 对接 OBS 文件系统

本章节适用于MRS 3.x及之后的版本。

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 配置委托。

1. 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”。
2. 单击集群名称，进入集群详情页面。
3. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步。
4. 单击委托右侧的“管理委托”，选择需要绑定的委托并单击“确定”进行绑定。

步骤2 创建OBS文件系统用于存放数据。

1. 登录OBS控制台。
2. 单击“并行文件系统”进入并行文件系统页面，单击“创建并行文件系统”。
3. 填写文件系统名称，例如“esdk-c-test-pfs1”，其他参数请根据需要填写。单击“立即创建”等待创建完成。
4. 在OBS控制台并行文件系统列表中，单击已新建的文件系统名称进入详情页面。
5. 在左侧导航栏选择“文件 > 新建文件夹”新建“testFlumeOutput”文件夹。

步骤3 准备properties.properties文件并将上传至“/opt/flumeInput”目录。

1. 在本地准备“properties.properties”文件，文件内容如下：

```
# source
server.sources = r1
# channels
server.channels = c1
# sink
server.sinks = obs_sink
# ----- define net source -----
server.sources.r1.type = seq
server.sources.r1.spooldir = /opt/flumeInput
# ---- define OBS sink ----
server.sinks.obs_sink.type = hdfs
server.sinks.obs_sink.hdfs.path = obs://esdk-c-test-pfs1/testFlumeOutput
server.sinks.obs_sink.hdfs.filePrefix = %[localhost]
server.sinks.obs_sink.hdfs.useLocalTimeStamp = true
# set file size to trigger roll
server.sinks.obs_sink.hdfs.rollSize = 0
server.sinks.obs_sink.hdfs.rollCount = 0
server.sinks.obs_sink.hdfs.rollInterval = 5
#server.sinks.obs_sink.hdfs.threadsPoolSize = 30
server.sinks.obs_sink.hdfs.fileType = DataStream
server.sinks.obs_sink.hdfs.writeFormat = Text
server.sinks.obs_sink.hdfs.fileCloseByEndEvent = false
```

```
# define channel
server.channels.c1.type = memory
server.channels.c1.capacity = 1000
# transaction size
server.channels.c1.transactionCapacity = 1000
server.channels.c1.byteCapacity = 800000
server.channels.c1.byteCapacityBufferPercentage = 20
server.channels.c1.keep-alive = 60
server.sources.r1.channels = c1
server.sinks.obs_sink.channel = c1
```

说明

参数“server.sinks.obs_sink.hdfs.path”中的值为**步骤2**中新建的OBS文件系统。

2. 使用**root**用户登录安装Flume客户端的节点。
3. 新建“/opt/flumeInput”目录，并在该目录下新建一个内容自定义的txt文件。
4. 登录FusionInsight Manager。
5. 选择“集群 > *待操作集群的名称* > 服务 > Flume > 配置”，在参数“flume.config.file”的“值”中单击“上传文件”，上传**步骤3.1**准备的“properties.properties”文件，单击“保存”。

步骤4 在OBS系统中查看结果。

1. 登录OBS控制台。
2. 单击“并行文件系统”，进入**步骤2**中创建的并行文件系统中的文件夹查看结果。

----结束

7.4.3 HDFS 客户端对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 以客户端安装用户登录安装了HDFS客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd *客户端安装目录*

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 在hdfs命令行显式添加要访问的OBS文件系统。

例如：

- 使用以下命令访问OBS文件系统。
  hdfs dfs -ls obs://*OBS并行文件系统名称|路径*
- 使用以下命令上传客户端节点“/opt/test.txt”文件到OBS文件系统路径下。
  hdfs dfs -put /opt/test.txt obs://*OBS并行文件系统名称/路径*

----结束

说明

OBS文件系统打印大量日志可能导致读写性能受影响，可通过调整OBS客户端日志级别优化，日志调整方式如下：

cd *客户端安装目录*/**HDFS/hadoop/etc/hadoop**

vi log4j.properties

在文件中添加OBS日志级别配置

log4j.logger.org.apache.hadoop.fs.obs=WARN

log4j.logger.com.obs=WARN

```
[root@node-master1AuKK hadoop]# tail -4 log4j.properties
log4j.logger.org.apache.commons.beanutils=WARN

log4j.logger.org.apache.hadoop.fs.obs=WARN
log4j.logger.com.obs=WARN
[root@node-master1AuKK hadoop]#
```

7.4.4 Hive 对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

建表时指定 Location 为 OBS 路径

步骤1 使用安装客户端用户登录客户端安装节点。

步骤2 执行如下命令初始化环境变量。

source *客户端安装目录*/**bigdata_env**

步骤3 如果是安全集群，执行以下命令进行用户认证（该用户需要具有Hive操作的权限），如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *Hive组件操作用户*

步骤4 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 配置 > 全部配置”。

在左侧的导航列表中选择“Hive > 自定义”。在自定义配置项中，给参数“hdfs.site.customized.configs”添加配置项“dfs.namenode.acls.enabled”，设置值为“false”。

步骤5 单击“保存”，保存配置。单击“概览”，选择“更多 > 重启服务”，输入当前用户密码，单击“确定”，并勾选“同时重启上层服务。”，单击“确定”，重启Hive服务。

步骤6 进入beeline客户端，在创建表时指定Location为OBS文件系统路径。

beeline

例如，创建一个表“test”，该表的Location为“obs://OBS并行文件系统名称/user/hive/warehouse/数据库名/表名”：

create table test(name string) location "obs://*OBS并行文件系统名称/user/hive/warehouse/数据库名/表名***";**

说明

需要添加组件操作用户到Ranger策略中的URL策略，URL填写对象在obs上的完整路径。权限选择Read, Write 权限，其他权限不涉及URL策略。

----结束

指定创建的 Hive 表默认 Location 为 OBS 路径

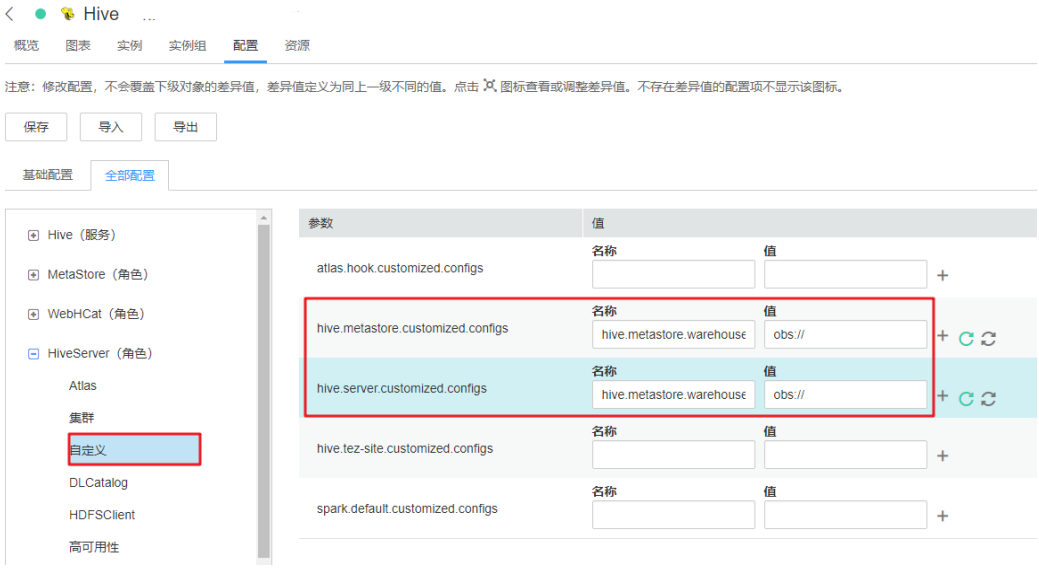
步骤1 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 配置 > 全部配置”。

步骤2 在左侧的导航列表中选择“MetaStore > 自定义”。在自定义配置项中，给参数“hive.metastore.customized.configs”添加配置项“hive.metastore.warehouse.dir”，设置值为OBS路径。

图 7-10 hive.metastore.warehouse.dir 配置

步骤3 在左侧的导航列表中选择“HiveServer > 自定义”。在自定义配置项中，给参数“hive.metastore.customized.configs”和“hive.server.customized.configs”添加配置项“hive.metastore.warehouse.dir”，设置值为OBS路径。

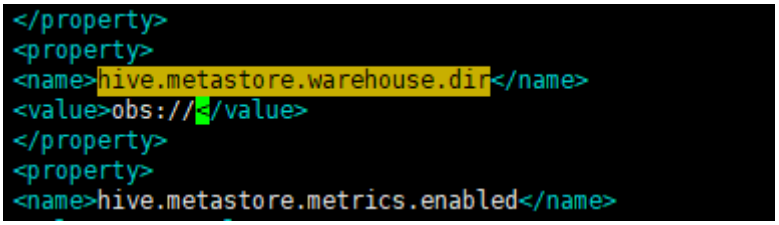
图 7-11 hive.metastore.warehouse.dir 配置



步骤4 保存并重启Hive服务。

步骤5 更新客户端配置文件。

1. 执行以下命令修改客户端Hive配置文件目录下的“hivemetastore-site.xml”。
   vim /opt/Bigdata/client/Hive/config/hivemetastore-site.xml
2. 将“hive.metastore.warehouse.dir”的值修改为对应的OBS路径。



步骤6 进入beeline客户端，创建表并确认Location为OBS路径。

beeline

create table test(name string);

desc formatted test;

说明

如果当前数据库Location已指向HDFS，那么在当前数据库下建表（不指定Location）默认也指向当前HDFS。如需修改默认建表策略可以修改数据库的Location重新指向OBS。操作如下：

1. 执行以下命令查看数据库Location。

   show create database *obs_test*;

   ```
   +----------------------------------------------------+
   |                    createdb_stmt                   |
   +----------------------------------------------------+
   | CREATE DATABASE `obs_test`                         |
   | LOCATION                                           |
   |   'hdfs://hacluster/user/hive/warehouse/obs_test.db'  |
   +----------------------------------------------------+
   3 rows selected (0.038 seconds)
   ```

2. 执行以下命令修改数据库Location。

   alter database *obs_test* **set location** '**obs://***OBS并行文件系统名称/user/hive/warehouse/数据库名*'

   执行命令**show create database** *obs_test*，查看数据库Location已经指向OBS。

   ```
   INFO  : Concurrency mode is disabled, not creating
   +----------------------------+
   |        createdb_stmt       |
   +----------------------------+
   | CREATE DATABASE `obs_test` |
   | LOCATION                   |
   |   'obs://test1231/'        |
   +----------------------------+
   3 rows selected (0.063 seconds)
   ```

3. 执行以下命令修改表的Location。

   alter table *user_info* **set location** '**obs://***OBS并行文件系统名称/user/hive/warehouse/数据库名/表名*'

   如果表已有业务数据，需要同步迁移原数据文件至修改后的Location地址。

----结束

7.4.5 MapReduce 对接 OBS 文件系统

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

步骤1 登录MRS管理控制台，单击集群名称进入集群详情页面。

步骤2 选择“组件管理 > Mapreduce”，进入Mapreduce服务“全部配置”页面，在左侧的导航列表中选择“Mapreduce > 自定义”。在自定义配置项中，给参数文件“core-site.xml”添加配置项“mapreduce.jobhistory.always-scan-user-dir”，设置值为“true”。

步骤3 保存配置，并重启Mapreduce服务。

----结束

7.4.6 Spark2x 对接 OBS 文件系统

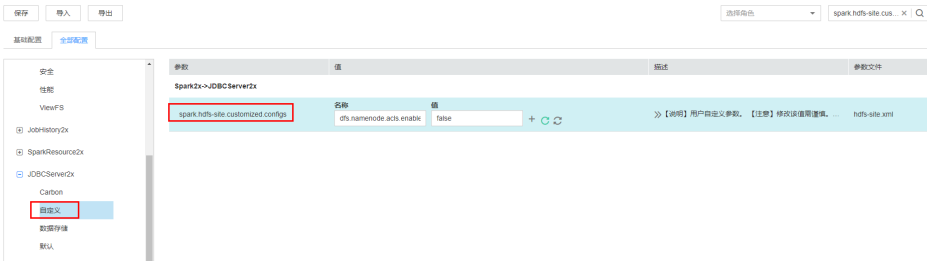
MRS集群支持Spark2x在集群安装完成后对接OBS文件系统。

使用本章节前已参考**配置存算分离集群（委托方式）**或**配置存算分离集群（AKSK方式）**完成存算分离集群配置。

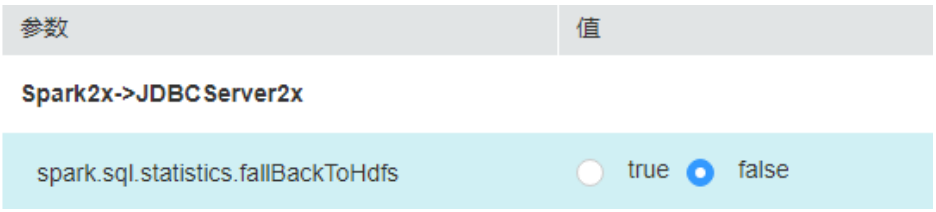
集群安装后使用 spark beeline

步骤1 登录FusionInsight Manager，选择“集群 > 服务 > Spark2x > 配置 > 全部配置”。

在左侧的导航列表中选择“JDBCServer2x > 自定义”。在参数“spark.hdfs-site.customized.configs”中添加配置项“dfs.namenode.acls.enabled”，值为“false”。



步骤2 在搜索框中搜索参数“spark.sql.statistics.fallBackToHdfs”，修改该参数值为“false”。



步骤3 保存配置并重启JDBCServer2x实例。

步骤4 使用安装客户端用户登录客户端安装节点。

步骤5 配置环境变量。

source *客户端安装目录*/**bigdata_env**

步骤6 如果是安全集群，使用以下命令用户进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

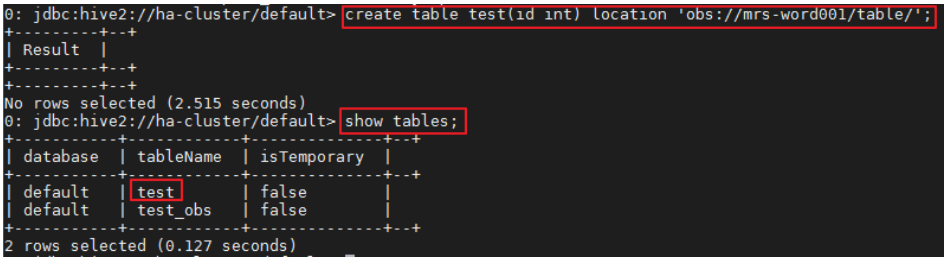
步骤7 在spark-beeline中访问OBS，例如在“obs://mrs-word001/table/”目录中创建表“test”。

create table test(id int) location '*obs://mrs-word001/table/*';

步骤8 执行如下命令查询所有表，返回结果中存在表test，即表示访问OBS成功。

show tables;

图 7-12 Spark2x 验证返回已创建的表名



步骤9 使用“Ctrl + C”退出spark beeline。

----结束

集群安装后使用 spark sql

步骤1 使用安装客户端用户登录客户端安装节点。

步骤2 配置环境变量。

source *客户端安装目录*/**bigdata_env**

步骤3 修改配置文件：

vim *客户端安装目录*/**Spark2x/spark/conf/hdfs-site.xml**

```
<property>
<name>dfs.namenode.acls.enabled</name>
<value>false</value>
</property>
```

步骤4 如果是安全集群，使用以下命令用户进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

步骤5 在spark-sql中访问OBS，例如在“obs://mrs-word001/table/”目录中创建表“test”。

步骤6 进入spark bin目录：**cd** *客户端安装目录*/**Spark2x/spark/bin**，执行./**spark-sql**登录spark-sql命令行。

步骤7 在spark-sql命令行执行以下命令：

create table test(id int) location '*obs://mrs-word001/table/*';

步骤8 执行语句**show tables;**查看表是否存在。

步骤9 执行**exit;**退出spark-sql命令行。

说明

OBS文件系统打印大量日志可能导致读写性能受影响，可通过调整OBS客户端日志级别优化，日志调整方式如下：

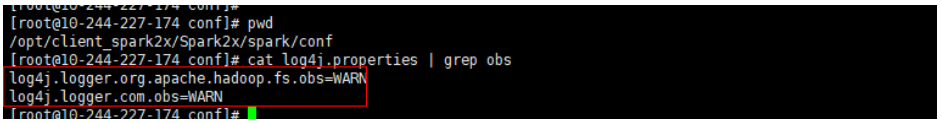
cd *客户端安装目录*/**Spark2x/spark/conf**

vi log4j.properties

在文件中添加OBS日志级别配置

log4j.logger.org.apache.hadoop.fs.obs=WARN

log4j.logger.com.obs=WARN



----结束

7.4.7 Sqoop 对接外部存储系统

sqoop export (HDFS 到 MySQL)

步骤1 登录客户端所在节点。

步骤2 执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤3 使用sqoop命令操作sqoop客户端。

sqoop export --connect jdbc:mysql://10.100.231.134:3306/test --username root --password xxxxxx --table component13 -export-dir hdfs://hacluster/user/hive/warehouse/component_test3 --fields-terminated-by ',' -m 1

表 7-1 参数说明

参数	说明
-direct	快速模式，利用了数据库的导入工具，如MySQL的mysqlimport，可以比jdbc连接的方式更为高效的将数据导入到关系数据库中。
-export-dir <dir>	存放数据的HDFS的源目录。
-m或-num-mappers <n>	启动n个map来并行导入数据，默认是4个，该值请勿高于集群的最大Map数。
-table <table-name>	要导入的目的关系数据库表。
-update-key <col-name>	后面接条件列名，通过该参数可以将关系数据库中已经存在的数据进行更新操作，类似于关系数据库中的update操作。

参数	说明
-update-mode <mode>	更新模式，有两个值updateonly和默认的allowinsert，该参数只能在关系数据表里不存在要导入的记录时才能使用，比如要导入的hdfs中有一条id=1的记录，如果在表里已经有一条记录id=2，那么更新会失败。
-input-null-string <null-string>	可选参数，如果没有指定，则字符串null将被使用。
-input-null-non-string <null-string>	可选参数，如果没有指定，则字符串null将被使用。
-staging-table <staging-table-name>	创建一个与导入目标表同样数据结构的表，将所有数据先存放在该表中，然后由该表通过一次事务将结果写入到目标表中。 该参数是用来保证在数据导入关系数据库表的过程中的事务安全性，因为在导入的过程中可能会有多个事务，那么一个事务失败会影响到其它事务，比如导入的数据会出现错误或出现重复的记录等等情况，那么通过该参数可以避免这种情况。
-clear-staging-table	如果该staging-table非空，则通过该参数可以在运行导入前清除staging-table里的数据。

----结束

sqoop import（MySQL 到 Hive 表）

步骤1 登录客户端所在节点。

步骤2 执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤3 使用sqoop命令操作sqoop客户端。

sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password xxxxxx --table component --hive-import --hive-table component_test2 --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile

表 7-2 参数说明

参数	说明
-append	将数据追加到hdfs中已经存在的dataset中。使用该参数，sqoop将把数据先导入到一个临时目录中，然后重新给文件命名到一个正式的目录中，以避免和该目录中已存在的文件重名。
-as-avrodatafile	将数据导入到一个Avro数据文件中。
-as-sequencefile	将数据导入到一个sequence文件中。

参数	说明
-as-textfile	将数据导入到一个普通文本文件中，生成该文本文件后，可以在hive中通过sql语句查询出结果。
-boundary-query <statement>	边界查询，在导入前先通过SQL查询得到一个结果集，然后导入的数据就是该结果集内的数据，格式如：**-boundary-query 'select id,creationdate from person where id = 3'**，表示导入的数据为id=3的记录，或者**select min(<split-by>), max(<split-by>) from <table name>**。 注意：查询的字段中不能有数据类型为字符串的字段，否则会报错：java.sql.SQLException: Invalid value for getLong()。
-columns<col,col,col…>	指定要导入的字段值，格式如：-columns id,username
-direct	快速模式，利用了数据库的导入工具，如MySQL的mysqlimport，可以比jdbc连接的方式更为高效的将数据导入到关系数据库中。
-direct-split-size	在使用上面direct直接导入的基础上，对导入的流按字节数分块，特别是使用直连模式从PostgreSQL导入数据时，可以将一个到达设定大小的文件分为几个独立的文件。
-inline-lob-limit	设定大对象数据类型的最大值。
-m或-num-mappers	启动n个map来并行导入数据，默认是4个，该值请勿高于集群的最大Map数。
-query，-e<statement>	从查询结果中导入数据，该参数使用时必须指定-target-dir、-hive-table，在查询语句中一定要有where条件且在where条件中需要包含$CONDITIONS。 示例：-query 'select * from person where $CONDITIONS ' -target-dir /user/hive/warehouse/person -hive-table person
-split-by<column-name>	表的列名，用来切分工作单元，一般后面跟主键ID。
-table <table-name>	关系数据库表名，数据从该表中获取。
-target-dir <dir>	指定hdfs路径。
-warehouse-dir <dir>	与**-target-dir**不能同时使用，指定数据导入的存放目录，适用于导入hdfs，不适合导入hive目录。
-where	从关系数据库导入数据时的查询条件，示例：-where 'id = 2'
-z,-compress	压缩参数，默认数据不压缩，通过该参数可以使用gzip压缩算法对数据进行压缩，适用于SequenceFile，text文本文件，和Avro文件。
–compression-codec	Hadoop压缩编码，默认为gzip。

参数	说明
–null-string <null-string>	替换null字符串，如果没有指定，则字符串null将被使用。
–null-non-string<null-string>	替换非String的null字符串，如果没有指定，则字符串null将被使用。
-check-column (col)	增量导入参数，用来作为判断的列名，如id。
-incremental (mode) append 或lastmodified	增量导入参数。 append：追加，比如对大于last-value指定的值之后的记录进行追加导入。 lastmodified：最后的修改时间，追加last-value指定的日期之后的记录。
-last-value (value)	增量导入参数，指定自从上次导入后列的最大值（大于该指定的值），也可以自己设定某一值。

----结束

Sqoop 使用样例

- sqoop import（MySQL到HDFS）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--query 'SELECT * FROM component where $CONDITIONS and component_id ="MRS 1.0_002"' --target-dir /tmp/component_test --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile**

- sqoop export （obs到MySQL）

  sqoop export --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--table component14 -export-dir obs://obs-file-bucket/xx/part-m-00000 --fields-terminated-by ',' -m 1**

- sqoop import（MySQL到obs）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--table component --target-dir obs://obs-file-bucket/xx --delete-target-dir --fields-terminated-by "," -m 1 --as-textfile**

- sqoop import（MySQL到Hive外obs表）

  sqoop import --connect jdbc:mysql://10.100.231.134:3306/test --username root --password *xxx* **--table component --hive-import --hive-table component_test01 --fields-terminated-by "," -m 1 --as-textfile**

7.4.8 Hudi 对接 OBS 文件系统

步骤1 使用安装客户端用户登录客户端安装节点。

步骤2 配置环境变量。

source *客户端安装目录*/**bigdata_env**

source *客户端安装目录*/**Hudi/component_env**

步骤3 修改配置文件：

vim *客户端安装目录*/**Hudi/hudi/conf/hdfs-site.xml**

```
<property>
<name>dfs.namenode.acls.enabled</name>
<value>false</value>
</property>
```

步骤4 如果是安全集群，使用以下命令用户进行用户认证，如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *用户名*

步骤5 启动spark-shell，执行下面的命令创建COW表存储到OBS中：

```
import org.apache.hudi.QuickstartUtils._
import scala.collection.JavaConversions._
import org.apache.spark.sql.SaveMode._
import org.apache.hudi.DataSourceReadOptions._
import org.apache.hudi.DataSourceWriteOptions._
import org.apache.hudi.config.HoodieWriteConfig._
val tableName = "hudi_cow_table"
val basePath = "obs://testhudi/cow_table/"
val dataGen = new DataGenerator
val inserts = convertToStringList(dataGen.generateInserts(10))
val df = spark.read.json(spark.sparkContext.parallelize(inserts, 2))
df.write.format("org.apache.hudi").
options(getQuickstartWriteConfigs).
option(PRECOMBINE_FIELD_OPT_KEY, "ts").
option(RECORDKEY_FIELD_OPT_KEY, "uuid").
option(PARTITIONPATH_FIELD_OPT_KEY, "partitionpath").
option(TABLE_NAME, tableName).
mode(Overwrite).
save(basePath);
```

步骤6 使用datasource查看表建立成功，数据正常。

```
val roViewDF = spark.
read.
format("org.apache.hudi").
load(basePath + "/*/*/*/*")
roViewDF.createOrReplaceTempView("hudi_ro_table")
```

spark.sql("select * from hudi_ro_table").show()

步骤7 执行**:q**退出spark-shell命令行。

----结束

8 访问 MRS 集群上托管的开源组件 Web 页面

8.1 开源组件 Web 站点

场景介绍

MRS集群默认在集群的Master节点或Core节点创建并托管了不同组件的Web站点，用户可以通过这些Web站点查看组件相关信息。

访问开源组件Web站点步骤：

1. 配置访问方式。

   MRS提供如下访问开源组件Web站点的方式：

   - **通过弹性公网IP访问**：推荐使用该方式，为集群绑定弹性公网IP，简便易操作。
   - **通过Windows弹性云服务器访问**：需要创建单独的ECS并进行相关配置。
   - **创建连接MRS集群的SSH隧道并配置浏览器**：当用户和MRS集群处于不同的网络中时可以使用该方式访问。

2. 访问站点。请参考的地址进行访问。

Web 站点一览

说明

对于开启Kerberos认证的集群，admin用户不具备各组件的管理权限，如需正常访问各组件的Web UI界面，请提前创建具有对应组件管理权限的用户。

表 8-1 开源组件 Web 站点地址

集群类型	站点类型	站点地址
全部类型	MRS Manager	• 适用于所有版本集群 https://Manager浮动IP地址:28443/web **说明** 1. 确保本地机器与MRS集群网络互通。 2. 远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示MRS Manager浮动IP地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到MRS Manager的浮动IP地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。 • MRS 3.x之前版本集群 https://<弹性公网IP>:9022/mrsmanager?locale=zh-cn 具体请参见**访问MRS Manager（MRS 2.x及之前版本）**。 • MRS 3.x及以后版本请参见**访问FusionInsight Manager（MRS 3.x及之后版本）**。
分析集群	HDFS NameNode	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > HDFS > NameNode WebUI > NameNode (主)” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > HDFS > NameNode WebUI > NameNode (*主机名称*，主)”
	HBase HMaster	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > HBase > HMaster WebUI > HMaster (主)” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > HBase > HMaster WebUI > HMaster (*主机名称*，主)”

集群类型	站点类型	站点地址
	MapReduce JobHistoryServer	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Mapreduce > JobHistoryServer WebUI > JobHistoryServer” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Mapreduce > JobHistoryServer WebUI > JobHistoryServer (*主机名称*，主)”
	YARN ResourceManager	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Yarn > ResourceManager WebUI > ResourceManager (主)” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Yarn > ResourceManager WebUI > ResourceManager (*主机名称*，主)”
	Spark JobHistory	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Spark > Spark WebUI > JobHistory” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Spark2x > Spark2x WebUI > JobHistory2x (*主机名称*)”
	Hue	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Hue > Hue WebUI > Hue (主)” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Hue > Hue WebUI > Hue (*主机名称*，主)” Loader页面是基于开放源代码Sqoop WebUI的图形化数据迁移管理工具，由Hue WebUI承载。
	Tez	• MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Tez > Tez WebUI > TezUI” • MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Tez > Tez WebUI > TezUI (*主机名称*)”

集群类型	站点类型	站点地址
	Presto	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Presto > Presto WebUI > Coordinator (主)” ● 在Manager页面选择“集群 > 服务 > Presto > Coordinator WebUI > Coordinator(Coordinator)”
	Ranger	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Ranger > Ranger WebUI > RangerAdmin (主)” ● MRS 3.x及以后版本集群，在Manager页面选择“集群 > 服务 > Ranger > Ranger WebUI > RangerAdmin”
流处理集群	Storm	● MRS 3.x之前版本集群，在集群详情页选择“组件管理 > Storm > Storm WebUI > UI” ● 在Manager页面选择“集群 > 服务 > Storm > Storm WebUI > UI (*主机名称*)”

8.2 开源组件端口列表

HBase 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.6.3以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.6.3以后版本)	端口说明
hbase.master.port	21300	16000	HMaster RPC端口。该端口用于HBase客户端连接到HMaster。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.6.3以后版本)	端口说明
hbase.master.info.port	21301	16010	HMaster HTTPS端口。该端口用于远程Web客户端连接到HMaster UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.regionserver.port	21302	16020	RS (RegoinServer) RPC端口。该端口用于HBase客户端连接到RegionServer。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.regionserver.info.port	21303	16030	Region server HTTPS端口。该端口用于远程Web客户端连接到RegionServer UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.thrift.info.port	21304	9095	Thrift Server的Thrift Server侦听端口。 该端口用于： 客户端链接时使用该端口侦听。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.6.3以后版本)	端口说明
hbase.regionserver.thrift.port	21305	9090	RegionServer的Thrift Server侦听端口 。 该端口用于： 客户端链接RegionServer时使用该端口侦听。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hbase.rest.info.port	21308	8085	RegionServer RESTServer原生Web界面的端口
-	21309	21309	RegionServer RESTServer的REST端口

HDFS 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
dfs.namenode.rpc.port	25000	● 9820（MRS 3.x之前版本） ● 8020（MRS 3.x及之后版本）	NameNode RPC 端口。 该端口用于： 1. HDFS客户端与Namenode间的通信。 2. Datanode与NameNode之间的连接。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
dfs.namenode.http.port	25002	9870	HDFS HTTP端口(NameNode)。 该端口用于: 1. 点对点的NameNode检查点操作。 2. 远程Web客户端连接NameNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.namenode.https.port	25003	9871	HDFS HTTPS端口(NameNode)。 该端口用于: 1. 点对点的NameNode检查点操作。 2. 远程Web客户端连接NameNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.ipc.port	25008	9867	Datanode IPC 服务器端口。 该端口用于: 客户端连接DataNode用来执行RPC操作。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
dfs.datanode.port	25009	9866	Datanode数据传输端口。 该端口用于： 1. HDFS客户端从DataNode传输数据或传输数据到DataNode。 2. 点对点的Datanode传输数据。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.http.port	25010	9864	Datanode HTTP端口。 该端口用于： 安全模式下，远程Web客户端连接DataNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.datanode.https.port	25011	9865	Datanode HTTPS端口。 该端口用于： 安全模式下，远程Web客户端连接DataNode UI。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.JournalNode.rpc.port	25012	8485	JournalNode RPC端口。 该端口用于： 客户端通信用于访问多种信息。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
dfs.journalnode.http.port	25013	8480	JournalNode HTTP端口。 该端口用于： 安全模式下，远程Web客户端链接JournalNode。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
dfs.journalnode.https.port	25014	8481	JournalNode HTTPS端口。 该端口用于： 安全模式下，远程Web客户端链接JournalNode。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
httpfs.http.port	25018	14000	HttpFS HTTP服务器侦听的端口。 该端口用于： 远程REST接口连接HttpFS。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Hive 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
templeton.port	21055	9111	WebHCat提供REST服务的端口。 该端口用于： WebHCat客户端与WebHCat服务端之间的通信。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hive.server2.thrift.port	21066	10000	HiveServer提供Thrift服务的端口。 该端口用于： HiveServer客户端与HiveServer之间的通信。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hive.metastore.port	21088	9083	MetaStore提供Thrift服务的端口。 该端口用于： MetaStore客户端与MetaStore之间的通信，即HiveServer与MetaStore之间通信。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
hive.server2.webui.port	-	10002	Hive的Web UI端口。 该端口用Web请求与Hive UI服务器进行HTTPS/HTTP通信。 该端口在MRS 1.9.x及之后版本支持。

Hue 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
HTTP_PORT	21200	8888	Hue提供HTTPS服务端口。 该端口用于：HTTPS方式提供Web服务，支持修改。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Kafka 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
port	21005	9092	Broker提供数据接收、获取服务
ssl.port	21008	9093	Broker提供数据接收、获取服务的SSL端口
sasl.port	21007	21007	Broker提供SASL安全认证端口，提供安全Kafka服务
sasl-ssl.port	21009	21009	Broker提供SASL安全认证和SSL通信的端口，提供安全认证及通信加密服务

Loader 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
LOADER_HTTPS_PORT	21351	21351	该端口用于提供Loader作业配置、运行的REST接口 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Manager 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后、3.x之前版本)	端口说明
-	8080	8080	WebService提供的供用户访问端口 该端口用于使用HTTP协议访问Web UI ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
-	28443	28443	WebService提供的供用户访问端口 该端口用于使用https协议访问Web UI ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

MapReduce 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
mapreduce.jobhistory.webapp.port	26012	19888	Job history服务器Web http端口。 该端口用于：查看Job History服务器的Web页面。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
mapreduce.jobhistory.port	26013	10020	Job history服务器端口。 该端口用于： 1. 用于MapReduce客户端恢复任务的数据。 2. 用于Job客户端获取任务报告。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
mapreduce.jobhistory.webapp.https.port	26014	19890	Job history服务器Web https端口。 该端口用于查看Job History服务器的Web页面。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Spark 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
hive.server2.thrift.port	22550	22550	JDBC thrift端口。 该端口用于： Spark2.1.0 CLI/JDBC与Spark2.1.0 CLI/JDBC服务器进行socket通信。 **说明** 如果hive.server2.thrift.port被占用，将抛端口被占用异常。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
spark.ui.port	22950	4040	JDBC的Web UI端口 该端口用于：Web请求与JDBC Server Web UI服务器进行HTTPS/HTTP通信。 **说明** 系统会根据端口的设置取值，并验证其有效性；如果无效，端口+1，直到取到有效值为止（上限16次，重试次数可以通过配置spark.port.maxRetries改变）。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
spark.history.ui.port	22500	18080	JobHistory Web UI端口 该端口用于：Web请求与Spark2.1.0 History Server间的HTTPS/HTTP通信 **说明** 系统会根据端口的设置取值，并验证其有效性；如果无效，端口+1，直到取到有效值为止（上限16次，重试次数可以通过配置spark.port.maxRetries改变）。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Storm 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
nimbus.thrift.port	29200	6627	nimbus提供thrift服务
supervisor.slots.ports	29200-29499	6700,6701,6702,6703	接收由其它服务器转发过来的请求
logviewer.https.port	29248	29248	logviewer提供HTTPS服务
ui.https.port	29243	29243	Storm UI提供Https服务(ui.https.port)

YARN 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
yarn.resourcemanager.webapp.port	26000	8088	ResourceManager服务的Web http端口。
yarn.resourcemanager.webapp.https.port	26001	8090	ResourceManager服务的Web https端口。 该端口用于：安全模式下，接入Resource Manager Web应用。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是
yarn.nodemanager.webapp.port	26006	8042	NodeManager Web http端口
yarn.nodemanager.webapp.https.port	26010	8044	NodeManager Web https端口。 该端口用于： 安全模式下，接入NodeManager web应用。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

ZooKeeper 常用端口

表中涉及端口的协议类型均为：TCP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
clientPort	24002	2181	ZooKeeper客户端端口。 该端口用于： ZooKeeper客户端连接ZooKeeper服务器。 **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Kerberos 常用端口

表中涉及端口的协议类型均为：UDP。(MRS 1.7.0及以后版本)

配置参数	默认端口 (MRS 1.6.3及之前版本)	默认端口 (MRS 1.7.0及以后版本)	端口说明
kdc_ports	21732	21732	Kerberos服务端端口 该端口用于： 组件向Kerberos服务认证。配置集群互信可能会用到； **说明** 端口的取值范围为一个建议值，由产品自己指定。在代码中未做端口范围限制。 ● 安装时是否缺省启用：是 ● 安全加固后是否启用：是

Opentsdb 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
tsd.network.port	4242	Opentsdb的WEB UI端口。 该端口用于：Web请求与Opentsdb UI服务器进行HTTPS/HTTP通信。

Tez 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
tez.ui.port	28888	Tez的WEB UI端口。

KafkaManager 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
kafka_manager_port	9099	KafkaManager的WEB UI端口。

Presto 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
http-server.http.port	7520	presto coordinator对外提供服务的HTTP端口。
http-server.https.port	7521	presto coordinator对外提供服务的HTTPS端口。
http-server.http.port	7530	presto worker对外提供服务的HTTP端口。
http-server.https.port	7531	presto worker对外提供服务的HTTPS端口。

Flink 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
jobmanager.web.port	32261-32325	Flink的WEB UI端口。 用于Client Web请求与Flink server进行HTTP/HTTPS通信。

ClickHouse 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
tcp_port	9000	业务客户端TCP接入端口。
http_port	8123	业务客户端HTTP接入端口。
https_port	8443	业务客户端HTTPS接入端口。
tcp_port_secure	9440	业务客户端TCP With SSL接入端口。默认仅在安全模式下开放。

Impala 常用端口

表中涉及端口的协议类型均为：TCP。

配置参数	默认端口	端口说明
--beeswax_port	21000	提供给impala-shell通信的端口。
--hs2_port	21050	提供给Impala应用通信的端口。
--hs2_http_port	28000	Impala对外提供HiveServer2协议的端口。

8.3 通过专线访问

MRS为您提供云专线（Direct Connect）方式访问MRS集群。云专线用于搭建用户本地数据中心与线上云VPC之间高速、低时延、稳定安全的专属连接通道，充分利用线上云服务优势的同时，继续使用现有的IT设施，实现灵活一体，可伸缩的混合云计算环境。

前提条件

云专线服务可用，并已打通本地数据中心到线上VPC的连接通道。

通过专线访问 MRS 集群

步骤1 登录MRS管理控制台。

步骤2 单击集群名称进入集群详情页。

步骤3 在集群详情页面的“概览”页签，单击“集群管理页面”右侧的“前往 Manager”。

步骤4 “访问方式”选择“专线访问”，并勾选“我确认已打通本地与浮动IP的网络，可使用专线直接访问MRS Manager。”。

浮动IP为MRS为您访问MRS Manager页面自动分配的IP地址，使用专线访问MRS Manager之前您确保云专线服务已打通本地数据中心到线上VPC的连接通道。

步骤5 单击“确定”，进入MRS Manager登录页面，用户名使用“admin”，密码为创建集群时设置的admin密码。

----结束

切换 MRS Manager 访问方式

为了便于用户操作，浏览器缓存会记录用户所选择的访问Manager的方式，如需切换访问Manager方式，参考如下步骤操作。

步骤1 登录MRS管理控制台。

步骤2 单击集群名称进入集群详情页。

步骤3 在集群详情页面的“概览”页签，单击“集群管理页面”右侧的 按钮。

步骤4 在弹出页面重新选择“访问方式”即可。

- 若由“EIP访问”切换为“专线访问”，请在专线网络互通的前提下，在弹出页面的“访问方式”选择“专线访问”并勾选“我确认已打通本地与浮动IP的网络，可使用专线直接访问MRS Manager。”后单击“确定”。
- 若由“专线访问”切换为“EIP访问”，在弹出页面的“访问方式”选择“EIP访问”并参考**通过弹性公网IP访问Manager**配置EIP。若集群已配置过公网IP，直接单击“确定”以EIP方式访问Manager。

----结束

8.4 通过弹性公网 IP 访问

为了方便用户访问开源组件的Web站点，MRS集群支持通过为集群绑定弹性公网IP的方式，访问MRS集群上托管的开源组件。该方式更加简便易操作，推荐使用该方式访问开源组件的Web站点。

为集群绑定弹性公网 IP 并添加安全组规则

1. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步，待IAM用户同步成功后，在集群详情页会出现“组件管理”页签。
2. 单击“集群管理页面”右侧的“前往 Manager”。
3. 弹出访问MRS Manager页面，绑定弹性公网IP并添加安全组规则。仅首次访问该集群的组件开源站点时，需要如下配置。
   a. 绑定弹性公网IP，在弹性公网IP下拉框中选择可用的弹性公网IP。若没有可用的弹性公网IP，请单击“管理弹性公网IP”创建弹性公网IP后在该页面引用。若创建集群时已绑定弹性公网IP，请跳过该步骤。
   b. 选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
   c. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问MRS组件。

d. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

e. 单击“确定”，进入登录页面，用户名使用“admin”，密码为创建集群时设置的admin密码。

4. 登录Manager页面，选择“集群 > 服务 > HDFS > NameNode WebUI > NameNode (*主机名称*，主)”，访问开源组件Web站点。此处仅以HDFS NameNode为例介绍，其他组件访问地址请参考**开源组件Web站点**页面提供的站点地址。

8.5 通过 Windows 弹性云服务器访问

MRS支持通过Windows弹性云服务器访问开源组件Web站点。该方式操作较为复杂，推荐不支持EIP功能的MRS集群使用。

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在 “现有集群” 列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“集群控制台地址”、“安全组”。

说明

集群控制台地址获取方式：远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示集群控制台地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到集群控制台地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。

步骤3 在ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个企业镜像“Enterprise_Windows_STD_2012R2_20170316-0(80GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和MRS集群的“安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为MRS集群的安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组中添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的账号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

Manager访问地址形式为**https://*集群控制台地址*:28443/web**。访问时需要输入MRS集群的用户名和密码，例如“admin”用户。

说明

- 集群控制台地址：远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示集群控制台地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到集群控制台地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。
- 如果使用其他MRS集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

步骤7 请参考**开源组件Web站点**页面提供的站点地址访问开源组件Web站点。

----结束

相关任务

配置集群节点名称与IP地址映射

步骤1 登录Manager，单击“主机管理”。

记录集群中所有节点的“主机名称”和“管理IP”。

步骤2 在工作环境使用“记事本”打开“hosts”文件，将节点名称与IP地址的对应关系填写到文件中。

每个对应关系填写一行，填写效果例如：

```
192.168.4.127 node-core-Jh3ER
192.168.4.225 node-master2-PaWVE
192.168.4.19 node-core-mtZ81
192.168.4.33 node-master1-zbYN8
192.168.4.233 node-core-7KoGY
```

保存修改。

----结束

8.6 创建连接 MRS 集群的 SSH 隧道并配置浏览器

操作场景

用户和MRS集群处于不同的网络中，需要创建一个SSH隧道连接，使用户访问站点的数据请求，可以发送到MRS集群并动态转发到对应的站点。

MAC系统暂不支持该功能访问MRS，请参考**通过弹性公网IP访问**内容访问MRS。

前提条件

- 准备一个SSH客户端用于创建SSH隧道，例如使用开源SSH客户端Git。请下载并安装。
- 已创建好集群，并准备pem格式的密钥文件或创建集群时的密码。
- 用户本地环境可以访问互联网。

操作步骤

步骤1 登录MRS管理控制台，选择“集群列表 > 现有集群”。

步骤2 单击指定名称的MRS集群。

记录集群的“安全组”。

步骤3 为集群Master节点的安全组添加一条需要访问MRS集群的IP地址的入规则，允许指定来源的数据访问端口“22”。

具体请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 查询集群的主管理节点，具体请参考**如何确认Manager的主备管理节点**。

步骤5 为集群的主管理节点绑定一个弹性IP地址。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤6 在本地启动Git Bash，执行以下命令登录集群的主管理节点：**ssh root@*弹性IP地址***或者**ssh -i *密钥文件路径* root@*弹性IP地址***

步骤7 执行以下命令查看数据转发配置：

cat /etc/sysctl.conf | grep net.ipv4.ip_forward

- 系统查询到“net.ipv4.ip_forward=1”表示已配置转发，则请执行**步骤9**。
- 系统查询到“net.ipv4.ip_forward=0”表示未配置转发，则请执行**步骤8**。
- 系统查询不到“net.ipv4.ip_forward”参数表示该参数未配置，则请执行以下命令后再执行**步骤9**。

  echo "net.ipv4.ip_forward = 1" >> /etc/sysctl.conf

步骤8 修改节点转发配置：

1. 执行以下命令切换**root**用户：

   sudo su - root
2. 执行以下命令，修改转发配置：

   echo 1 > /proc/sys/net/ipv4/ip_forward

   sed -i "s/net.ipv4.ip_forward=0/net.ipv4.ip_forward = 1/g" /etc/sysctl.conf

   sysctl -w net.ipv4.ip_forward=1
3. 执行以下命令，修改sshd配置文件：

   vi /etc/ssh/sshd_config

   按I进入编辑模式，查找“AllowTcpForwarding”和“GatewayPorts”，并删除注释符号，修改内容如下，然后保存并退出：

   ```
   AllowTcpForwarding yes
   GatewayPorts yes
   ```

4. 执行以下命令，重启sshd服务：

   service sshd restart

步骤9 执行以下命令查看浮动IP地址：

ifconfig

系统显示的“eth0:FI_HUE”表示为Hue的浮动IP地址，“eth0:wsom”表示Manager浮动IP地址，请记录“inet”的实际参数值。

然后退出登录：**exit**

步骤10 在本地机器执行以下命令创建支持动态端口转发的SSH隧道：

使用命令**ssh -i *密钥文件路径* -v -ND *本地端口地址* root@*弹性IP地址***或者***ssh -v -ND 本地端口地址 root@弹性IP地址***，然后输入创建集群时的密码。

其中，“本地端口地址”需要指定一个用户本地环境未被使用的端口，建议选择8157。

创建后的SSH隧道，通过“-D”启用动态端口转发功能。默认情况下，动态端口转发功能将启动一个SOCKS代理进程并侦听用户本地端口，端口的数据将由SSH隧道转发到集群的主管理节点。

步骤11 执行如下命令配置浏览器代理。

1. 进入本地Google Chrome浏览器客户端安装目录。
2. 按住“shift+鼠标右键”，选择“在此处打开命令窗口”，打开CMD窗口后输入如下命令：

   chrome --proxy-server="socks5://localhost:8157" --host-resolver-rules="MAP * 0.0.0.0 , EXCLUDE localhost" --user-data-dir=c:/tmppath --proxy-bypass-list="*google*com,*gstatic.com,*gvt*.com,*:80"

   说明

   - 8157为**步骤10**中配置的本地代理端口。
   - 若本地操作系统为Windows 10，请打开Windows操作系统“开始”菜单，输入**cmd**命令，打开一个命令行窗口执行**步骤11.2**中的命令。若该方式不能成功，请打开Windows操作系统“开始”菜单后，在搜索框中输入并执行**步骤11.2**中的命令。

步骤12 在新弹出的浏览器地址栏，输入Manager的访问地址。

Manager访问地址形式为**https://*Manager浮动IP地址*:28443/web**。

访问启用Kerberos认证的集群时，需要输入MRS集群的用户名和密码，例如“admin”用户。未启用Kerberos认证的集群则不需要。

第一次访问时，请根据浏览器提示，添加站点信任以继续打开页面。

步骤13 准备站点的访问地址。

1. 参考**Web站点一览**，获取Web站点的地址格式及对应的角色实例。
2. 单击“服务管理”。
3. 单击指定的服务名称，例如HDFS。
4. 单击“实例”，查看NameNode的主角色实例“NameNode(主)”的“业务IP”。

步骤14 在浏览器输入访问Web站点真实地址并访问。

步骤15 退出访问Web站点时，请终止并关闭SSH隧道。

----结束

9 Jupyter Notebook 采用自定义 Python 对接 MRS

9.1 简介

在MRS服务中可以配合Jupyter Notebook使用Pyspark，能够提高机器学习、数据探索和ETL应用开发效率。

本文指导用户如何在MRS中配置Jupyter Notebook来使用Pyspark。具体流程如下：

1. **安装集群外节点客户端**
2. **安装Python3**
3. **配置MRS客户端**
4. **安装Jupyter Notebook**
5. **验证Jupyter Notebook访问MRS**

说明

本操作仅适用MRS 3.x及之后版本。

9.2 安装集群外节点客户端

步骤1 准备一台集群外Linux弹性云服务器，需要满足的条件，请参考**在集群外节点安装客户端前提条件**。

步骤2 参考**集群外节点安装客户端**，安装集群外节点客户端，例如安装目录为“/opt/client”。

步骤3 确认集群是否开启了Kerberos认证。

- 是，执行**步骤4**。
- 否，执行**安装Python3**。

步骤4 **访问FusionInsight Manager（MRS 3.x及之后版本）**，登录到Manager界面。

步骤5 创建用户，用户组选择hadoop，主组选择hadoop，角色选择Manager_operator，例如创建用户mrs-test。

步骤6 使用root用户，登录到安装的客户端节点，执行如下命令，配置环境变量，安全认证。

source /opt/client/bigdata_env

kinit *mrs-test*

说明

第一次认证需要修改用户密码。

----结束

9.3 安装 Python3

步骤1 使用root用户，登录集群外客户端节点，执行如下命令，检查是否安装Python3。

python3 --version

```
[root@ecs-notebook FusionInsight_Cluster_1_Services_ClientConfig]# python3 --version
-bash: python3: command not found
```

- 是，执行**配置MRS客户端**。
- 否，执行**步骤2**。

步骤2 安装Python，此处以Python3.6.6为例。

1. 执行如下命令，安装依赖：

   yum install zlib zlib-devel zip -y

   yum install gcc-c++

   yum install openssl-devel

   yum install sqlite-devel -y

   如果pandas库需要额外安装如下依赖：

   yum install -y xz-devel

   yum install bzip2-devel

2. 下载对应Python版本源码：**wget https://www.python.org/ftp/python/3.6.6/Python-3.6.6.tgz**
3. 执行如下命令，解压python源码压缩包，例如下载在“opt”目录下：

   cd /opt

   tar –xvf Python-3.6.6.tgz

4. 创建Python的安装目录，此处以“/opt/python36”为例：

   mkdir /opt/python36

5. 编译Python。

   cd /opt/python-3.6.6

   ./configure --prefix=/opt/python36

   执行成功，显示结果如下：

```
configure: creating ./config.status
config.status: creating Makefile.pre
config.status: creating Modules/Setup.config
config.status: creating Misc/python.pc
config.status: creating Misc/python-config.sh
config.status: creating Modules/ld_so_aix
config.status: creating pyconfig.h
creating Modules/Setup
creating Modules/Setup.local
creating Makefile


If you want a release build with all stable optimizations active (PGO, etc),
please run ./configure --enable-optimizations
```

   执行命令：**make -j8**，执行成功，显示结果如下：

```
creating build/scripts-3.6
copying and adjusting /tmp/python366/Python-3.6.6/Tools/scripts/pydoc3 -> build/scripts-3.6
copying and adjusting /tmp/python366/Python-3.6.6/Tools/scripts/idle3 -> build/scripts-3.6
copying and adjusting /tmp/python366/Python-3.6.6/Tools/scripts/2to3 -> build/scripts-3.6
copying and adjusting /tmp/python366/Python-3.6.6/Tools/scripts/pyvenv -> build/scripts-3.6
changing mode of build/scripts-3.6/pydoc3 from 644 to 755
changing mode of build/scripts-3.6/idle3 from 644 to 755
changing mode of build/scripts-3.6/2to3 from 644 to 755
changing mode of build/scripts-3.6/pyvenv from 644 to 755
renaming build/scripts-3.6/pydoc3 to build/scripts-3.6/pydoc3.6
renaming build/scripts-3.6/idle3 to build/scripts-3.6/idle3.6
renaming build/scripts-3.6/2to3 to build/scripts-3.6/2to3-3.6
renaming build/scripts-3.6/pyvenv to build/scripts-3.6/pyvenv-3.6
```

   执行命令：**make install**，执行成功，显示结果如下：

```
rm -f /opt/python36/share/man/man1/python3.1
(cd /opt/python36/share/man/man1; ln -s python3.6.1 python3.1)
if test "xupgrade" != "xno"  ; then \
        case upgrade in \
                upgrade) ensurepip="--upgrade" ;; \
                install|*) ensurepip="" ;; \
        esac; \
         ./python -E -m ensurepip \
                $ensurepip --root=/ ; \
fi
Looking in links: /tmp/tmp6ldv525m
Collecting setuptools
Collecting pip
Installing collected packages: setuptools, pip
Successfully installed pip-10.0.1 setuptools-39.0.1
```

6. 执行如下命令，配置python环境。

   export PYTHON_HOME=/opt/python36

   export PATH=$PYTHON_HOME/bin:$PATH

7. 执行命令**python3 --version**，显示结果如下，表示Python已经安装完成。

```
[root@ecs-notebook Python-3.6.6]# python3 --version
Python 3.6.6
```

步骤3 验证Python3。

pip3 install helloword

python3

import helloworld

helloworld.say_hello("test")

```
[root@ecs-notebook Python-3.6.6]# pip3 install helloword
Collecting helloword
  Downloading https://files.pythonhosted.org/packages/1b/bf/f0f69f122158e0e98b5d95987a7ef5add3f8a340c6eb78d5871f855ca04e/helloword-0.0.1-py3-none-any.whl
Installing collected packages: helloword
Successfully installed helloword-0.0.1
You are using pip version 10.0.1, however version 21.3.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
[root@ecs-notebook Python-3.6.6]# python3
Python 3.6.6 (default, Dec 15 2021, 06:12:40)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import helloworld
helloworld.say_hello("test")Hello, Sara!
>>>
'Hello, test!'
>>>
```

步骤4 测试安装第三方Python库（如pandas, sklearn）。

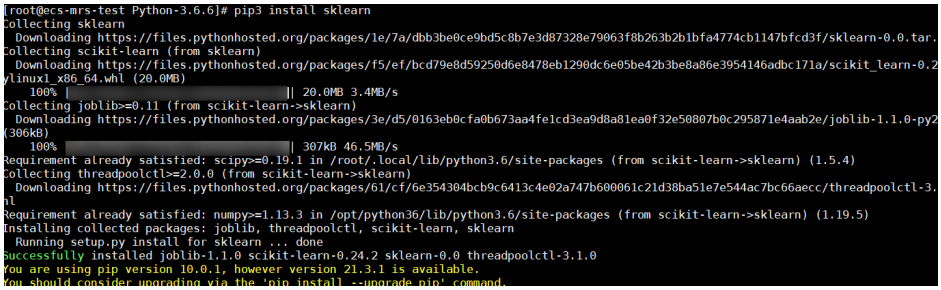
pip3 install pandas

```
[root@ecs-mrs-test Python-3.6.6]# pip3 install pandas
Collecting pandas
  Downloading https://files.pythonhosted.org/packages/c3/e2/00cacecafbab071c787019f00ad84ca3185952f6bb9bca9550ed83870d4d/pandas-1.1.5-cp36-cp36m-man
_x86_64.whl (9.5MB)
    100% |                                | 9.5MB 6.5MB/s
Collecting pytz>=2017.2 (from pandas)
  Downloading https://files.pythonhosted.org/packages/60/2e/dec1cc18c51b8df33c7c4d0a321b084cf38e1733b98f9d15018880fb4970/pytz-2022.1-py2.py3-none-an
503kB)
    100% |                                | 512kB 47.2MB/s
Collecting python-dateutil>=2.7.3 (from pandas)
  Downloading https://files.pythonhosted.org/packages/36/7a/87837f39d0296e723bb9b62bbb257d0355c7f6128853c78955f57342a56d/python_dateutil-2.8.2-py2.p
-any.whl (247kB)
    100% |                                | 256kB 54.5MB/s
Collecting numpy>=1.15.4 (from pandas)
  Downloading https://files.pythonhosted.org/packages/45/b2/6c7545bb7a38754d63048c7696804a0d947328125d81bf12beaa692c3ae3/numpy-1.19.5-cp36-cp36m-man
_x86_64.whl (13.4MB)
    100% |                                | 13.4MB 4.2MB/s
Collecting six>=1.5 (from python-dateutil>=2.7.3->pandas)
  Downloading https://files.pythonhosted.org/packages/d9/5a/e7c31adbe875f2abbb91bd84cf2dc52d792b5a01506781dbcf25c91daf11/six-1.16.0-py2.py3-none-any
Installing collected packages: pytz, six, python-dateutil, numpy, pandas
Successfully installed numpy-1.19.5 pandas-1.1.5 python-dateutil-2.8.2 pytz-2022.1 six-1.16.0
You are using pip version 10.0.1, however version 21.3.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

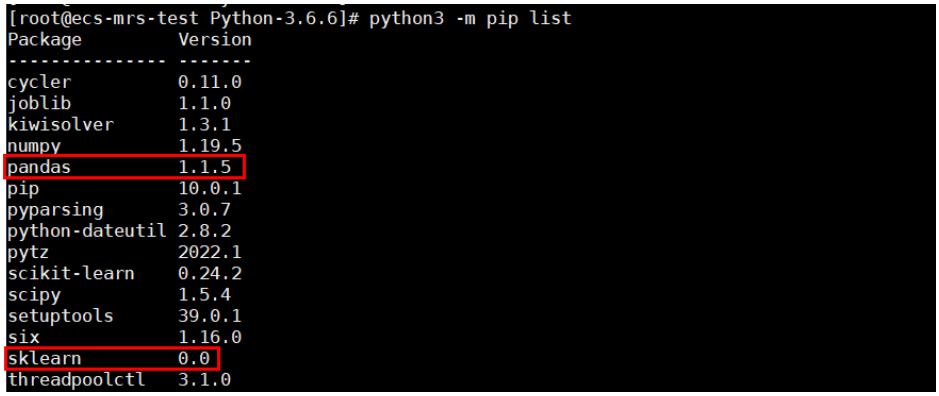
pip3 install backports.lzma

```
[root@ecs-mrs-test Python-3.6.6]# pip3 install backports.lzma
Collecting backports.lzma
  Using cached https://files.pythonhosted.org/packages/21/0f/1a9990233076d48aa2084100ba289ca162975e73a688f3a56c0ee2bb441a/backports.lzma-0.0.14.tar.gz
Installing collected packages: backports.lzma
  Running setup.py install for backports.lzma ... done
Successfully installed backports.lzma-0.0.14
You are using pip version 10.0.1, however version 21.3.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

pip3 install sklearn



步骤5 执行命令**python3 -m pip list**，查看安装结果。



步骤6 打包Python.zip

cd /opt/python36/

zip -r python36.zip ./*

步骤7 创建并上传到HDFS指定目录作为后续使用。

hdfs dfs -mkdir /user/python

hdfs dfs -put python36.zip /user/python

----结束

9.4 配置 MRS 客户端

1. 登录Spark客户端节点，进入Spark客户端安装目录“/opt/client/Spark2x/spark/conf”

   cd /opt/client/Spark2x/spark/conf

2. 在spark-defaults.conf配置文件如下参数：

   ```
   spark.pyspark.driver.python=/usr/bin/python3
   spark.yarn.dist.archives=hdfs://hacluster/user/python/python36.zip#Python
   ```

9.5 安装 Jupyter Notebook

步骤1 使用root用户登录客户端节点，执行如下命令安装Jupyter Notebook。

pip3 install jupyter notebook

显示结果如下，表示安装成功：

步骤2 为了安全，我们需要生成一个密文密码用于登录Jupyter，放到Jupyter Notebook的配置文件中。

执行如下命令，需要输入两次密码：（进行到Out[3]退出）

ipython

```
[root@ecs-notebook python36]# ipython
Python 3.6.6 (default, Dec 20 2021, 09:32:25)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.16.2 -- An enhanced Interactive Python. Type '?' for help.
In [1]: from notebook.auth import passwd
In [2]: passwd()
Enter password:
Verify password:
Out[2]: 'argon2:$argon2id$v=19$m=10240,t=10,p=8$g14BqLddl927n/unsyPlLQ
$YmoKJzbUfNG7LcxylJzm90bgbKWUIiHy6ZV+ObTzdcA
```

步骤3 执行如下命令生成jupyter配置文件：

jupyter notebook --generate-config

步骤4 修改配置文件：

vi ~/.jupyter/jupyter_notebook_config.py

添加如下配置：

```
# -*- coding: utf-8 -*-
c.NotebookApp.ip='*' #此处填写ecs对应的内网IP
c.NotebookApp.password = u'argon2:$argon2id$v=19$m=10240,t=10,p=8$NmoAVwd8F6vFP2rX5ZbV7w
$SyueJoC0a5TbCuHYzqfSx1vQcFvOTTryR+0uk2MNNZA' # 填写步骤2，Out[2]密码生成的密文
c.NotebookApp.open_browser = False # 禁止自动打开浏览器
c.NotebookApp.port = 9999 # 指定端口号
c.NotebookApp.allow_remote_access = True
```

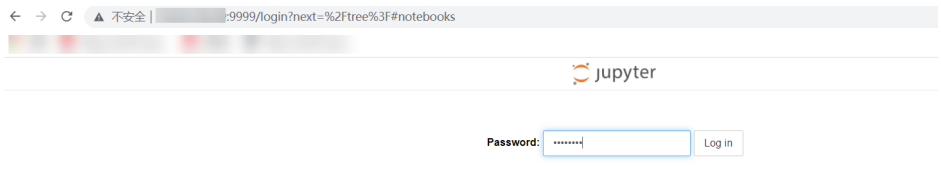
----结束

9.6 验证 Jupyter Notebook 访问 MRS

步骤1 在客户端节点执行如下命令，启动Jupyter Notebook：

PYSPARK_PYTHON=./Python/bin/python3 PYSPARK_DRIVER_PYTHON=jupyter-notebook PYSPARK_DRIVER_PYTHON_OPTS="--allow-root" pyspark --master yarn --executor-memory 2G --driver-memory 1G

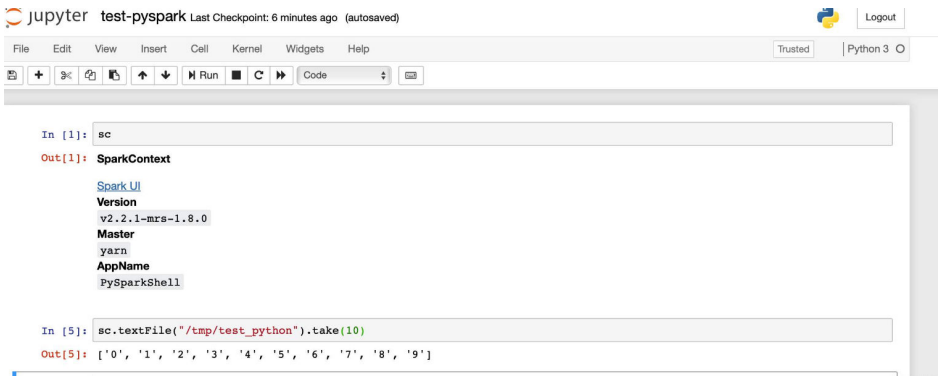
步骤2 使用*弹性IP*:9999，登录到jupyter webui（保证ECS的安全组对外放通本地公网IP和9999端口），登录密码为步骤2设置的密码。



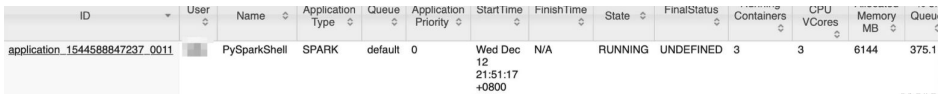
步骤3 创建代码。

创建个新的python3任务，使用Spark读取文件。

结果如下：

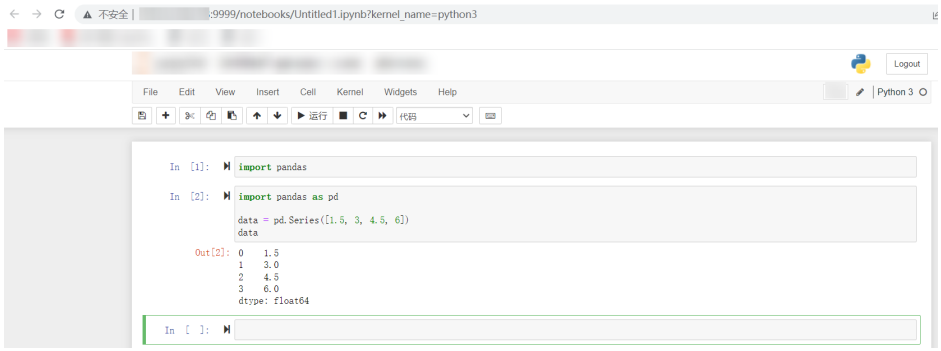


登录到Manager界面，在Yarn的WebUI页面上查看提交的pyspark应用：



ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Containers	CPU VCores	Memory MB	Queu
application_1544588847237_0011	[redacted]	PySparkShell	SPARK	default	0	Wed Dec 12 21:51:17 +0800	N/A	RUNNING	UNDEFINED	3	3	6144	375.1

步骤4 验证pandas库调用。

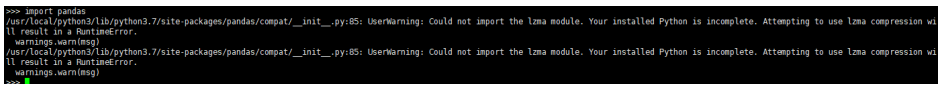


----结束

9.7 常见问题

问题

pandas本地import使用时，有如下告警：



处理步骤

步骤1 执行命令**python -m pip install backports.lzma**安装lzma模块，如下图所示：

```
[root@master ~]# python -m pip install backports.lzma
Looking in indexes: http://mirrors.aliyun.com/pypi/simple/
Requirement already satisfied: backports.lzma in /usr/local/python3/lib/python3.7/site-packages (0.0.14)
You are using pip version 10.0.1, however version 19.3.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

步骤2 进入“/usr/local/python3/lib/python3.6”目录（机器不同，目录也有所不同，可以通过**which**命令来查找当前运行python是使用的那个目录的），然后编辑lzma.py文件。

将：

```
from _lzma import *
from _lzma import _encode_filter_properties, _decode_filter_properties
```

更改为：

```
try:
    from _lzma import *
    from _lzma import _encode_filter_properties, _decode_filter_properties
except ImportError:
    from backports.lzma import *
    from backports.lzma import _encode_filter_properties, _decode_filter_properties
```

修改前：

```
"""Interface to the liblzma compression library.

This module provides a class for reading and writing compressed files,
classes for incremental (de)compression, and convenience functions for
one-shot (de)compression.

These classes and functions support both the XZ and legacy LZMA
container formats, as well as raw compressed data streams.
"""

__all__ = [
    "CHECK_NONE", "CHECK_CRC32", "CHECK_CRC64", "CHECK_SHA256",
    "CHECK_ID_MAX", "CHECK_UNKNOWN",
    "FILTER_LZMA1", "FILTER_LZMA2", "FILTER_DELTA", "FILTER_X86", "FILTER_IA64",
    "FILTER_ARM", "FILTER_ARMTHUMB", "FILTER_POWERPC", "FILTER_SPARC",
    "FORMAT_AUTO", "FORMAT_XZ", "FORMAT_ALONE", "FORMAT_RAW",
    "MF_HC3", "MF_HC4", "MF_BT2", "MF_BT3", "MF_BT4",
    "MODE_FAST", "MODE_NORMAL", "PRESET_DEFAULT", "PRESET_EXTREME",

    "LZMACompressor", "LZMADecompressor", "LZMAFile", "LZMAError",
    "open", "compress", "decompress", "is_check_supported",
]

import builtins
import io
import os
from _lzma import *
from _lzma import _encode_filter_properties, _decode_filter_properties
import _compression
```

修改后：

```
These classes and functions support both the XZ and legacy LZMA
container formats, as well as raw compressed data streams.
"""

__all__ = [
    "CHECK_NONE", "CHECK_CRC32", "CHECK_CRC64", "CHECK_SHA256",
    "CHECK_ID_MAX", "CHECK_UNKNOWN",
    "FILTER_LZMA1", "FILTER_LZMA2", "FILTER_DELTA", "FILTER_X86", "FILTER_IA64",
    "FILTER_ARM", "FILTER_ARMTHUMB", "FILTER_POWERPC", "FILTER_SPARC",
    "FORMAT_AUTO", "FORMAT_XZ", "FORMAT_ALONE", "FORMAT_RAW",
    "MF_HC3", "MF_HC4", "MF_BT2", "MF_BT3", "MF_BT4",
    "MODE_FAST", "MODE_NORMAL", "PRESET_DEFAULT", "PRESET_EXTREME",

    "LZMACompressor", "LZMADecompressor", "LZMAFile", "LZMAError",
    "open", "compress", "decompress", "is_check_supported",
]

import builtins
import io
import os
#from _lzma import *
#from _lzma import _encode_filter_properties, _decode_filter_properties
try:
    from _lzma import *
    from _lzma import _encode_filter_properties, _decode_filter_properties
except ImportError:
    from backports.lzma import *
    from backports.lzma import _encode_filter_properties, _decode_filter_properties
import _compression
```

步骤3 保存退出，然后再次import，如下图所示：

```
[root@master python3.7]# python
Python 3.7.0 (default, Oct 26 2019, 01:19:22)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-36)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas
>>>
```

----结束

10 访问集群 Manager

10.1 访问 FusionInsight Manager（MRS 3.x 及之后版本）

操作场景

MRS 3.x及之后版本的集群使用FusionInsight Manager对集群进行监控、配置和管理。用户在集群安装后可使用账号登录FusionInsight Manager。

说明

如果不能正常登录组件的WebUI页面，请参考**通过ECS访问FusionInsight Manager**方式访问FusionInsight Manager。

通过弹性 IP 访问 FusionInsight Manager

步骤1 登录MRS管理控制台页面。

步骤2 单击“集群列表 > 现有集群”，在集群列表中单击指定的集群名称，进入集群信息页面。

步骤3 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中配置弹性IP信息。

1. 若创建MRS集群时暂未绑定弹性公网IP，在“弹性公网IP”下拉框中选择可用的弹性公网IP。若用户创建集群时已经绑定弹性公网IP，直接执行**步骤3.2**。

   说明

   - 如果没有弹性公网IP，可先单击“管理弹性公网IP”创建弹性公网IP后，然后在弹性公网IP下拉框中选择创建的弹性公网IP。
   - 如果在使用完后需要解绑或释放弹性公网IP，请登录“弹性公网IP”界面，在待操作的弹性公网IP后，单击“操作”列的“解绑”或“更多 > 释放”。
   - 如果已创建弹性公网IP，但在绑定时无法找到，可能是由于该弹性公网IP被其他集群绑定，请先在弹性公网IP界面解绑，然后再为当前集群绑定。

2. 在“安全组”中选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
3. 添加安全组规则，默认填充的是用户访问弹性IP地址的规则，如需开放多个IP段为可信范围用于访问Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

4. 勾选确认信息后，单击“确定”。

步骤4 单击“确定”，进入Manager登录页面。

步骤5 输入默认用户名“admin”及创建集群时设置的密码，单击“登录”进入Manager页面。

步骤6 在MRS管理控制台，在“现有集群” 列表，单击指定的集群名称，进入集群信息页面。

说明

如需给其他用户开通访问Manager的权限，请执行**步骤6**～**步骤9**，添加对应用户访问公网的IP地址为可信范围。

步骤7 单击弹性公网IP后边的“添加安全组规则”。

步骤8 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的*公网IP*/*端口*号是可信任的公网访问IP范围，我了解使用0.0.0.0/0会带来安全风险”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤6**-**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤9 单击“确定”完成安全组规则添加。

----结束

通过 ECS 访问 FusionInsight Manager

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在“现有集群”列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“集群管理页面”、“安全组”。

步骤3 在管理控制台首页服务列表中选择“弹性云服务器”，进入ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个标准镜像“Windows Server 2012 R2 Standard 64bit(40GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和Master节点的“默认安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为Master节点的默认安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的账号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

Manager访问地址为“集群管理页面”地址。访问时需要输入集群的用户名和密码，例如“admin”用户。

说明

- 如果使用其他集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。请咨询MRS集群管理员。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

步骤7 注销用户退出Manager时移动鼠标到右上角 ，然后单击“注销”。

----结束

10.2 访问 MRS Manager（MRS 2.x 及之前版本）

操作场景

MRS使用Manager对集群进行监控、配置和管理，用户可以在MRS控制台页面打开Manager管理页面，使用创建集群时设置的admin账号和密码登录Manager。

通过弹性公网 IP 访问 Manager

步骤1 登录MRS管理控制台页面。

步骤2 单击“集群列表 > 现有集群”，在集群列表中单击指定的集群名称，进入集群信息页面。

步骤3 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中“访问方式”选择“EIP访问”。专线访问请参考**通过专线访问**。

1. 若用户创建集群时暂未绑定弹性公网IP，在弹性公网IP下拉框中选择可用的弹性公网IP。若用户创建集群时已经绑定弹性公网IP，直接执行**步骤3.2**。

   说明

   - 如果没有弹性公网IP，可先单击“管理弹性公网IP”创建弹性公网IP后，然后在弹性公网IP下拉框中选择创建的弹性公网IP。
   - 如果在使用完后需要解绑或释放弹性公网IP，请登录“弹性公网IP”界面，在待操作的弹性公网IP后，单击“操作”列的“解绑”或“更多 > 释放”。
   - 如果已创建弹性公网IP，但在绑定时无法找到，可能是由于该弹性公网IP被其他集群绑定，请先在弹性公网IP界面解绑，然后再为当前集群绑定。

2. 选择待添加的安全组规则所在安全组，该安全组在创建群时配置。
3. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则，如需开放多个IP段为可信范围用于访问MRS Manager页面，请参考**步骤6**～**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

   说明

   - 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
   - 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问MRS Manager服务。

4. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

步骤4 单击“确定”，进入MRS Manager登录页面。

步骤5 输入默认用户名“admin”及创建集群时设置的密码，单击“登录”进入MRS Manager页面。

步骤6 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

说明

如需给其他用户开通访问MRS Manager的权限，请执行**步骤6**-**步骤9**，添加对应用户访问公网的IP地址为可信范围。

步骤7 单击弹性公网IP后边的“添加安全组规则”。

步骤8 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的授权对象是可信任的公网访问IP范围，禁止使用0.0.0.0/0,否则会有安全风险。”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤6**-**步骤9**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤9 单击“确定”完成安全组规则添加。

----结束

通过 ECS 访问 Manager

步骤1 在MRS管理控制台，单击“集群列表”。

步骤2 在 “现有集群” 列表中，单击指定的集群名称。

记录集群的“可用区”、“虚拟私有云”、“安全组”。

步骤3 在ECS管理控制台，创建一个新的弹性云服务器。

- 弹性云服务器的“可用区”、“虚拟私有云”、“安全组”，需要和待访问集群的配置相同。
- 选择一个Windows系统的公共镜像。例如，选择一个企业镜像“Enterprise_Windows_STD_2012R2_20170316-0(80GB)”。
- 其他配置参数详细信息，请参见“弹性云服务器 > 用户指南 > 快速入门 > 创建并登录Windows弹性云服务器”。

说明

如果ECS的安全组和MRS集群的“默认安全组”不同，用户可以选择以下任一种方法修改配置：

- 将ECS的安全组修改为MRS集群的默认安全组，请参见“弹性云服务器 > 用户指南 > 安全组 > 更改安全组”。
- 在集群Master节点和Core节点的安全组中添加两条安全组规则使ECS可以访问集群，“协议”需选择为“TCP”，“端口”需分别选择“28443”和“20009”。请参见“虚拟私有云 > 用户指南 > 安全性 > 安全组 > 添加安全组规则”。

步骤4 在VPC管理控制台，申请一个弹性IP地址，并与ECS绑定。

具体请参见“虚拟私有云 > 用户指南 > 弹性公网IP > 为弹性云服务器申请和绑定弹性公网IP”。

步骤5 登录弹性云服务器。

登录ECS需要Windows系统的账号、密码，弹性IP地址以及配置安全组规则。具体请参见“弹性云服务器 > 用户指南 > 实例 > 登录弹性云服务器 > 登录Windows弹性云服务器”。

步骤6 在Windows的远程桌面中，打开浏览器访问Manager。

例如Windows 2012操作系统可以使用Internet Explorer 11。

Manager访问地址形式为**https://*集群控制台地址*:28443/web**。访问时需要输入MRS集群的用户名和密码，例如“admin”用户。

说明

- 集群控制台地址：远程登录Master2节点，执行“ifconfig”命令，系统回显中“eth0:wsom”表示集群控制台地址，请记录“inet”的实际参数值。如果在Master2节点无法查询到集群控制台地址，请切换到Master1节点查询并记录。如果只有一个Master节点时，直接在该Master节点查询并记录。
- 如果使用其他MRS集群用户访问Manager，第一次访问时需要修改密码。新密码需要满足集群当前的用户密码复杂度策略。
- 默认情况下，在登录时输入5次错误密码将锁定用户，需等待5分钟自动解锁。

步骤7 注销用户退出Manager时移动鼠标到右上角 ，然后单击“注销”。

----结束

为集群更换弹性公网 IP

步骤1 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤2 查看“弹性公网IP”。

步骤3 登录“虚拟私有云 VPC”管理控制台。

步骤4 选择“弹性公网IP和带宽 > 弹性公网IP”。

步骤5 查找MRS集群所绑定的弹性公网IP，并在“操作”列单击“解绑”解绑MRS集群绑定的弹性公网IP。

步骤6 登录MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

此时，集群详情页面“弹性公网IP”显示“暂未绑定”。

步骤7 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中“访问方式”选择“EIP访问”。

步骤8 在弹性公网IP下拉框中选择新的弹性公网IP并配置他参数，具体请参考**通过弹性公网IP访问Manager**。

----结束

为其他用户开通访问 MRS Manager 的权限

步骤1 在MRS管理控制台，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

步骤2 单击弹性公网IP后边的“添加安全组规则”。

步骤3 进入“添加安全组规则”页面，添加需要开放权限用户访问公网的IP地址段并勾选“我确认这里设置的授权对象是可信任的公网访问IP范围，禁止使用0.0.0.0/0,否则会有安全风险。”

默认填充的是用户访问公网的IP地址，用户可根据需要修改IP地址段，如需开放多个IP段为可信范围，请重复执行**步骤1**-**步骤4**。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

步骤4 单击“确定”完成安全组规则添加。

----结束

11 FusionInsight Manager 操作指导（适用于 3.x）

11.1 从这里开始

11.1.1 FusionInsight Manager 入门指导

概述

MRS为用户提供海量数据的管理及分析功能，快速从结构化和非结构化的海量数据中挖掘您所需要的价值数据。开源组件结构复杂，安装、配置、管理过程费时费力，使用FusionInsight Manager将为您提供企业级的集群的统一管理平台：

- 提供集群状态的监控功能，您能快速掌握服务及主机的运行状态。
- 提供图形化的指标监控及定制，您能及时的获取系统的关键信息。
- 提供服务属性的配置功能，满足您实际业务的性能需求。
- 提供集群、服务、角色实例的操作功能，满足您一键启停等操作需求。
- 提供权限管理及审计功能，您能设置访问控制及管理操作日志。

系统界面简介

FusionInsight Manager提供统一的集群管理平台，帮助您快捷、直观的完成集群的运行维护。

界面最上方为操作栏，中部为显示区，最下方为任务栏。

操作栏各操作入口的详细功能如表11-1所示。

表 11-1 界面操作入口功能描述

入口	功能描述
主页	提供柱状图、折线图、表格等多种图表方式展示集群的主要监控指标、主机的状态统计。您可以定制关键监控信息面板，并拖动到任意位置。系统概览支持数据自动刷新，请参见**主页**。
集群	提供各集群内服务监控、服务操作向导以及服务配置，帮助您对服务进行统一管理。请参见**集群**。
主机	提供主机监控、主机操作向导，帮助您对主机进行统一管理。请参见**主机**。
运维	提供告警查询、告警处理指导功能。帮助您及时发现产品故障及潜在隐患，并进行定位排除，以保证系统正常运行。请参见**运维**。
审计	提供审计日志查询及导出功能。帮助您查阅所有用户活动及操作。请参见**审计**。
租户资源	提供统一租户管理平台。请参见**租户资源**。
系统	提供对FusionInsight Manager的系统管理设置，例如用户权限设置。请参见**系统设置**。

11.1.2 查询 FusionInsight Manager 版本号

通过查看FusionInsight Manager版本号，可以进行下一步的系统升级及日常维护操作。

- 界面方式

  登录FusionInsight Manager，在主页界面，单击右上角的 ，在下拉框中单击“关于”，在弹框中查看FusionInsight Manager版本号。

- 命令方式

  a. 以**root**用户登录FusionInsight Manager主管理节点。

  b. 执行如下命令，查看FusionInsight Manager的版本号及平台信息。

  su - omm

  cd ${BIGDATA_HOME}/om-server/om/sbin/pack

  ./queryManager.sh

  显示如下：

  ```
  Version          Package                          Cputype
  ***              FusionInsight_Manager_***        x86_64
  ```

  说明

  ***表示版本号，具体以实际查询的版本号为准。

11.1.3 登录管理系统

操作场景

该任务指导用户在Manager安装完成后使用账号登录FusionInsight Manager。

操作步骤

步骤1 获取FusionInsight Manager的网络地址。

步骤2 打开页面后，输入系统用户和密码。

步骤3 新用户登录需要修改密码。

用户密码策略：

- 密码字符长度必须为8～64个字符。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符`~!@#$%^&*()-_=+|[{}];',<.>/\?中的4种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与当前密码相同。

步骤4 将光标移动到FusionInsight Manager右上角的，在弹出窗口中单击“注销”，单击“确定”后可退出当前登录用户。

----结束

11.1.4 登录管理节点

操作场景

部分运维操作的脚本与命令需要或只支持在主管理节点上运行。系统管理员可以根据以下指导确认并登录主或备管理节点。

在 Manager 查看主备管理节点并登录

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > OMS”。

在“基本信息”区域，“当前主用”表示主管理节点的主机名，“当前备用”表示备管理节点的主机名。

单击主机名可进入对应的主机详情页面。记录主机的IP地址信息。

步骤3 以**root**用户登录主或备管理节点。

----结束

执行脚本确定主备管理节点并登录

步骤1 以**root**用户登录任意部署Manager的节点。

步骤2 执行以下命令确认主备管理节点。

su - omm

sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh

界面打印信息中“HAActive”参数值为“active”的节点为主管理节点（如下例中“Master1”为主管理节点），参数值为“standby”的节点为备管理节点（如下例中“Master2”为备管理节点）。

```
HAMode
double
NodeName            HostName        HAVersion       StartTime           HAActive
HAAllResOK          HARunPhase
192-168-0-30        Master1         V100R001C01     2021-09-01 07:12:05     active
normal              Actived
192-168-0-24        Master2         V100R001C01     2021-09-01 07:14:02     standby
normal              Deactived
```

步骤3 执行如下命令获取主备管理节点IP地址。

cat /etc/hosts

获取的主备管理节点IP地址示例如下：

```
127.0.0.1       localhost
192.168.0.30    Master1
192.168.0.24    Master2
```

步骤4 以**root**用户登录主或备管理节点。

----结束

11.2 主页

11.2.1 主页概述

登录FusionInsight Manager以后，界面将默认显示“主页”内容，“综述”页面提供各集群服务状态预览区及监控状态报表，“告警分析”页面展示TOP告警统计及分析。

- 主页右侧可查看集群的不同级别告警个数、运行任务个数、当前用户和帮助信息等内容。
  - 单击 可查看“任务管理中心”中近100次操作任务的任务名、集群、状态、进度、开始时间和结束时间。

    说明

    对于启动、停止、重启以及滚动重启操作，在任务执行过程中，单击任务列表中的对应任务名称，单击“中止”按钮，根据界面提示输入系统管理员密码后，用户可中止该任务。中止后，任务将不再继续执行。

  - 单击 可获得帮助信息，如表11-2所示。

    表 11-2 帮助信息一览表

项目	描述
“关于”	提供FusionInsight Manager版本号信息。

- 主页底部任务栏显示FusionInsight Manager的语言选项和当前集群时间及时区信息，可切换系统语言。

服务状态预览区

主页界面的左侧展示各集群主机个数及已安装服务个数，可通过单击 ，展开对应集群的全部服务信息，查看当前集群已安装各服务的状态和告警情况。

通过单击 ，对当前集群进行基本的运维管理操作，详情请参考**表11-3**。

每个服务名称左侧的表示当前该服务运行状态良好，表示当前服务启动失败，表示当前服务未启动。

同时服务名称右侧可查看当前该服务是否产生了告警，如果存在告警，则以图标区分告警的级别并显示告警数。

对于支持多服务特性的组件，若在同一集群中安装了多个服务，服务的右侧会显示安装的个数。

如果服务右侧显示则表示该服务配置已过期。

监控状态报表

主页界面的右侧为图表区，包含关键监控状态的报表，例如集群中所有主机的状态、主机CPU使用率、主机内存使用率等。用户可以自定义在图表区展示的监控报表，管理监控指标请参考**管理监控指标数据报表**。

监控图表的数据来源可在图表的左下方查看，每个监控报表可以放大查看具体数值，也可以关闭不再显示。

告警分析

“告警分析”页面展示“Top20告警统计”表和“Top3告警分析”图。单击“Top20告警统计”中的告警名称，可以在告警分析中只展示该告警信息。该功能支持告警统计，可以展示TOP告警以及发生的时间规律，可以有针对性地解决告警，提升系统稳定性。

11.2.2 管理监控指标数据报表

操作场景

FusionInsight Manager支持用户自定义在主页进行展示的监控项，也可以导出监控数据。

说明

历史报表根据所自定义的时间长度不同，图表横轴中每个时间间隔也会不同，具体监控数据的规则如下：

- 0～25小时：每个间隔5分钟，要求集群至少安装10分钟以上，最多保留15天监控数据。
- 25小时～150小时：每个间隔30分钟，要求集群至少安装30分钟以上，最多保留3个月监控数据。
- 150小时～300小时：每个间隔1小时，要求集群至少安装1小时以上，最多保留3个月监控数据。
- 300小时～300天：每个间隔1天，要求集群至少安装1天以上，最多保留6个月监控数据。
- 300天以上：每个间隔7天，要求集群安装7天以上，最多保留一年的监控数据。
- 如果FusionInsight Manager存储所用的GaussDB所在分区的磁盘使用率超过80%时，会清理实时监控数据和周期为5分钟的监控数据。
- 若为“租户资源”下的“存储资源(HDFS)”表，0小时～300小时：每个间隔1小时，要求集群至少安装1小时以上，最多保留3个月监控数据。

自定义监控指标报表

步骤1 登录FusionInsight Manager。

步骤2 单击“主页”。

步骤3 在图表区的右上角，单击 ▼ ，在弹出菜单中选择“定制”。

说明

监控时段以5分钟为单位，显示最近1小时的监控数据；从进入“实时监控”页面后，在监控图右侧以5分钟为单位显示实时监控数据。

步骤4 在窗口左侧分类中，选择一项监控资源主体。

步骤5 在右侧监控列表勾选一个或多个监控指标。

步骤6 单击“确定”。

----结束

导出全部监控数据

步骤1 登录FusionInsight Manager。

步骤2 单击“主页”。

步骤3 在所需要操作的集群的图表区的右上角，选择一个时间范围获取监控数据，例如“1周”。

默认为实时数据，无法导出。单击可以自定义监控数据时间范围。

步骤4 在图表区的右上角，单击 ▼ ，在弹出菜单中选择“导出”。

----结束

导出指定监控项数据

步骤1 登录FusionInsight Manager。

步骤2 单击“主页”。

步骤3 在所需要操作的集群的图表区任意一个监控报表窗格的右上角，单击。

步骤4 选择一个时间范围获取监控数据，例如“1周”。

默认为实时数据，无法导出。单击可以自定义监控数据时间范围。

步骤5 单击“导出”。

----结束

11.3 集群

11.3.1 管理集群

11.3.1.1 集群管理概述

总览

登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 概览”可以查看当前集群的主要状态信息。

在“概览”页面上可对当前集群进行基本管理操作，如启动、停止、滚动重启、同步配置等，具体如**表11-3**所示。

表 11-3 维护管理功能

操作入口	说明
“启动”	将集群中所有服务启动。
“停止”	将集群中所有服务停止。
“更多 > 重启”	将集群中所有服务重启。
“更多 > 滚动重启”	为集群中所有服务提供不中断业务的重启操作，具体可参考**滚动重启集群**。
“更多 > 同步配置”	为集群中所有服务启用新的配置参数。
“更多 > 重启配置过期的实例”	为集群中所有服务重启配置过期的实例，具体可参考**管理配置过期**。
“更多 > 健康检查”	为OMS、集群所有服务和所有节点进行健康检查，健康检查可以包含三方面检查项：各检查对象的运行状态、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“运行状态”。 健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。如果发现问题，可以单击“查看帮助”。
“ 更多 > 下载客户端”	为用户下载默认的客户端，具体可参考**下载客户端**。

操作入口	说明
“ 更多 > 导出安装模板”	将集群所有安装配置批量导出，例如集群认证模式、节点信息、服务配置等，可用于相同环境下集群重新安装的场景。
“更多 > 导出配置”	将集群所有服务的配置批量导出。
“更多 > 进入维护模式/退出维护模式”	配置集群进入/退出维护模式。
“更多 > 维护模式视图”	查看集群进入维护状态的服务或主机。

11.3.1.2 滚动重启集群

操作场景

滚动重启指当集群中服务角色升级更新或修改配置后，在尽可能不中断业务的前提下的重启操作。

如果需要批量为集群中所有服务进行重启且不中断业务，可执行集群滚动重启操作。

说明

- 部分服务不支持滚动重启，在执行滚动重启集群的过程中，不支持滚动重启的服务将进行普通重启，业务可能会中断。请根据界面提示是否可以执行操作。
- 如果修改了端口类等需要尽快生效的配置（例如服务端的端口），则不建议通过滚动重启的方式使之生效，建议采用普通重启。

对系统的影响

与普通重启相比，滚动重启不会导致服务业务中断，但是滚动重启将比普通重启要花费更长的时间，且对应服务的吞吐量、性能等可能会受到影响。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 滚动重启”。

步骤3 输入当前登录的用户密码确认身份，单击“确定”。

步骤4 根据实际情况调整相关参数，如表11-4所示。

表 11-4 滚动重启参数

参数名称	描述
“只重启集群内配置过期的实例”	是否只重启集群内修改过配置的实例。

参数名称	描述
“启用机架策略”	是否启用机架并发滚动重启策略，只对满足机架策略滚动重启的角色（角色支持机架感知功能，且角色下的实例归属于2个或2个以上的机架）生效。 **说明** 该参数仅在滚动重启HDFS、Yarn时可设置。
“数据节点滚动重启并发数”	采用分批并发滚动重启策略的数据节点实例每一个批次重启的实例数，默认为1。 **说明** ● 该参数仅对同时满足“采用并发滚动策略”和“实例为数据节点”两个条件时才有效。 ● 当启用机架策略时，该参数将失效，集群以机架策略默认配置的最大实例数（默认值为20）作为一个机架内分批并发重启的最大实例数。 ● 该参数仅在滚动重启HDFS、HBase、Yarn、Kafka、Storm、Flume时可设置。 ● HBase的RegionServer滚动重启的并发数不支持手动配置，会根据RegionServer的节点数自行调整，调整规则为：30节点以内，每个批次1个节点；300节点以内，每个批次2个节点；300节点以上(含300节点)，每个批次1%(向下取整)个节点。
“批次时间间隔”	滚动重启实例批次之间的间隔时间，默认为0。
“退服超时时间”	角色实例在滚动重启过程中的退服等待时间，默认为1800s。 部分角色（例如HiveServer、JDBCServer）在滚动重启前会暂时停止提供服务，该状态下的实例不可再接入新的客户端连接，而已经存在的连接需要等待一段时间才能完成，配置合适的超时时间参数能尽可能地保证业务不中断。 **说明** 该参数仅在滚动重启Hive、Spark2x时可设置。
“批次容错阈值”	滚动重启实例批次执行失败容错次数，默认为0，即表示任意一个批次的实例重启失败后，滚动重启任务终止。

说明

“数据节点滚动重启并发数”、“批次时间间隔”、“批次容错阈值”等高级参数需要根据实际情况合理设置，否则可能导致服务业务中断或者严重影响性能，请谨慎调整。

例如：

- “数据节点滚动重启并发数”过大，同时重启多个实例导致服务业务中断或者由于剩余工作实例较少严重影响性能。
- “批次容错阈值”过大，某一批次实例失败后继续重启下一批次实例，导致服务业务中断。

步骤5 单击“确定”，等待滚动重启完成。

----结束

11.3.1.3 管理配置过期

操作场景

某个新的配置需要同时下发到集群所有服务，或修改某项配置后导致多个不同服务的“配置状态”为“配置过期”或“失败”时，表示这些服务的配置参数值未同步且未生效，集群管理员可以对集群执行同步配置功能，并在同步配置后重启相关服务实例，使所有服务启用新的配置参数。

若集群中服务配置均已同步但未生效，需重启配置过期的实例。

对系统的影响

- 集群执行同步配置后，需要重启配置过期的服务。重启时对应的服务不可用。
- 重启配置过期的实例时，该实例不可用。

操作步骤

同步配置

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 概览”。

步骤3 选择“更多 > 同步配置”。

步骤4 在弹出窗口中单击“确定”，开始为当前集群同步配置。

----结束

重启配置过期的实例

步骤1 选择“更多 > 重启配置过期的实例”。

步骤2 在弹出窗口中输入当前登录的用户密码确认身份，然后单击“确定”。

步骤3 在确认重启实例的对话框中单击“确定”。

支持单击“查看实例”打开所有配置已过期的实例列表，确认可以执行重启任务。

----结束

11.3.1.4 下载客户端

操作场景

MRS集群提供了默认的客户端，用户可以通过客户端执行管理操作、运行业务或进行二次开发。使用客户端前需要下载客户端软件包。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 下载客户端”。

界面显示“下载集群客户端”对话框。

步骤3 在“选择客户端类型”选择一个类型。

- “完整客户端”表示下载包中包含了脚本、编译文件和配置文件。
- “仅配置文件”表示下载包仅包含客户端配置文件。

  一般适用于应用开发任务。例如完整客户端已下载并安装后，集群管理员通过Manager界面修改了服务配置，开发人员需要更新客户端配置文件的场景。

  说明

  平台类型包括x86_64和aarch64两种，可分别在x86和TaiShan节点上安装使用。默认情况下，下载的客户端平台类型和服务端保持一致。

步骤4 是否在集群的节点中生成客户端软件包文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件。

  文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client/”。支持修改为其他目录且**omm**用户拥有目录的读、写与执行权限。如果路径中已存在客户端文件，会覆盖路径下已有的客户端文件。

  等待文件生成后，使用**omm**用户或客户端安装用户将获取的下载包复制到其他目录，例如“/opt/Bigdata/hadoopclient”。
- 否，单击“确定”，下载客户端文件至本地。

  开始下载客户端软件包，并等待下载完成。

客户端下载成功后，参考**安装客户端**进行客户端的安装。

----结束

11.3.1.5 修改集群属性

操作场景

FusionInsight Manager支持用户在集群安装完成后查看基本属性。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 集群属性”。

默认可查看集群名称、集群描述、产品类型、集群ID、认证模式、创建时间和已安装部件信息。

步骤3 修改“集群名称”。

1. 单击，填入新的名称。

   支持的命名规则：集群名称只能包含汉字、字母、数字、下划线（_）、中划线（-）和空格，仅以汉字、字母、数字、下划线（_）或中划线（-）开头，只能在中间包含空格，并且最小长度为2个字符，最大长度不能超过199个字符。
2. 单击“确定”使新的集群名称生效。

步骤4 修改“集群描述”。

1. 单击，填入新的描述信息。

   只能包含汉字、英文字母、数字、中英文逗号、中英文句号、下划线（_）、空格和换行符，并且不能超过199个字符。

2. 单击“确定”使新的描述生效。

----结束

11.3.1.6 管理集群配置

操作场景

FusionInsight Manager支持一键查看集群内各服务配置参数的变动情况，方便用户快速排查定位问题，提升配置管理效率。

集群管理员可通过配置界面快速查看集群内各服务所有非初始默认值、同一角色实例之间非统一值、集群配置修改的历史记录、集群内当前配置状态为过期的参数。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 配置”。

步骤3 根据操作场景，选择对应操作页面：

- 查看所有非默认值：
  a. 单击“所有非默认值”，界面将显示当前集群内各服务、角色或实例的配置参数中，与初始默认值不一致的参数项。

     单击参数值后面的图标可快速恢复配置项的参数值至系统默认值，单击图标可查看该配置项的历史修改记录。

     配置参数较多时，可通过界面右上角的服务过滤框进行筛选，或者在搜索框中直接搜索关键字。
  b. 如需修改配置项参数值，根据参数描述修改配置后，单击“保存”，在弹出的窗口中单击“确定”。
- 查看所有非统一值：
  a. 单击“所有非统一值”，界面将显示当前集群内角色级别、服务级别、实例组级别或实例级别的存在差异化配置的配置项。

     单击参数值后面的 图标，在弹出的窗口中可查看具体的差异项。

  b. 如需修改配置项参数值，可单击 取消下层的配置差异化或手动调整，然后单击“确定”，再单击“保存”，在弹出的窗口中单击“确定”。
- 查看过期配置：
  a. 单击“过期配置”，界面将显示当前集群内配置过期的配置项。
  b. 可通过界面上方的服务过滤框进行筛选，查看不同服务的过期配置，或者在搜索框中直接搜索关键字。
  c. 处于过期状态的配置项并未完全生效，在不影响业务情况下，请及时重启配置过期的服务或实例。
- 查看历史配置记录：
  a. 单击“历史配置”，界面将显示当前集群的历史配置变更记录，用户可查看具体的参数值变动详情，包括所属服务、修改前与修改后的参数值、参数文件等内容。

b. 如需还原某次配置变更，可单击记录所在行“操作”列的“还原配置”按钮，在弹出的窗口中单击“确定”。

说明

部分配置项在修改参数值后需重启对应服务才会生效，在保存配置后请及时重启配置过期的服务或实例。

----结束

11.3.1.7 静态服务池

11.3.1.7.1 静态服务资源

简介

集群分配给各个服务的资源是静态服务资源，这些服务包括Flume、HBase、HDFS、IoTDB、Kafka（Kafka组件仅MRS 3.2.0及之后版本支持静态服务池）和Yarn。每个服务的计算资源总量固定，不与其他服务共享，是静态的。租户通过独占或共享一个服务来获取这个服务运行时需要的资源。

静态服务池

静态服务池用来指定服务资源的配置。

在服务级别上，静态服务池对各服务可使用的资源进行统一管理：

- 限制服务使用的资源总量，支持配置Flume、HBase、HDFS、IoTDB、Kafka（Kafka组件仅MRS 3.2.0及之后版本支持静态服务池）和Yarn在部署节点可使用的CPU、I/O和内存总量。
- 实现服务级别的资源隔离，可将集群中的服务与其他服务隔离，使一个服务上的负载对其他服务产生的影响有限。

调度机制

静态服务资源支持基于时间的动态调度机制，可以在不同时间段为服务配置不同的资源量，优化客户业务运行环境，提高集群的效率。

在一个复杂的集群环境中，多种服务共享使用集群资源，但是各服务的资源使用周期可能会有比较大的区别。

例如以下业务场景，对于一个银行客户：

- 在白天HBase查询服务的业务多。
- 在晚上查询服务的业务少而Hive分析服务业务多。

如果只给每个服务设置固定的资源可能会导致：

- 白天查询服务的资源不够用，分析服务的资源空闲。
- 晚上分析服务的资源不够用，查询服务的资源空闲。

集群资源利用率不高，而且服务能力也打了折扣。因此：

- 白天多配置HBase服务资源。
- 晚上多配置Hive服务资源。

这种基于时间的动态调度机制可以更高效的利用资源，运行任务。

11.3.1.7.2 配置集群静态资源

操作场景

当需要控制集群服务可以使用节点资源的情况，或者控制集群服务在不同时间段节点可用配额的CPU与I/O资源时，集群管理员可以在FusionInsight Manager调整资源基数，并自定义资源配置组。

对系统的影响

- 配置静态服务池后，受影响的服务的“配置状态”将显示为“配置过期”，需要重启服务，重启期间服务不可用。
- 配置静态服务池后，各服务及角色实例使用的最大资源将不能超过限制。

操作步骤

修改资源调整基数

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 静态服务池”。

步骤2 单击右上角“配置”，进入静态资源池配置向导。

步骤3 在“系统资源调整基数”分别修改参数“CPU（%）”和“Memory（%）”。

修改“系统资源调整基数”将限制服务能够使用节点的最大物理CPU和内存资源百分比。如果多个服务部署在同一节点，则所有服务使用的最大物理资源百分比不能超过此参数值。

步骤4 单击“下一步”。

需要重新修改参数，可单击“上一步”返回。

修改资源池默认“default”配置组

步骤5 单击“default”，在“权重配置”表格中各服务对应的“CPU LIMIT(%)”、“CPU SHARE(%)”、“I/O(%)”和“Memory(%)”填写各服务的资源使用百分比数量。

说明

- 所有服务使用的“CPU LIMIT(%)”和 “CPU SHARE(%)”资源配置总和可以大于100%。
- 所有服务使用的“I/O(%)”资源配置总和可以大于100%，不能为0。
- 所有服务使用的“Memory(%)”资源配置总和可以小于或等于100%，也可以大于100%。
- “Memory(%)”不支持动态生效，仅在“default”配置组中可以修改。
- “CPU LIMIT(%)”用于配置服务可使用的CPU核数与节点可分配的CPU核数占比。
- “CPU SHARE(%)”用于配置服务在与其他服务使用同一个CPU核的时间占比，即多个服务在使用同一个CPU核发生争抢时的时间占比。

步骤6 单击“根据权重配置生成详细配置”，FusionInsight Manager将根据集群硬件资源与分配情况，生成资源池实际参数配置值。

步骤7 单击“确定”。

在弹出窗口单击“确定”，确认保存配置。

添加自定义资源配置组

步骤8 是否需要在不同时间段自动调整资源配置?

- 是，执行**步骤9**。
- 否，只需要使用“default”在所有时间段生效，任务结束。

步骤9 单击“配置”，修改“系统资源调整基数”，然后单击“下一步”。

步骤10 单击“添加”增加新的资源配置组。

步骤11 在“第一步：调度时间”，单击“配置”显示时间策略配置页面。

根据业务需要修改以下参数，并单击“确定”保存：

- “重复”：勾选时表示此资源配置组按调度周期重复运行。不勾选时请设置一个资源配置组应用的日期与时间。
- “重复策略”：支持“每天”、“每周”和“每月”。仅在“重复”模式中生效。
- “在”：表示资源配置组应用的开始与结束时间。请设置一个唯一的时间区间，如果与已有配置组的时间区间有重叠，则无法保存。

说明

- “default”配置组会在所有未定义的时间段内生效。
- 新增加的配置组属于动态生效的配置项集合，在配置组应用的时间区间内可直接生效。
- 新增加的配置组可以被删除。最多增加4个动态生效的配置组。
- 选择任一种“重复策略”，如果结束时间小于开始时间，默认标识为第二天的结束时间。例如“22:00”到“6:00”表示调度时间为当天22点到第二天6点。
- 若多个配置组的“重复策略”类型不相同，则时间区间可以重叠，且生效的策略优先级从低到高的顺序为“每天”、“每周”、“每月”。例如，有“每月”与“每天”的调度配置组，时间区间分别为4:00到7:00，6:00到8:00，此时以每月的配置组为准。
- 若多个配置组的“重复策略”类型相同，当日期不相同时，则时间区间可以重叠。例如，有两个“每周”的调度配置组，可以分别指定时间区间为周一和周三的4:00到7:00。

步骤12 在“第二步：权重配置”修改各服务资源配置。

步骤13 单击“根据权重配置生成详细配置”，FusionInsight Manager将根据集群硬件资源与分配情况，生成资源池实际参数配置值。

步骤14 单击“确定”。

在弹出窗口单击“确定”，确认保存配置。

----结束

11.3.1.7.3 查看集群静态资源

操作场景

大数据管理平台支持通过静态服务资源池对没有运行在Yarn上的服务资源进行管理和隔离。系统支持基于时间的静态服务资源池自动调整策略，使集群在不同的时间段自动调整参数值，从而更有效地利用资源。

系统管理员可以在FusionInsight Manager查看静态服务池各个服务使用资源的监控指标结果，包含监控指标如下：

- 服务总体CPU使用率

- 服务总体磁盘IO读速率
- 服务总体磁盘IO写速率
- 服务总体内存使用大小

说明

启用多实例功能后，支持管理HBase所有服务实例使用的CPU、I/O和内存总量。

操作步骤

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 静态服务池”。

步骤2 在“配置组列表”，单击一个配置组，例如“default”。

步骤3 查看系统资源调整基数。

- “系统资源调整基数”表示集群中每个节点可以被集群服务使用的最大资源。如果节点只有一个服务，则表示此服务独占节点可用资源。如果节点有多个服务，则表示所有服务共同使用节点可用资源。
- “CPU”表示节点中服务可使用的最大CPU。
- “Memory”表示节点中服务可使用的最大内存。

步骤4 在图表区域，查看集群服务资源使用状态指标数据图表。

说明

- 可通过“为图表添加服务”，将特定服务的静态服务资源数据添至图表，最多可选择12个服务。
- 管理单个图表的操作，可参见**管理监控指标数据报表**。

----结束

11.3.1.8 客户端管理

11.3.1.8.1 管理客户端

操作场景

FusionInsight Manager支持统一管理集群的客户端安装信息，用户下载并安装客户端后，界面可自动记录已安装（注册）客户端的信息，方便查询管理。同时系统支持手动添加、修改未自动注册的客户端信息（如历史版本已安装的客户端）。

操作步骤

查看客户端信息

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 >*待操作集群的名称* > 客户端管理”，即可查看当前集群已安装的客户端信息。

用户可查看客户端所在节点的IP地址、安装路径、组件列表、注册时间及安装用户等信息。

在当前最新版本集群下载并安装客户端时，客户端信息会自动注册。

添加客户端信息

步骤3 如需手动添加已安装好的客户端信息，单击“添加”，根据界面提示手动添加客户端的IP地址、安装路径、用户、平台信息、注册信息等内容。

步骤4 配置好客户端信息，单击“确定”，添加成功。

修改客户端信息

步骤5 手动注册的客户端信息可以手动修改。

在“客户端管理”界面选择待修改的客户端，单击“修改”。修改信息后，单击“确定”完成修改。

删除客户端信息

步骤6 在“客户端管理”界面选择待删除的客户端，单击“删除”，在弹出的窗口中单击“确定”，即可删除客户端信息。

如需删除多个客户端信息，勾选待删除的客户端，单击“批量删除”，在弹出的窗口中单击“确定”，即可删除客户端信息。

导出客户端信息

步骤7 在“客户端管理”界面选择待操作的客户端，单击“导出全部”可导出所有已注册的客户端信息到本地。

说明

客户端管理界面上组件列表栏只展示有真实客户端的组件，因此部分没有客户端的组件和客户端特殊的组件不会显示在组件列表栏。

不显示的组件有：

LdapServer、KrbServer、DBService、Hue、Mapreduce、Flume

----结束

11.3.1.8.2 批量升级客户端

操作场景

在FusionInsight Manager界面上下载的客户端包中包含客户端批量升级工具，当集群升级或扩容后需要对多个客户端进行升级时，可以使用该工具对客户端进行批量一键升级。同时客户端批量升级工具提供了轻量级的批量刷新客户端所在节点“/etc/hosts”文件的功能。

操作步骤

客户端升级前准备

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 >*待操作集群的名称* > 更多 > 下载客户端”，下载完整客户端到服务端指定目录。

具体操作参考**下载客户端**。

解压新下载的客户端，在解压后的目录找到batch_upgrade目录，例如“/tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig/batch_upgrade”。

步骤3 选择“集群 >*待操作集群的名称* > 客户端管理”，进入客户端管理界面，单击“导出全部”，将所选的客户端信息导出到本地。

步骤4 解压导出的客户端信息，将client-info.cfg文件上传到客户端解压目录的batch_upgrade目录下。

步骤5 参见**参考信息**，补全“client-info.cfg”中缺失的密码。

批量升级客户端

步骤6 执行**sh client_batch_upgrade.sh -u -f /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar -g /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig/batch_upgrade/client-info.cfg**，进行升级。

须知

由于配置了密码信息，执行完升级后建议尽快删除**client-info.cfg**文件。

步骤7 升级执行完成后确认结果。确保客户端升级无误后执行**sh client_batch_upgrade.sh -c**，确认升级结果。

步骤8 如果客户端升级后存在问题，可以执行**sh client_batch_upgrade.sh -s**，回滚客户端。

说明

- 客户端批量升级工具本身是将原客户端move至备份目录，然后再使用**-f**参数指定的客户端包再次安装客户端。因此若原客户端中有定制的内容，请在执行**-c**命令之前，将定制的内容从备份目录手动保存或者移至升级后的客户端目录。客户端备份路径为：*{原客户端路径}*_backup。
- 参数**-u**是**-c**和**-s**的前提，必须在**-u**命令执行了升级之后，才能选择是要执行**-c**进行提交还是**-s**进行回滚。
- 升级命令（**-u**）可以多次执行，每次执行只升级前面升级失败的客户端，跳过升级成功的客户端。
- 客户端批量升级工具也支持升级之前的旧客户端。
- 执行非root用户安装的客户端升级时，请确保相应用户在目标节点客户端所在目录及父目录的读写权限，否则会升级失败。
- **-f**参数输入的客户端包必须为全量客户端，不支持单组件或部分组件客户端包作为输入。

----结束

参考信息

批量升级客户端前，需手动配置远程登录客户端节点的用户密码信息：

执行**vi client-info.cfg**命令，添加用户密码信息。

例如：

```
clientIp,clientPath,user,password
10.10.10.100,/home/omm/client /home/omm/client2,omm,密码
```

配置文件各字段含义如下：

- clientIp：表示客户端所在节点IP地址。

- clientPath：客户端安装路径，可以包含多个路径，以空格分隔多个路径。注意路径不要以“/”结尾。
- user：节点用户名。
- password：节点用户密码信息。

说明

如果执行失败，请在执行目录的**work_space/log_*XXX***下查看**node.log**日志。

11.3.1.8.3 批量刷新 hosts 文件

操作场景

在FusionInsight Manager界面上下载的客户端包中包含客户端批量升级工具，该工具在提供批量升级客户端功能的同时，也提供了轻量级的批量刷新客户端所在节点“/etc/hosts”文件的功能。

前提条件

更新前准备请参考**批量升级客户端**章节“客户端升级前准备”步骤。

批量更新 hosts 文件

步骤1 检查需要更新“/etc/hosts”文件的节点的配置用户是否为“root”。

- 是，执行**步骤2**。
- 否，更改配置用户为“root”，再执行**步骤2**。

步骤2 执行**sh client_batch_upgrade.sh -r -f /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_Client.tar -g /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig/batch_upgrade/client-info.cfg**，批量刷新客户端所在节点的“/etc/hosts”文件。

说明

- 执行批量刷新“/etc/hosts”文件时，输入的客户端包可以是完整客户端，也可以是仅包含配置文件的客户端软件包，推荐使用仅包含配置文件的客户端软件包。
- 需要更新“/etc/hosts”文件的主机所配置的用户必须为**root**用户，否则会刷新失败。

----结束

11.3.2 管理服务

11.3.2.1 服务管理概述

总览

登录FusionInsight Manager以后，选择“集群 > *待操作集群的名称* > 服务”后，打开服务管理页面，包含功能区和服务列表。

功能区

服务管理页面的功能区支持选择视图类型，以及通过服务类型筛选和搜索服务。通过高级搜索可以根据“运行状态”和“配置状态”选择所需要的服务。

服务列表

服务管理页面的服务列表包含了集群中所有已安装的服务。如果选择“平铺视图”，则显示为窗格样式；如果选择“列表视图”，则显示为表格样式。

说明

本章节默认以“平铺视图”进行介绍。

服务列表可显示每个服务的运行状态、配置状态、角色的类型以及对应的实例个数。同时可以执行部分服务维护任务，例如启动、停止、重启服务等。

表 11-5 服务运行状态

状态	说明
良好	表示服务当前运行正常。
故障	表示服务当前无法正常工作。
亚健康	表示服务部分增强功能无法正常工作。
未启动	表示服务已停止。
未知	表示服务的初始状态信息无法检测。
正在启动	表示服务正在执行启动过程。
正在停止	表示服务正在执行停止过程。
启动失败	表示服务启动操作失败。
停止失败	表示服务停止操作失败。

说明

- 服务的运行状态为“故障”，会触发告警，请根据告警信息处理。
- HBase、Hive、Spark和Loader可显示“亚健康”（Subhealthy）状态。
  - Yarn已安装且不正常时，HBase处于“亚健康”状态。如启用多实例功能，则已安装的所有HBase服务实例处于“亚健康”状态。
  - HBase已安装且状态不正常时，Hive、Spark和Loader处于“亚健康”状态。
  - 启用多实例功能后，任意一个HBase服务实例已安装且不正常时，Loader处于“亚健康”状态。
  - 启用多实例功能后，某一个HBase服务实例已安装且不正常时，对应的Hive和Spark服务实例处于“亚健康”状态，即HBase2已安装且不正常时，Hive2和Spark2为“亚健康”状态。

表 11-6 服务配置状态

状态	说明
已同步	表示服务所有参数配置已在集群内全部生效。
配置过期	表示修改服务参数后，最新的配置未同步且未生效，需要同步配置且重启相应服务。可单击配置状态后的 图标查看过期的配置项。
失败	表示同步参数配置过程中出现通信或读写异常等操作。尝试使用“同步配置”恢复。
正在同步	表示正在同步服务参数配置。
未知	表示服务配置的初始状态信息无法检测。

服务列表中单击服务对应菜单，可对服务进行简单的维护管理操作，具体如表11-7所示。

表 11-7 基本维护管理功能

操作入口	说明
“启动服务”	启动集群中指定服务。
“停止服务”	将集群中指定服务停止。
“重启服务”	将集群中指定服务重启。 **说明** 某个服务可能被其他服务依赖，重启该服务则导致其他服务不可用，需要勾选“同时重启上层服务”。请根据对话框的服务列表确认是否可以执行操作，集群中由于依赖关系服务的重启为串行进行。单个服务的重启时长如表11-8所示。
“滚动重启服务”	为集群中指定服务提供不中断业务的重启操作，具体参数配置可参考表11-4。
“同步配置”	• 为集群中指定服务启用新的配置参数。 • 为集群中“配置状态”为“配置过期”的服务，下发新的配置参数。 **说明** 部分服务同步配置后需重启服务使配置生效。

表 11-8 重启时长

服务名称	重启时长	启动时长	附加说明
IoTDB	3min	IoTDBServer：3min	-

服务名称	重启时长	启动时长	附加说明
CDL	2min	CDLConnector：1min CDLService：1min	-
ClickHouse	4min	ClickHouseServer：2min ClickHouseBalancer：2min	-
HDFS	10min+x	NameNode：4min+x DataNode：2min JournalNode：2min Zkfc：2min	x为NameNode元数据加载时长，每千万文件大约耗时2分钟，例如5000万文件x为10min。由于受DataNode数据块上报影响启动时间有一定浮动。
Yarn	5min+x	ResourceManager：3min+x NodeManager：2min	x为ResourceManager保留任务数恢复时长，每1万保留任务大约需要1分钟
Mapreduce	2min+x	JobHistoryServer：2min+x	x为历史任务扫描时长，每10万任务大约2.5min
ZooKeeper	2min+x	quorumpeer：2min+x	x为加载znode节点时长，每100万znode大约1min
Hive	3.5min	HiveServer：3min MetaStore：1min30s WebHcat：1min Hive整体服务：3min	-
Spark2x	5min	JobHistory2x：5min SparkResource2x：5min JDBCServer2x：5min	-

服务名称	重启时长	启动时长	附加说明
Flink	4min	FlinkResource：1min FlinkServer：3min	-
Kafka	2min+x	Broker：1min+x	x为数据恢复时长，单实例20000 partition启动所需时长大约2mins。
Storm	6min	Nimbus：3mins UI：1min Supervisor：1min Logviewer：1min	-
Flume	3min	Flume：2 min MonitorServer：1min	-

11.3.2.2 其他服务管理操作

11.3.2.2.1 服务详情概述

总览

登录FusionInsight Manager以后，选择“集群 > *待操作集群的名称* > 服务”，在服务列表单击指定的服务名称打开服务详情页面，包含“概览”、“实例”、“实例组”和“配置”等页面，以及功能区。部分服务还支持显示自定义的管理工具页面，具体支持列表如表11-9所示。

表 11-9 自定义管理工具名称一览表

工具名称	对应服务	说明
Flume配置工具	Flume	用于为Flume的服务端和客户端配置采集参数。
Flume客户端管理工具	Flume	查看Flume客户端监控信息。
Kafka Topic监控工具	Kafka	用于为Kafka的Topic提供监控与管理。

其中“概览”为默认页，包含基本信息、角色列表、依赖关系表和监控图表等，右上角可对服务进行管理，基本管理如启动、停止、滚动重启、同步配置请参考**表11-7**，其他服务管理操作如**表11-10**所示：

表 11-10 服务管理操作

操作入口	说明
“更多 > 健康检查”	为当前服务进行健康检查，健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“运行状态”。 健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。如果发现问题，可以单击“查看帮助”。
“更多 > 下载客户端”	为用户下载默认的仅包含具体服务的客户端，通过客户端执行管理操作、运行业务或进行二次开发，具体可参考**下载客户端**。
“更多 > 修改服务名称”	修改当前服务名称。
“更多 > 执行*角色名称*倒换”	具体请参考**执行角色实例主备倒换**。
“更多 > 进入维护模式/退出维护模式”	配置服务进入/退出维护模式。
“配置 > 导入/导出”	在迁移服务到新集群场景或者重新部署相同服务的场景下，为具体服务的所有配置数据做导入或者导出操作，实现配置结果的快速复制。

基本信息

“概览”的基本信息包含该服务的基本状态数据，即运行状态、配置状态、版本，还包含各个服务自身关键信息。如果服务支持开源WebUI，则在基本信息区域可通过WebUI的链接访问开源WebUI。

说明

当前版本“admin”用户没有权限访问服务的开源WebUI完整功能。请另外创建组件业务管理员并访问WebUI地址。

角色列表

“概览”页面的角色列表包含了该服务中所有的角色。角色列表可显示每个角色的运行状态和角色的实例个数。

依赖关系表

“概览”页面的依赖关系表支持展示该服务依赖的服务，以及依赖此服务的其他服务。

告警和事件的历史记录

告警和事件的历史记录区显示了当前服务上报的关键告警与事件记录，系统最大可显示20条历史记录。

图表

“概览”页面的右侧展示图表区，包含该服务的各个关键监控指标报表。用户可以自定义在图表区展示的监控报表、可以打开监控指标的解释说明或导出监控数据。对于定制类别为资源贡献类的图表，支持放大后切换趋势图和分布图。

说明

集群中部分服务提供服务级别的资源监控项，具体请参考**资源监控**。

11.3.2.2.2 执行角色实例主备倒换

操作场景

部分服务的角色以主备高可用的模式进行部署，在需要对主实例进行维护不能提供服务，或者其他维护需要时，可以手动触发实例主备倒换。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在服务详情页面单击“更多”，选择“执行*角色实例*倒换”。

步骤5 输入当前登录的用户密码确认身份，单击“确定”。

步骤6 在弹出界面单击“确定”，执行角色实例主备倒换。

说明

- Manager部件包支持的主备倒换角色实例的服务有：DBService。
- HD部件包支持的主备倒换角色实例的服务有：HDFS、Yarn、Storm、HBase、Mapreduce。
- HDFS的角色NameNode在进行主备倒换时，需要指定NameService。
- Porter部件包支持的主备倒换角色实例的服务有：Loader。
- 其他角色实例则不支持此功能。

----结束

11.3.2.2.3 资源监控

登录FusionInsight Manager以后，选择“集群 > *待操作集群的名称* > 服务”后，选择待操作的服务，单击“资源”，进入资源监控页面。

集群中部分服务提供服务级别的资源监控项，默认显示12小时的监控数据。用户可单击自定义时间区间，缺省时间区间包括：12小时、1天、1周、1月。单击可导出相应报表信息，无数据的监控项无法导出报表。支持资源监控的服务及监控项如**表11-11**所示。

表 11-11 服务资源监控

服务	监控指标	说明
HDFS	资源使用（按租户）	• 按租户统计HDFS的资源使用情况。 • 可选择按“容量”或“文件对象数”观察。
	资源使用（按用户）	• 按用户统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件对象数”观察。
	资源使用（按目录）	• 按目录统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件对象数”观察。 • 单击 配置空间监控，可以指定HDFS文件系统目录进行监控。
	资源使用（按副本）	• 按副本数统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件数”观察。
	资源使用（按文件大小）	• 按文件大小统计HDFS的资源使用情况。 • 可选择按“已使用容量”或“文件数”观察。
	回收站（按用户）	• 按用户统计HDFS回收站的使用情况。 • 可选择按“回收站容量”或“文件对象数”观察。
	操作数	• 统计HDFS中操作数。
	自动balance	• 统计HDFS自动balancer的执行速度以及本次balancer当前迁移的总容量大小。
	NameNode RPC连接数（按用户）	• 按用户统计连接到NameNode的Client RPC请求中，各个用户的连接数。
	慢DataNode节点	集群中数据传输或处理慢的DataNode节点。
	慢磁盘	集群中DataNode节点上数据处理慢的磁盘。
HBase	表级别操作请求次数	所有RegionServer上的所有表中put、delete、get、scan、increment、append操作请求次数。

服务	监控指标	说明
	RegionServer级别操作请求次数	RegionServer中put、delete、get、scan、increment、append操作请求次数以及所有操作请求次数。
	服务级别操作请求次数	RegionServer上所有Region中put、delete、get、scan、increment、append操作请求次数。
	RegionServer级别HFile数	所有RegionServer中HFile数。
Hive	HiveServer2-Background-Pool线程数（按IP）	周期内统计并显示Top用户的HiveServer2-Background-Pool线程数。
	HiveServer2-Handler-Pool线程数（按IP）	周期内统计并显示Top用户的HiveServer2-Handler-Pool数监控。
	MetaStore使用数（按IP）	Hive周期内统计并显示Top用户的MetaStore使用数。
	Hive的Job数	Hive周期内统计并显示用户相关的Job数目。
	Split阶段访问的文件数	统计Hive周期内Split阶段访问底层文件存储系统（默认：HDFS）的文件数。
	Hive基本操作时间	Hive周期内统计底层创建目录（mkdirTime）、创建文件（touchTime）、写文件（writeFileTime）、重命名文件（renameTime）、移动文件（moveTime）、删除文件（deleteFileTime）、删除目录（deleteCatalogTime）所用的时间。
	表分区个数	Hive所有表分区个数监控，返回值的格式为：数据库#表名，表分区个数。
	HQL的Map数	Hive周期内执行的HQL与执行过程中调用的Map数统计，展示的信息包括：用户、HQL语句、Map数目。
	HQL访问次数	周期内HQL访问次数统计信息。
Kafka	Kafka磁盘使用率分布	Kafka集群的磁盘使用率分布统计。
Spark2x	HQL访问次数	周期内HQL访问次数统计信息，展示信息包括用户名，HQL语句，执行该语句的次数。
Yarn	资源使用（按任务）	• 任务使用的CPU核数和内存。 • 可选择“按内存”或“按CPU”观察。

服务	监控指标	说明
	资源使用（按租户）	• 租户所使用的CPU核数和内存。 • 可选择“按内存”或“按CPU”观察。
	资源使用比例（按租户）	• 租户所使用的CPU核数和内存的比例。 • 可选择“按内存”或“按CPU”观察。
	任务耗时排序	对Yarn任务耗时进行排序显示。
	ResourceManager RPC连接数（按用户）	统计连接到RM的Client RPC请求中，各个用户的连接数。
	操作数	统计Yarn每种操作类型对应的操作数及占比。
	队列中任务资源使用排序	• 在界面上选择某个队列（租户）后，显示在该队列中正在运行任务的消耗资源排序。 • 可选择“按内存”或“按CPU”观察。
	队列中用户资源使用排序	• 在界面上选择某个队列（租户）后，显示在该队列中正在运行任务的用户消耗的资源排序。 • 可选择“按内存”或“按CPU”观察。
ZooKeeper	资源使用（按二级Znode）	• ZooKeeper服务二级znode资源状况。 • 可选择“按Znode数量”或“按容量”观察
	连接数（按客户端IP）	ZooKeeper客户端连接资源状况。

11.3.2.2.4 采集堆栈信息

操作场景

为了满足实际业务的需求，集群管理员可以在FusionInsight Manager中采集指定角色或实例的堆栈信息，保存到本地目录，并支持下载。采集内容包括：

1. jstack栈信息。
2. jmap -histo堆统计信息。
3. jmap -dump堆信息快照。
4. 对于jstack和jmap-histo信息，支持连续采集以便对比。

操作步骤

采集堆栈

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > *待收集服务的名称*”。

步骤3 选择“更多 > 采集堆栈”。

说明

- 采集多个实例的堆栈信息：进入实例列表，勾选要采集的实例名称，选择“更多 > 采集堆栈”。
- 采集单个实例的堆栈信息：单击要采集的实例，选择“更多 > 采集堆栈”。

步骤4 根据界面提示，在弹框中选择需要采集的角色，采集内容，配置高级选项（若无特殊需求，保持默认配置即可），单击“确定”。

步骤5 采集成功后，单击“下载”。

下载堆栈信息

步骤6 选择“集群 > *待操作集群的名称* > 服务 > *待操作服务的名称*”。选择右上角“更多 > 下载堆栈信息”。

步骤7 选择需要下载的角色和内容，单击“下载”，可直接下载相关堆栈信息到本地。

清理堆栈信息

步骤8 选择“集群 > *待操作集群的名称* > 服务 > *待操作服务的名称*”。

步骤9 选择右上角“更多 > 清理堆栈信息”。

步骤10 选择需要清理的角色和内容，并配置“文件目录”。单击“确定”执行清理操作。

----结束

11.3.2.2.5 切换 Ranger 鉴权

操作场景

新安装的安全模式集群默认即安装了Ranger服务并启用了Ranger鉴权，用户可以通过组件的权限插件对组件资源的访问设置细粒度的安全访问策略。若不需使用Ranger进行鉴权，集群管理员可在服务页面手动停用Ranger鉴权，停用Ranger鉴权后，访问组件资源时系统将继续基于FusionInsight Manager的角色模型进行权限控制。

从历史版本升级的集群，用户访问组件资源时默认不使用Ranger鉴权，集群管理员可在安装了Ranger服务后手动启用Ranger鉴权。

说明

- 安全模式集群中，支持使用Ranger鉴权的组件包括：HDFS、Yarn、Kafka、Hive、HBase、Storm、Spark2x、Impala、CDL。
- 非安全模式集群中，Ranger可以支持基于OS用户进行组件资源的权限控制，支持启用Ranger鉴权的组件包括：HBase、HDFS、Hive、Spark2x、Yarn。
- 启用Ranger鉴权后，该组件所有鉴权将由Ranger统一管理，原鉴权插件设置的权限将会失效（HDFS与Yarn的组件ACL规则仍将生效），请谨慎操作，建议提前在Ranger上做好权限部署。
- 停用Ranger鉴权后，该组件所有鉴权将由组件自身权限插件管理，Ranger上设置的权限将会失效，请谨慎操作，建议提前在Manager上做好权限部署。

启用 Ranger 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在服务详情页面单击“更多”，选择“启用Ranger鉴权”。

步骤5 输入当前登录的用户密码确认身份，单击“确定”。

步骤6 在服务列表，重启配置过期的服务。

----结束

停用 Ranger 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 在服务详情页面单击“更多”，选择“停用Ranger鉴权”。

步骤5 输入当前登录的用户密码确认身份，单击“确定”，在弹出框中单击“确定”。

步骤6 在服务列表，重启配置过期的服务。

----结束

11.3.2.3 服务配置

11.3.2.3.1 修改服务配置参数

操作场景

为了满足实际业务的需求，集群管理员可以在FusionInsight Manager中快速查看及修改服务默认的配置。请务必参照配置描述中的建议进行参数配置。

说明

集群中只剩下一个DBService角色实例时，不支持修改DBService服务的参数。

对系统的影响

- 配置服务属性后，如果服务状态为“配置过期”，则需要重启此服务，重启期间该服务不可用。
- 修改服务配置参数并重启生效后，需要重新下载并安装客户端，或者下载配置文件刷新客户端。例如HBase、HDFS、Hive、Spark、Yarn、Mapreduce。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 单击“配置”。

默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。

步骤5 在导航树选择指定的参数分类，并在右侧修改参数值。

说明

对于端口类参数值请从右侧描述中的取值范围中选取，请确保同一个服务中所有参数项配置的值均在取值范围内且唯一，否则会导致服务启动失败。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

步骤6 单击“保存”，并在确认对话框中单击“确定”。

等待界面提示“操作成功”，单击“完成”，配置已修改。

说明

- 更新Yarn服务队列的配置且不重启服务时，选择“更多 > 刷新队列”更新队列使配置生效。
- 配置Flume参数“flume.config.file”时，支持“上传文件”和“下载文件”功能。上传配置文件后旧文件将被覆盖，再下载文件只能获取新文件。如果未保存配置并重启服务，那么新文件设置未生效，请及时保存配置。
- 修改服务配置参数后如需重启服务使配置生效，可在服务页面单击右上角“更多 > 重启服务”。
- 配置参数前包含图标时，表示该参数为动态生效，保存配置后参数值会自动刷新到配置文件中。（动态生效图标仅MRS 3.2.0及之后版本支持）

----结束

11.3.2.3.2 修改服务自定义配置参数

操作场景

MRS集群各个组件支持开源的所有参数，其中部分关键使用场景的参数支持在FusionInsight Manager界面进行修改，且部分组件的客户端可能不包含开源特性的所有参数。如果需要修改其他Manager未直接支持的组件参数，集群管理员可以在Manager通过自定义配置项功能为组件添加新参数。添加的新参数最终将保存在组件的配置文件中并在重启后生效。

对系统的影响

- 配置服务属性后，如果服务状态为“配置过期”，则需要重启此服务，重启期间该服务不可用。
- 修改服务配置参数并重启生效后，需要重新下载并安装客户端，或者下载配置文件刷新客户端。

前提条件

集群管理员已充分了解需要新添加的参数意义、生效的配置文件以及对组件的影响。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中指定的服务名称。

步骤4 选择“配置 > 全部配置”。

步骤5 在左侧导航栏定位到某个一级节点，并选择“自定义”，Manager将显示当前组件的自定义参数。

“参数文件”显示保存集群管理员新添加的自定义参数的配置文件。每个配置文件中可能支持相同名称的开源参数，设置不同参数值后生效结果由组件加载配置文件的顺序决定。自定义参数支持服务级别与角色级别，请根据业务实际需要选择。不支持单个角色实例添加自定义参数。

步骤6 在对应参数项所在行“名称”列输入组件支持的参数名，在“值”列输入此参数的参数值。

支持单击“+”或“-”增加或删除一条自定义参数。

步骤7 单击“保存”，在弹出的“保存配置”窗口中确认修改参数，单击“确定”。界面提示“操作成功。”，单击“完成”，配置保存成功。

保存完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

任务示例（配置 Hive 自定义参数）

Hive依赖于HDFS，默认情况下Hive访问HDFS使用的是HDFS的客户端，生效的配置参数统一由HDFS控制。例如HDFS参数“ipc.client.rpc.timeout”影响所有客户端连接HDFS服务端的RPC超时时间，如果集群管理员需要单独修改Hive连接HDFS的超时时间，可以使用自定义配置项功能进行设置。在Hive的“core-site.xml”文件增加此参数可被Hive服务识别并代替HDFS的设置。

步骤1 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务”。

步骤2 选择“Hive > 配置 > 全部配置”。

步骤3 在左侧导航栏选择Hive服务级别“自定义”，Manager将显示Hive支持的服务级别自定义参数。

步骤4 在“core-site.xml”对应参数“core.site.customized.configs”的“名称”输入“ipc.client.rpc.timeout”，“值”输入新的参数值，例如“150000”。单位为毫秒。

步骤5 单击“保存”，在弹出的“保存配置”窗口中确认修改参数并单击“确定”。界面提示“操作成功。”，单击“完成”，配置保存成功。

保存完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

11.3.3 管理实例

11.3.3.1 实例管理概述

总览

登录FusionInsight Manager以后，例如选择“集群 > *待操作集群的名称* > 服务 > KrbServer > 实例”，进入实例管理页面，包含功能区和角色实例列表。

功能区

在功能区勾选需要操作的实例后，可对角色实例执行相关维护管理任务，例如启动或停止实例等，主要操作如**表11-12**所示。

表 11-12 实例维护管理功能

操作入口	说明
“ 启动实例”	将集群中指定实例启动。适用于操作状态为“未启动”、“停止失败”或“启动失败”角色实例，以使用该角色实例。
“ 更多 > 停止实例”	将集群中指定实例停止。适用于不再使用或异常的角色实例。
“ 更多 > 重启实例”	将集群中指定实例重启。适用于状态异常的角色实例，以恢复角色实例功能。
“ 更多 > 滚动重启实例”	为集群中指定实例提供不中断业务的重启操作，具体参数配置可参考**滚动重启集群**。
“ 更多 > 入服/退服”	为集群中指定实例执行入服务或退服的操作，变更实例的业务可用状态方式，具体可参考**入服与退服实例**。 **说明** 仅HDFS的角色DataNode、Yarn的角色NodeManager、HBase的角色RegionServer支持此操作。
“*待操作实例名称* > 更多 > 同步配置”	当某个角色实例的“配置状态”为“配置过期”，表示该角色实例修改配置后还未重启生效，新的配置仅保存在FusionInsight Manager。将新的配置下发至指定实例。 **说明** • 同步角色实例配置后需要重启配置过期的角色实例。重启时对应的角色实例不可用。 • 完成同步配置后，完成后请重启实例以使配置生效。

操作入口	说明
“*待操作实例名称* > 实例配置”	具体请参考**管理实例配置**。

功能区支持按角色或运行状态进行快速筛选。

说明

单击“高级搜索”，支持指定其他筛选条件搜索指定的实例，例如主机名称、管理IP、业务IP和实例组等。

角色实例列表

角色实例列表包含了该服务中所有的角色在集群中的实例情况，列表可显示每个实例的运行状态、配置状态、实例对应的主机以及相关的IP地址信息等。

表 11-13 实例运行状态

状态	说明
良好	表示实例当前运行正常。
故障	表示实例当前无法正常工作。
已退服	表示实例处于退服状态。
未启动	表示实例已停止。
未知	表示实例的初始状态信息无法检测。
正在启动	表示实例正在执行启动过程。
正在停止	表示实例正在执行停止过程。
正在恢复	表示实例可能存在异常正在自动修复。
正在退服	表示实例正在执行退服过程。
正在入服	表示实例正在执行入服过程。
启动失败	表示实例启动操作失败。
停止失败	表示实例停止操作失败。

实例详情

单击实例名称可进入实例详情页面，可查看实例基本信息、配置文件、实例日志以及该实例相关的监控指标图表。

11.3.3.2 入服与退服实例

操作场景

部分角色实例以分布式并行工作的方式对外部业务提供服务，服务会单独保存每个实例是否可以使用的信息，所以需要使用FusionInsight Manager为这些实例执行入服或退服的操作，变更实例的业务可用状态方式。

不支持该此功能的实例，默认无法执行任务。

说明

当前支持退服和入服操作的角色有：HDFS的DataNode、Yarn的NodeManager、HBase的RegionServer。

- 当DataNode数量少于或等于HDFS的副本数时，不能执行退服操作。若HDFS副本数为3时，则系统中少于4个DataNode，将无法执行退服，Manager在执行退服操作时会等待30分钟后报错并退出执行。
- 由于Mapreduce任务执行时，会生成一些副本数为10的文件，此时若DataNode实例数少于10时，将无法进行退服操作。
- 如果退服前，DataNode节点的机架数（机架数由各DataNode节点所配置的“机架”的名称数量决定）大于1；而退服部分DataNode后，剩余的DataNode节点的机架数变为1，则此次退服将会失败。所以需要在退服前评估退服操作对机架数的影响，以调整退服的DataNode节点。
- 在退服多个DataNode时，如果每个DataNode存储的数据量较大，如果执行选择多个DataNode同时退服，则很有可能会因超时而退服失败。为了避免这种情况，建议每次退服仅退服1个DataNode，进行多次退服操作。

操作步骤

步骤1 DataNode节点退服前需要进行健康检查，步骤如下：

1. 使用客户端用户登录客户端安装节点，并切换到客户端安装目录。
2. 如果是安全集群，需要使用**hdfs**用户进行权限认证。

```
source bigdata_env          #配置客户端环境变量
kinit hdfs                  #设置kinit认证
Password for hdfs@HADOOP.COM:    #输入hdfs用户登录密码
```

3. 执行**hdfs fsck / -list-corruptfileblocks**，检查返回结果。
   - 如果结果是“...has 0 CORRUPT files”，执行**步骤2**。
   - 如果结果不是“...has 0 CORRUPT files”，并返回损坏的文件名称，执行**步骤1.4**。
4. 执行**hdfs dfs -rm** *损坏的文件名称*，删除损坏的文件。

   说明

   删除文件为高危操作，在执行操作前请务必确认对应文件是否不再需要。

步骤2 登录FusionInsight Manager。

步骤3 选择“集群 > *待操作集群的名称* > 服务”。

步骤4 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤5 勾选指定的待退服角色实例。

步骤6 在“更多”选择“退服”或“入服”。

输入当前登录的用户密码确认身份，单击“确定”。

勾选“我确定退服这些实例，并接受服务性能下降的结果。”，单击“确定”，执行相应的操作。

说明

实例退服操作未完成时在其他浏览器或窗口重启集群中实例对应的服务，FusionInsight Manager将提示停止退服，实例的“操作状态”显示为“启动”。实际上后台已将该实例退服，请重新执行退服操作同步状态。

----结束

11.3.3.3 管理实例配置

操作场景

每个单独的角色实例可以修改配置参数在迁移实例到新集群场景或者重新部署相同服务的场景下，集群管理员可以在FusionInsight Manager中将某服务所有配置数据导入或者导出，实现配置结果的快速复制。

FusionInsight Manager支持管理单个角色实例的配置参数，修改配置参数、导出实例配置或导入实例配置时不影响其他实例。

对系统的影响

修改角色实例配置后，如果实例状态为“配置过期”，则需要重启此实例。重启时对应的实例不可用。

修改实例配置

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作的集群名称* > 服务”。

步骤3 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤4 单击指定的实例，选择“实例配置”。

默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该实例支持的所有参数分类。

步骤5 在导航树选择指定的参数分类，并在右侧修改参数值。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

步骤6 单击“保存”，并在确认对话框中单击“确定”。

等待界面提示“操作成功”，单击“完成”，配置已修改。

说明

角色实例配置参数修改后，如果实例状态为“配置过期”则需要重启此实例，可在“实例”页面勾选“配置过期”的实例，选择“更多 > 重启实例”。

配置参数前包含 图标时，表示该参数为动态生效，保存配置后参数值会自动刷新到配置文件中。（动态生效图标 仅MRS 3.2.0及之后版本支持）

----结束

导出导入实例配置

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中待操作的服务名称，并选择“实例”页签。

步骤4 单击待操作的实例，选择“实例配置”。

步骤5 单击“导出”，导出配置参数文件到本地。

步骤6 在实例配置页面单击“导入”，在弹出的配置文件选择框中定位到实例的配置参数文件，即可导入所有配置。

----结束

11.3.3.4 查看实例配置文件

操作场景

FusionInsight Manager支持在管理页面上直接查看实例节点上实际的环境变量、角色配置等配置文件内容，运维人员在需要快速排查实例对应配置项是否配置错误或者需要查看部分隐藏类型的配置项时，可直接在FusionInsight Manager上进行查看，帮助用户快速分析配置问题。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作的集群名称* > 服务”。

步骤3 单击服务视图中指定的服务名称，并选择“实例”页签。

步骤4 单击需要查看配置的实例名称，在概览页面的“配置文件”区域内，系统会显示该实例相关的配置文件列表。

步骤5 单击要查看的配置文件的名称，可查看配置文件内具体的配置参数值内容。

如需获取该配置文件，可单击“下载至本地”按钮，将该配置文件内容下载到本地PC。

说明

集群内的节点故障时，将无法查看配置文件，请修复故障的节点后再查看。

----结束

11.3.3.5 实例组

11.3.3.5.1 管理实例组

操作场景

FusionInsight Manager支持对多个实例组的管理功能，即用户可以按照具有相同硬件配置的节点或者其他原则将同一角色内的多个实例进行分组。针对实例组进行的配置参数修改，将同时对组内所有的实例生效。

在大集群场景中，通过实例组将提升大集群下异构环境批量实例的管理能力，分配好实例组后，后续可反复配置，减少实例配置项的冗余，提升系统性能。

创建实例组

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中待操作的服务名称。

步骤4 选择“实例组”。

单击 ，按照界面提示填写参数。

表 11-14 实例组配置参数

参数名	说明
组名称	实例组名称只能包含字母、数字、下划线（_）、中划线（-）和空格，仅以字母、数字、下划线（_）或中划线（-）开头，只能在中间包含空格，并且不能超过99个字符。
角色	表示实例组包含哪个角色的实例。
复制源	指从指定的实例组复制配置值到新组，若为空，则新组对应的各配置值为系统默认值。
描述	只可以包含汉字、英文字母、数字、中英文逗号、中英文句号、下划线（_）、空格和换行符，并且不能超过200个字符。

说明

- 每个实例必须且只能属于一个实例组，实例首次安装时默认属于的实例组为“*角色名*-DEFAULT”。
- 多余或者不再使用的实例组可以删除，删除前需要将组内的实例全部迁移至其他实例组，然后参照**删除实例组**对实例组进行删除，系统默认的实例组不可删除。

步骤5 单击“确定”完成创建实例组。

----结束

修改实例组属性

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中待操作的服务名称。

步骤4 在“实例组”页签定位到指定的实例组。

单击 ，按照界面提示修改参数。

步骤5 单击“确定”完成修改。

默认实例组不支持修改。

----结束

删除实例组

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中待操作的服务名称。

步骤4 在“实例组”页签定位到指定的实例组。

步骤5 单击 。

步骤6 在弹出窗口单击“确定”。

默认实例组不支持删除。

----结束

11.3.3.5.2 查看实例组信息

操作场景

集群管理员可以在FusionInsight Manager查看指定服务的实例组。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中待操作的服务名称。

步骤4 单击“实例组”。

步骤5 在导航栏选择一个角色，在“基本”页签，查看该实例组的全部实例。

说明

需要将某个实例从一个实例组移动到另一个实例组中时，可以根据以下操作：

1. 勾选需要移动到新实例组的实例，然后单击“移动”。
2. 在弹出窗口选择一个目标的实例组。
   迁移时将自动继承新实例组的配置，如果该实例之前修改过配置，将以自身的配置优先。
3. 单击“确定”。
   完成后请重新启动配置过期的服务或实例以使配置生效。

----结束

11.3.3.5.3 配置实例组参数

操作场景

在大集群场景中，用户可以在FusionInsight Manager通过实例组可批量配置多个实例的参数，减少实例配置项的冗余，提升系统性能。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务视图中待操作的服务名称。

步骤4 选择“实例组”。

步骤5 在导航栏选择一个角色的实例组名称，切换至“配置”页签，调整需要修改的配置参数后单击“保存”，相关配置将对实例组内所有实例生效。

----结束

11.4 主机

11.4.1 主机管理页面

11.4.1.1 查看主机列表

总览

登录FusionInsight Manager以后，单击“主机”后，打开主机管理页面，可查看主机列表及基本信息。

用户可切换视图类型，以及设置条件筛选和搜索主机。

切换视图

单击“角色视图”，可直观查看各主机上当前已部署的角色。如果该角色支持主备模式，则角色名称显示为加粗。

主机列表

主机管理页面的主机列表包含了所有集群中所有主机，并支持对主机进行相关运维操作。

在主机管理页面，可通过节点类型或所属集群筛选主机，对主机类型的筛选规则为：

- 管理节点为部署了OMS的节点，同时管理节点上也可能部署控制角色和数据角色。
- 控制节点为部署控制角色的节点，同时控制节点上也可能部署数据角色。
- 数据节点为仅部署数据角色的节点。

系统默认为“主机视图”，可显示每个主机的IP地址信息、机架规划信息、运行状态、所归属集群以及硬件资源等使用情况。

表 11-15 主机运行状态

状态	说明
良好	表示主机当前状态正常。
故障	表示主机当前无法正常工作。
未知	表示主机的初始状态信息无法检测。
已隔离	表示主机处于隔离的状态。
已停机	表示主机处于停机的状态。

11.4.1.2 查看主机概览

总览

登录FusionInsight Manager以后，单击“主机”，在主机列表单击指定的主机名称，可以访问主机详情页面，主要包含基本信息区、磁盘状态区、角色列表区和监控图表等。

基本信息区

主机详情页面的基本信息包含该主机的各个关键信息，例如管理IP地址、业务IP地址、主机类型、机架、防火墙、CPU核数、操作系统等信息。

磁盘状态区

磁盘状态区包含了该主机所有为集群配置的磁盘分区，并显示每个磁盘分区的使用情况。

实例列表区

实例列表区显示了该主机所有安装的角色实例，并显示每个角色实例的状态，单击角色实例名称后的日志文件，可在线查看该实例对应日志文件内容。

告警和事件的历史记录

告警和事件的历史记录区显示了当前主机上报的关键告警与事件记录，系统最多可显示20条历史记录。

图表

主机详情页面的右侧展示图表区，包含该主机的各个关键监控指标报表。

用户可以单击右上角的“ ▼ > 定制”，自定义在图表区展示的监控报表。选择时间区间后，单击“ ▼ > 导出”，可以导出指定时间区间内的详细监控指标数据。

单击监控指标标题后的可以打开监控指标的解释说明。

单击主机的“图表”页签，可直接查看该主机的全量监控图表信息。

GPU 卡状态区

主机有配置GPU卡时，GPU卡状态区显示了当前主机安装的GPU卡型号、位置及状态信息。

11.4.1.3 查看主机进程及资源

总览

登录FusionInsight Manager页面，单击“主机”，在主机列表中选择指定的主机名称，进入主机详情页面，单击“进程”和“资源”页签进入相关页面。

主机进程

进程页面显示了当前主机上已部署服务实例的角色进程信息，例如进程状态、PID、进程运行时间等，并可直接在线查看各进程的日志文件内容。

主机资源

主机资源页面显示了当前主机上已部署服务实例的详细资源使用情况，包括CPU，内存，磁盘和端口情况。

11.4.2 主机维护操作

11.4.2.1 启动、停止主机上的所有实例

操作场景

当主机发生故障状态异常时，用户可能需要停止主机上的所有角色，对主机进行维护检查。故障清除后，启动主机上的所有角色恢复主机业务。Manager支持在主机管理页面或者主机详情页面进行相关操作，以下根据主机管理页面为例进行指导。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待操作主机前的复选框。

步骤4 在“更多”选择“启动所有实例”或“停止所有实例”执行相应操作。

----结束

11.4.2.2 执行主机健康检查

操作场景

如果某个主机节点的运行状态不是良好，用户可以执行主机健康检查，快速确认某些基本功能是否存在异常。在日常运维中，管理员也可以执行主机健康检查，以保证主机上各角色实例的配置参数以及监控没有异常、能够长时间稳定运行。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待操作主机前的复选框。

步骤4 在“更多”选择“健康检查”启动任务。

健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。如果发现问题，可以单击“查看帮助”。

----结束

11.4.2.3 分配机架

操作场景

大型集群的所有主机通常分布在多个机架上，不同机架间的主机通过交换机进行数据通信，且同一机架上的不同机器间的网络带宽要远大于不同机架机器间的网络带宽。在这种情况下网络拓扑规划应满足以下要求：

- 为了提高通信速率，希望不同主机之间的通信能够尽量发生在同一个机架之内，而不是跨机架。
- 为了提高容错能力，分布式服务的进程或数据需要尽可能存在多个机架的不同主机上。

Hadoop使用一种类似于文件目录结构的方式来表示主机。

由于HDFS不能自动判断集群中各个DataNode的网络拓扑情况，管理员需设置机架名称来确定主机所处的机架，NameNode才能绘出DataNode的网络拓扑图，并尽可能将DataNode的数据备份在不同机架中。同理，YARN需要获取机架信息，在可允许的范围内将任务分配给不同的NodeManager执行。

当集群网络拓扑发生变化时，需要使用FusionInsight Manager为主机重新分配机架，相关服务才会自动调整。

对系统的影响

修改主机机架名称，将影响HDFS的副本存放策略、Yarn的任务分配及Kafka的Partition存储位置。修改后需重启HDFS、Yarn和Kafka，使配置信息生效。

不合理的机架配置会导致集群的节点之间的负载（包括CPU、内存、磁盘、网络）不平衡，降低集群的可靠性，影响集群的稳定运行。所以在分配机架之前，需要进行全局的统筹，合理地设置机架。

机架分配策略

说明

物理机架：主机所在的真实的机架。

逻辑机架：在FusionInsight Manager中给主机设置的机架名称。

策略 1：每个逻辑机架包含的主机个数基本一致。

策略 2：主机所设置的逻辑机架要尽量符合其所在的物理机架。

策略 3：如果一个物理机架的主机个数很少，则需要和其他的主机较少的物理机架合并为一个逻辑机架，以满足策略1。不能将两个机房的主机合并为一个逻辑机架，否则会引起性能问题。

策略 4：如果一个物理机架的主机个数很多，则需要将其分隔为多个逻辑机架，以满足策略1。不建议物理机架中包含的主机有太大的差异，这样会降低集群的可靠性。

策略 5：建议机架的第一层为默认的“default”或其他值，但在集群中保持一致。

策略 6：每个机架所包含的主机个数不能小于3。

策略 7：一个集群的逻辑机架数，不建议多于50个（过多则不便于维护）。

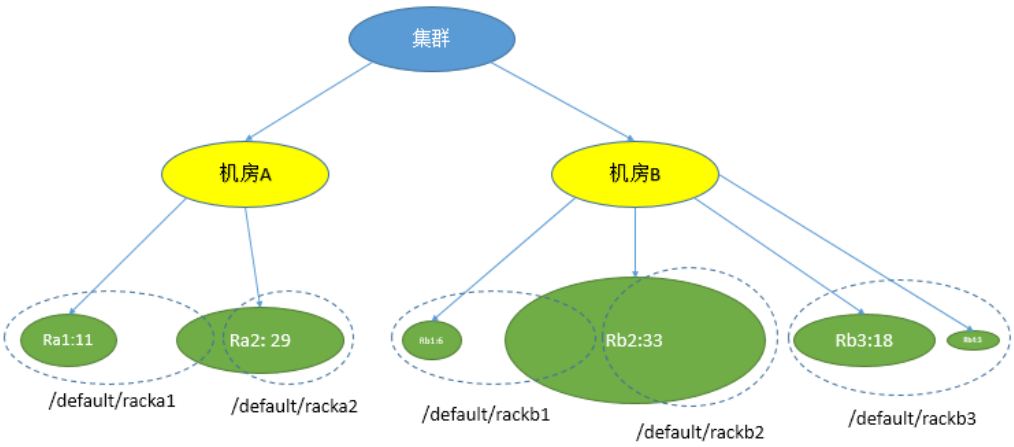
最佳实践示例

假设一个集群，共有主机100台，分别在两个机房中：机房A有40台主机，机房B有60台主机。在机房A中，物理机架Ra1有11台主机，物理机架Ra2有29台。在机房B中，物理机架Rb1有6台主机，物理机架Rb2有33台主机，物理机架Rb3有18台主机，物理机架Rb4有3台主机。

根据以上的“机架分配策略”，设置每个逻辑机架包含20个主机，具体分配如下：

- 逻辑机架 /default/racka1: 包含物理机架Ra1的11台主机，Ra2的9台主机。
- 逻辑机架 /default/racka2: 包含物理机架Ra2的剩余的20台主机。
- 逻辑机架 /default/rackb1: 包含物理机架Rb1的6台主机，Rb2的13台主机。
- 逻辑机架 /default/rackb2: 包含物理机架Rb2的剩余的20台主机。
- 逻辑机架 /default/rackb3: 包含物理机架Rb3的18台主机，Rb4的3台主机。

机架划分示例如下：



操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待操作主机前的复选框。

步骤4 在“更多”选择“设置机架”。

- 机架名称需遵循实际网络拓扑结构，以层级形式表示；各层级间以斜线“/”隔开。
- 机架命名规则为：“/level1/level2/…”，级别至少为一级，名称不能为空。机架名称由字母、数字及下划线“_”组成，且总长度不超过200个字符。

  例如“/default/rack0”。
- 如果待修改机架中所包含的主机中有DataNode实例，请确保所有DataNode实例所在主机的机架名称的层级一致。否则，会导致配置下发失败。

步骤5 单击“确定”，完成机架分配设置。

----结束

11.4.2.4 隔离主机

操作场景

某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。

说明

隔离主机仅支持隔离非管理节点。

对系统的影响

- 主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。
- 主机隔离后部分服务的实例不再工作，服务的配置状态可能过期。
- 主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。
- 待操作节点的SSH端口需保持默认（22），否则将导致本章节任务操作失败。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 勾选待隔离主机前的复选框。

步骤4 在“更多”选择“隔离”。

在弹出窗口中，输入当前登录的用户密码确认管理员身份，单击“确定”。

步骤5 在确认隔离的对话框中勾选“我确定隔离所选主机，接受可能出现的服务故障等后果。”单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功隔离，“运行状态”显示为“已隔离”。

步骤6 以**root**用户登录到被隔离主机上，执行**pkill -9 -u omm**命令终止节点上的omm用户的进程，然后执行**ps -ef | grep 'container' | grep '${BIGDATA_HOME}' | awk '{print $2}' | xargs -I '{}' kill -9 '{}'** 命令查找并终止container的进程。

步骤7 管理员已排除主机的异常或故障后，需要将主机隔离状态取消才能继续使用该主机。

在“主机”界面勾选已隔离的主机，选择“更多 > 取消隔离”。

说明

取消隔离后，主机上所有角色实例默认不启动。若需要启动主机上角色实例，可以在“主机”页面勾选目标主机，然后选择“更多 > 启动所有实例”。

----结束

11.4.2.5 导出主机信息

操作场景

管理员可以在FusionInsight Manager导出所有主机的信息。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 单击“主机”。

步骤3 在右上角的下拉菜单中选择所需主机的类型，也可以通过“高级搜索”进一步筛选所需主机。

步骤4 单击“导出全部”，在“保存类型”选择“TXT”或“CSV”。单击“确定”开始导出。

----结束

11.4.3 资源概况

11.4.3.1 分布

登录到FusionInsight Manager页面，选择“主机 > 资源概况 > 分布”，可查看各集群的资源分布监控，如图11-1所示，默认显示1小时的监控数据。用户可单击自定义时间区间，缺省时间区间包括：1小时、2小时、6小时、12小时、1天、1周、1月。

图 11-1 资源分布概况

- 单击“选择指标”可以自定义所需查看的指标项，详细指标项如表11-16所示。选择指标后，页面会显示在各个区间的主机分布图。

- 鼠标停留在某个色块时，会显示处于当前区间的主机数量，如图11-1所示。单击色块，页面会显示处于当前区间的主机列表。
  - 单击列表中某主机“主机名称”，会跳转至该主机的详细信息页面；
  - 单击列表中某主机“查看趋势”，会显示当前指标项整个集群的最大值、平均值、最小值、当前主机值。当前集群中，当指标为“主机CPU-内存-磁盘使用率”时，不能进行“查看趋势”操作。
- 单击“导出数据”，可以导出当前指标项集群中所有节点在选中的时间区域内的最大值、最小值、平均值。

表 11-16 指标项

指标分类	指标项
进程	• 运行的进程总数 • 进程总数 • omm进程总数 • D状态进程总数
网络状态	• 主机网络数据包冲突数 • LAST_ACK状态数量 • CLOSING状态数量 • LISTENING状态数量 • CLOSED状态数量 • ESTABLISHED状态数量 • SYN_RECV状态数量 • TIME_WAITING状态数量 • FIN_WAIT2状态数量 • FIN_WAIT1状态数量 • CLOSE_WAIT状态数量 • DNS解析时长 • TCP临时端口使用率 • 主机网络数据包帧错误数
网络读信息	• 主机网络读包数 • 主机网络读包丢包数 • 主机网络读包错误数 • 主机网络接收速率
磁盘	• 主机磁盘写速率 • 主机磁盘已使用大小 • 主机磁盘未使用大小 • 主机磁盘读速率 • 主机磁盘使用率

指标分类	指标项
内存	• 未使用内存 • 缓存内存大小 • 内核缓存的内存总量 • 共享内存大小 • 主机内存使用率 • 已使用内存
网络写信息	• 主机网络写包数 • 主机网络写包错误数 • 主机网络发送速率 • 主机网络写包丢包数
CPU	• 改变过优先级的进程占CPU的百分比 • 用户空间占用CPU百分比 • 内核空间占用CPU百分比 • 主机CPU使用率 • CPU总时间 • CPU闲置时间
主机状态	• 主机文件句柄使用率 • 每1分钟系统平均负载 • 每5分钟系统平均负载 • 每15分钟系统平均负载 • 主机PID使用率

11.4.3.2 趋势

选择“主机 > 资源概况 > 趋势”，可查看所有集群或者单个集群的资源趋势监控页面，如图11-2所示。默认显示1小时的监控数据。用户可单击自定义时间区间，缺省时间区间包括：1小时、2小时、6小时、12小时、1天、1周、1月。各指标趋势图默认显示整个集群的最大值、最小值、平均值。

图 11-2 资源趋势

- 单击“为图表添加主机”，可在定制显示的趋势指标图中，添加个别节点的指标趋势线，最多可添加12个主机。
- 单击 ▼ ，选择“定制”，可以自定义需要在页面上显示的指标项，详细指标项参考分布中表11-16。
- 选择“导出数据”，可以导出集群中所有节点，在所有选中的指标项下，选中时间范围内的最大值、最小值、平均值。

11.4.3.3 集群

选择“主机 > 资源概况 > 集群”，可同时查看FusionInsight Manager内各集群的资源监控页面，如图11-3所示。

系统默认显示1小时的监控数据。用户可单击 自定义时间区间，缺省时间区间包括：1小时、2小时、6小时、12小时、1天、1周、1月。

图 11-3 集群资源概况

- 单击“指定集群”，可定制需要显示的集群。
- 单击 ▼ ，选择“定制”，可以自定义需要在页面上显示的指标项，详细指标项参考分布中表11-16。

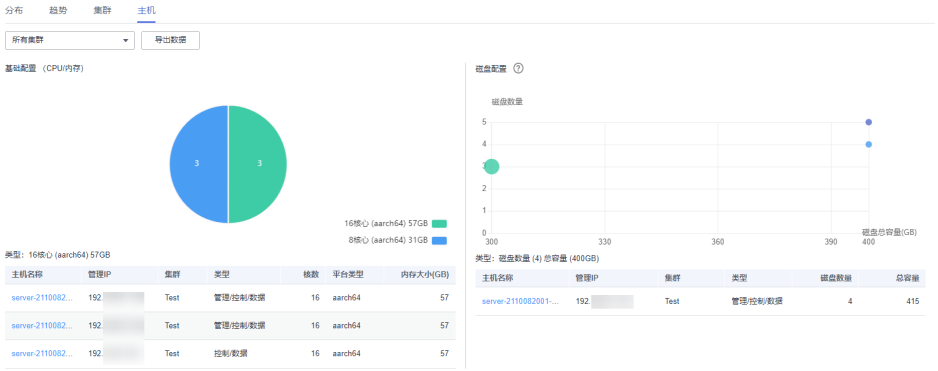
- 选择“导出数据”，可以导出各个集群在所有选中的指标项下，选中时间范围内的指标值。

11.4.3.4 主机

选择“主机 > 资源概况 > 主机”，可查看主机资源概况，分为基础配置（CPU/内存）和磁盘配置两部分，如图11-4所示。

单击“导出数据”，可导出集群中所有主机的配置列表，包括主机名称、管理IP、主机类型、核数、平台类型、内存容量、磁盘大小等。

图 11-4 主机资源概况



基础配置（CPU/内存）

鼠标放置饼图上会显示当前区域集群中各节点不同硬件配置下的配置信息及数量，格式为：*核数（平台类型）内存大小：数量*。

单击相应区域，会在下方显示相应的主机列表。

磁盘配置

横轴为节点上磁盘总容量（包含OS盘），纵轴为逻辑磁盘数量（包含OS盘）。

鼠标放置圆点上会显示处于当前配置状态下的磁盘信息，包括磁盘数量、总容量、主机数。

单击相应圆点，会在下方显示相应的主机列表。

11.5 运维

11.5.1 告警

11.5.1.1 告警与事件概述

告警

登录FusionInsight Manager，选择“运维 > 告警 > 告警”，用户可以查看FusionInsight Manager中各集群上报的告警信息，包括告警名称、ID、级别、产生时间等信息，默认每页显示最近的十条告警。

单击指定告警名称左侧的，展开完整告警信息参数，各项说明如表11-17所示。

表 11-17 告警参数

告警参数	说明
告警ID	告警信息的ID。
告警名称	告警信息的名称。
告警级别	包含紧急、重要、次要、提示四项级别。
产生时间	产生告警的时间。
清除时间	告警检测到已清除的时间。如果未清除，则显示为“--”。
来源	集群名称。
对象	触发告警的服务、进程或模块。
是否自动清除	能够在问题修复后自动清除告警。
告警状态	告警当前状态，包含自动清除、手动清除、未清除。
告警原因	告警可能的原因提示。
序列号	系统产生的告警计数。
附加信息	相关报错信息。
定位信息	定位告警的详细信息。主要包含以下信息： ● 来源：产品告警的集群 ● 服务名：产生告警的服务名称 ● 角色名：产生告警的角色名称 ● 主机名：产生告警的主机名

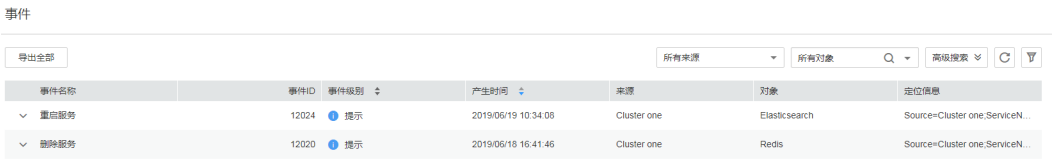
管理告警

- 单击“导出全部”可导出全部告警详情。
- 如果有多个告警已完成处理，可选中一个或多个待清除的告警，单击“清除告警”，批量清除告警。每次最多批量清除300条告警。
- 单击手动刷新当前页面，也可单击 修改告警表格显示的列。
- 支持通过指定对象或集群来筛选指定的告警。
- 单击“高级搜索”显示告警搜索区域，搜索条件包括告警ID、告警名称、告警状态、告警级别、开始时间和结束时间。单击“搜索”显示过滤后的告警，再次单击“高级搜索”，会显示已经填写的搜索条件数量。
- 单个告警支持“清除”、“屏蔽”以及“查看帮助”操作。
- 告警条目较多时，可单击“归类视图”，系统会将未恢复的告警按照告警ID进行归类，方便用户查看。归类后单击告警名称后的未恢复条数，即可查看具体的告警详情。

事件

登录FusionInsight Manager，选择“运维 > 告警 > 事件”，进入事件界面，用户可以查看集群中所有事件信息，包括名称、ID、级别、产生时间、来源、对象、定位信息，每页默认显示最近的十条事件。

图 11-5 FusionInsight Manager 事件管理



单击指定事件名称左侧的，展开完整信息参数，各项说明如表11-18所示。

表 11-18 事件参数

事件参数	说明
事件ID	事件信息的ID。
事件名称	事件信息的名称。
事件级别	包含紧急、重要、次要、提示共4项级别。
产生时间	事件产生的时间。
对象	事件可能的原因提示。
序列号	系统产生的事件计数。
定位信息	定位事件的详细信息。主要包含以下信息： ● 来源：产生事件的集群名称 ● 服务名：产生事件的服务名称 ● 角色名：产生事件的角色名称 ● 主机名：产生事件的主机名
附加信息	相关报错信息。
事件原因	事件可能的原因提示。
来源	集群名称。

管理事件：

- 单击“导出全部”可导出全部事件详情。
- 单击手动刷新当前页面，也可在 修改事件表格显示的列。
- 支持通过指定对象或集群来筛选指定的事件。
- 单击“高级搜索”显示事件搜索区域，搜索条件包括事件ID、事件名称、事件级别、开始时间和结束时间。

11.5.1.2 配置阈值

操作场景

FusionInsight Manager支持配置监控指标阈值用于关注各指标的健康情况，如果出现异常的数据并满足预设条件后，系统将会触发一条告警信息，并在告警页面中出现此告警信息。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 告警 > 阈值设置”。

步骤3 在监控分类中选择集群内指定主机或服务的监控指标。

图 11-6 配置指标阈值

例如“主机内存使用率”，界面显示此阈值的信息：

- 发送告警开关指示为 表示将触发告警。
- “告警ID”和“告警名称”包含阈值将触发的告警信息。
- Manager会检查监控指标数值是否满足阈值条件，若连续检查且不满足的次数等于“平滑次数”设置的值则发送告警，支持自定义。
- “检查周期（秒）”表示Manager检查监控指标的时间间隔。
- 规则列表中的条目为触发告警的规则。

步骤4 单击“添加规则”，可以新增指标的监控行为。

表 11-19 监控指标规则参数

参数名	参数解释	参数值
规则名称	规则名称	CPU_MAX（举例）
告警级别	告警级别 • 紧急 • 重要 • 次要 • 提示	• 紧急 • 重要 • 次要 • 提示

参数名	参数解释	参数值
阈值类型	选择某指标的最大值或最小值，类型为“最大值”表示指标的实际值大于设置的阈值时系统将产生告警，类型为“最小值”表示指标的实际值小于设置的阈值时系统将产生告警。	• 最大值 • 最小值
日期	设置规则生效的日期，即哪一天运行规则。	• 每天 • 每周 • 其他
添加日期	仅在“日期”模式为“其他”时可见，设置规则运行的自定义日期，支持多选。	09-30
阈值设置	设置规则运行的具体时间范围。	起止时间：00:00-8:30
	设置规则监控指标的阈值	阈值：10

说明

支持单击或设置多个阈值时间条件。

步骤5 单击“确定”保存规则。

步骤6 在新添加规则所在的行，单击“操作”中的“应用”，此时规则的“生效状态”变成“生效”。

当前已创建的规则单击“取消应用”后，才能应用新规则。

----结束

监控指标参考

FusionInsight Manager转告警监控指标可分为节点信息指标与集群服务指标。**表11-20**表示节点中可配置阈值的指标。

表 11-20 节点信息监控指标转告警列表

监控指标组名称	监控指标名称	指标含义	默认阈值
CPU	主机CPU使用率	描述周期内当前集群的运算和控制能力，可通过观察该统计值，了解集群整体资源的使用情况。	90.0%
磁盘	磁盘使用率	描述主机磁盘的使用率。	90.0%
	磁盘inode使用率	统计采集周期内磁盘inode使用率。	80.0%
内存	主机内存使用率	统计当前时间点的内存平均使用率。	90.0%
主机状态	主机文件句柄使用率	统计采集周期内该主机的文件句柄使用率。	80.0%
	主机PID使用率	主机PID使用率。	90%
网络状态	TCP临时端口使用率	统计采集周期内该主机的TCP临时端口使用率。	80.0%
网络读信息	读包错误率	统计采集周期内该主机上该网口的读包错误率。	0.5%
	读包丢包率	统计采集周期内该主机上该网口的读包丢包率。	0.5%
	读吞吐率	统计周期内网口的平均读吞吐率（MAC层）。	80%
网络写信息	写包错误率	统计采集周期内该主机上该网口的写包错误率。	0.5%
	写包丢包率	统计采集周期内该主机上该网口的写包丢包率。	0.5%
	写吞吐率	统计周期内网口的平均写吞吐率（MAC层）。	80%
进程	D状态进程总数	统计周期内主机上D状态进程数量。	0

监控指标组名称	监控指标名称	指标含义	默认阈值
	omm进程使用率	统计周期内omm进程使用率。	90

表 11-21 集群监控指标转告警列表

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
DBService	数据库	数据库连接数使用率	数据库连接数使用率统计。	90%
		数据目录磁盘空间使用率	数据目录磁盘空间使用率统计。	80%
Flume	Agent	Flume堆内存使用率	Flume堆内存使用百分比统计。	95.0%
		Flume直接内存使用率	Flume直接内存使用百分比统计。	80.0%
		Flume非堆内存使用率	Flume非堆内存使用百分比统计。	80.0%
		Flume垃圾回收（GC）总时间	Flume垃圾回收（GC）总时间。	12000ms
HBase	GC	GC中回收old区所花时长	RegionServer的总GC时间。	5000ms
		GC中回收old区所花时长	HMaster的总GC时间。	5000ms
	CPU和内存	RegionServer直接内存使用率统计	RegionServer直接内存使用率统计。	90%
		RegionServer堆内存使用率统计	RegionServer堆内存使用率统计。	90%
		HMaster直接内存使用率统计	HMaster直接内存使用率统计。	90%
		HMaster堆内存使用率统计	HMaster堆内存使用率统计。	90%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	服务	单个RegionServer的region数目	单个RegionServer的Region数目。	2000
		处在RIT状态达到阈值时长的region数	处在RIT状态达到阈值时长的region数。	1
	容灾	容灾同步失败次数	同步容灾数据失败次数。	1
		主集群等待同步的日志文件数量	主集群等待同步的日志文件数量。	128
		主集群等待同步的HFile文件数量	主集群等待同步的HFile文件数量。	128
	队列	Compaction操作队列大小	Compaction操作队列大小。	100
HDFS	文件和块	HDFS缺失的块数量	HDFS文件系统中缺少副本块数量。	0
		需要复制副本的块总数	NameNode需要复制副本的块总数。	1000
	RPC	主NameNode RPC处理平均时间	NameNode RPC处理平均时间。	100ms
		主NameNode RPC队列平均时间	NameNode RPC队列平均时间。	200ms
	磁盘	HDFS磁盘空间使用率	HDFS磁盘空间使用率。	80%
		DataNode磁盘空间使用率	HDFS文件系统中DataNode可以使用的磁盘空间率。	80%
		总副本预留磁盘空间所占比率	总副本预留磁盘空间占DataNode总未使用磁盘空间的百分比。	90%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	资源	故障的DataNode总数	出故障的DataNode节点数量。	3
		NameNode非堆内存使用百分比统计	NameNode非堆内存使用百分比统计。	90%
		NameNode直接内存使用百分比统计	NameNode直接内存使用百分比统计。	90%
		NameNode堆内存使用百分比统计	NameNode堆内存使用百分比统计。	95%
		DataNode直接内存使用百分比统计	DataNode直接内存使用百分比统计。	90%
		DataNode堆内存使用百分比统计	DataNode堆内存使用百分比统计。	95%
		DataNode非堆内存使用百分比统计	DataNode非堆内存使用百分比统计。	90%
	垃圾回收	垃圾回收时间统计（GC）	NameNode每分钟的垃圾回收（GC）所占用的时间。	12000ms
		垃圾回收时间统计（GC）	DataNode每分钟的垃圾回收（GC）所占用的时间。	12000ms
Hive	HQL	Hive执行成功的HQL百分比	Hive执行成功的HQL百分比。	90.0%
	Background	Background线程使用率	Background线程使用率。	90%
	GC	MetaStore的总GC时间	MetaStore的总GC时间。	12000ms
		HiveServer的总GC时间	HiveServer的总GC时间。	12000ms

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	容量	Hive已经使用的HDFS空间占可使用空间的百分比	Hive已经使用的HDFS空间占可使用空间的百分比。	85.0%
	CPU和内存	MetaStore直接内存使用率统计	MetaStore直接内存使用率统计。	95%
		MetaStore非堆内存使用率统计	MetaStore非堆内存使用率统计。	95%
		MetaStore堆内存使用率统计	MetaStore堆内存使用率统计。	95%
		HiveServer直接内存使用率统计	HiveServer直接内存使用率统计。	95%
		HiveServer非堆内存使用率统计	HiveServer非堆内存使用率统计。	95%
		HiveServer堆内存使用率统计	HiveServer堆内存使用率统计。	95%
	Session	连接到HiveServer的session数占最大允许session数的百分比	连接到HiveServer的session数占最大允许session数的百分比。	90.0%
Kafka	分区	未完全同步的Partition百分比	未完全同步的Partition数占Partition总数的百分比。	50%
	其他	Partition不可用百分比	Kafka各个Topic的Partition不可用占比。	40%
		broker上用户连接数使用率	broker上用户连接数使用率。	80%
	磁盘	Broker磁盘使用率	Broker数据目录所在磁盘的磁盘使用率。	80.0%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		Broker磁盘IO使用率	Broker数据目录所在磁盘的磁盘IO使用率。	80%
	进程	Broker每分钟的垃圾回收时间统计（GC）	Broker进程每分钟垃圾回收（GC）所占用的时间。	12000ms
		Kafka堆内存使用率	Kafka堆内存使用百分比统计。	95%
		Kafka直接内存使用率	Kafka直接内存使用百分比统计。	95%
Loader	内存	Loader堆内存使用率	Loader堆内存使用率。	95%
		Loader直接内存使用率统计	Loader直接内存使用率统计。	80.0%
		Loader非堆内存使用率	Loader非堆内存使用率。	80%
	GC	Loader的总GC时间	Loader的总GC时间。	12000ms
Mapreduce	垃圾回收	垃圾回收时间统计（GC）	垃圾回收时间统计（GC）。	12000ms
	资源	JobHistoryServer直接内存使用百分比统计	JobHistoryServer直接内存使用百分比统计。	90%
		JobHistoryServer非堆内存使用百分比统计	JobHistoryServer非堆内存使用百分比统计。	90%
		JobHistoryServer堆内存使用百分比统计	JobHistoryServer堆内存使用百分比统计。	95%
Oozie	内存	Oozie堆内存使用率	Oozie堆内存使用率。	95.0%
		Oozie直接内存使用率	Oozie直接内存使用率。	80.0%
		Oozie非堆内存使用率	Oozie非堆内存使用率。	80%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	GC	Oozie垃圾回收（GC）总时间	Oozie垃圾回收（GC）总时间。	12000ms
Spark2x	内存	JDBCServer2x堆内存使用率统计	JDBCServer2x堆内存使用率统计。	95%
		JDBCServer2x直接内存使用率统计	JDBCServer2x直接内存使用率统计。	95%
		JDBCServer2x非堆内存使用率统计	JDBCServer2x非堆内存使用率统计	95%
		JobHistory2x直接内存使用率统计	JobHistory2x直接内存使用率统计。	95%
		JobHistory2x非堆内存使用率统计	JobHistory2x非堆内存使用率统计。	95%
		JobHistory2x堆内存使用率统计	JobHistory2x堆内存使用率统计。	95%
		IndexServer2x直接内存使用率统计	IndexServer2x直接内存使用率统计。	95%
		IndexServer2x堆内存使用率统计	IndexServer2x堆内存使用率统计。	95%
		IndexServer2x非堆内存使用率统计	IndexServer2x非堆内存使用率统计。	95%
	GC次数	JDBCServer2x的Full GC次数	JDBCServer2x进程的Full GC次数。	12
		JobHistory2x的Full GC次数	JobHistory2x进程的Full GC次数。	12
		IndexServer2x的Full GC次数	IndexServer2x进程的Full GC次数。	12

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
	GC时间	JDBCServer2x的总GC时间	JDBCServer2x的总GC时间。	12000ms
		JobHistory2x的总GC时间	JobHistory2x的总GC时间。	12000ms
		IndexServer2x的总GC时间	IndexServer2x的总GC时间。	12000ms
Storm	集群	Supervisor数	统计周期内集群中可用的Supervisor数目。	1
		已用Slot比率	统计周期内集群中可用的slot使用率。	80.0%
	Nimbus	Nimbus堆内存使用率	Nimbus堆内存使用百分比统计。	80%
Yarn	资源	NodeManager直接内存使用百分比统计	NodeManager直接内存使用百分比统计。	90%
		NodeManager堆内存使用百分比统计	NodeManager堆内存使用百分比统计。	95%
		NodeManager非堆内存使用百分比统计	NodeManager非堆内存使用百分比统计。	90%
		ResourceManager直接内存使用百分比统计	ResourceManager直接内存使用百分比统计。	90%
		ResourceManager堆内存使用百分比统计	ResourceManager堆内存使用百分比统计。	95%
		ResourceManager非堆内存使用百分比统计	ResourceManager非堆内存使用百分比统计。	90%
	垃圾回收	垃圾回收时间统计（GC）	NodeManager每分钟的垃圾回收（GC）所占用的时间。	12000ms

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		垃圾回收时间统计（GC）	ResourceManager每分钟的垃圾回收（GC）所占用的时间。	12000ms
	其他	root队列下失败的任务数	root队列下失败的任务数。	50
		root队列下被杀死的任务数	root队列下被杀死的任务数。	50
	CPU和内存	挂起的内存量	挂起的内存量。	83886080MB
	任务	正在挂起的任务	正在挂起的任务。	60
ZooKeeper	连接	ZooKeeper连接数使用率	ZooKeeper连接数使用百分比统计。	80%
	CPU和内存	ZooKeeper堆内存使用率	ZooKeeper堆内存使用百分比统计。	95%
		ZooKeeper直接内存使用率	ZooKeeper直接内存使用百分比统计。	80%
	GC	ZooKeeper每分钟的垃圾回收时间统计（GC）	ZooKeeper每分钟的垃圾回收时间统计（GC）。	12000ms
meta	OBS数据写操作	OBS数据写操作接口调用成功率	OBS数据写操作接口调用成功率。	99.0%
	OBS元数据操作	OBS元数据接口调用平均时间	OBS元数据接口调用平均时间。	500ms
		OBS元数据接口调用成功率	OBS元数据接口调用成功率。	99.0%
	OBS数据读操作	OBS数据读操作接口调用成功率	OBS数据读操作接口调用成功率。	99.0%
Ranger	GC	UserSync垃圾回收（GC）时间	UserSync垃圾回收（GC）时间。	12000ms

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		RangerAdmin垃圾回收（GC）时间	RangerAdmin垃圾回收（GC）时间。	12000ms
		TagSync垃圾回收（GC）时间	TagSync垃圾回收（GC）时间。	12000ms
	CPU和内存	UserSync非堆内存使用率	UserSync非堆内存使用百分比统计。	80.0%
		UserSync直接内存使用率	UserSync直接内存使用百分比统计。	80.0%
		UserSync堆内存使用率	UserSync堆内存使用百分比统计。	95.0%
		RangerAdmin非堆内存使用率	RangerAdmin非堆内存使用百分比统计。	80.0%
		RangerAdmin堆内存使用率	RangerAdmin堆内存使用百分比统计。	95.0%
		RangerAdmin直接内存使用率	RangerAdmin直接内存使用百分比统计。	80.0%
		TagSync直接内存使用率	TagSync直接内存使用百分比统计。	80.0%
		TagSync非堆内存使用率	TagSync非堆内存使用百分比统计。	80.0%
		TagSync堆内存使用率	TagSync堆内存使用百分比统计。	95.0%
ClickHouse	集群配额	Clickhouse服务在ZooKeeper的数量配额使用率	ClickHouse服务在ZooKeeper上目录的数量配额使用百分比。	90%

服务	监控指标组名称	监控指标名称	指标含义	默认阈值
		Clickhouse服务在ZooKeeper的容量配额使用率	ClickHouse服务在ZooKeeper上目录的容量配额使用百分比。	90%

11.5.1.3 配置告警屏蔽状态

操作场景

如果如下特定场景中不希望看到FusionInsight Manager上报指定的告警，可以手动设置屏蔽。

- 使用过程中，不想关注某些不重要的告警，屏蔽次要告警。
- 第三方产品集成FusionInsight产品时，部分告警与产品自身的告警信息重复，屏蔽重复告警。
- 部署环境特殊时，可能存在特定告警误报，屏蔽误报的告警。

某种告警被屏蔽后，与该告警ID相同的新告警将不再出现在“告警管理”页面中，也不会被统计。已经上报的告警仍然显示。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 告警 > 屏蔽设置”。

步骤3 在“屏蔽设置”区域，选择指定的服务或模块。

步骤4 在告警列表中选择指定的告警。

图 11-7 屏蔽告警

界面显示此告警的信息，包含名称、ID、级别、屏蔽状态和操作：

- 屏蔽状态包含：“屏蔽”和“显示”。
- 操作包含：“屏蔽”和“查看帮助”。

说明

在屏蔽列表上方可筛选指定的告警。

步骤5 设置已选中告警的屏蔽状态：

- 单击“屏蔽”后在弹出的对话框中单击“确定”，修改告警的屏蔽状态为“屏蔽”。
- 单击“取消屏蔽”后在弹出的对话框中单击“确定”，修改告警的屏蔽状态为“显示”。

----结束

11.5.2 日志

11.5.2.1 在线检索日志

操作场景

FusionInsight Manager支持在线检索并显示组件的日志内容，用于问题定位等其他日志查看场景。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 日志 > 在线检索”。

步骤3 根据所需查询日志分别填写**表11-22**各参数，用户可以根据需要选择所需查阅日志时长，缺省时间区间包括：半小时、1小时、2小时、6小时、12小时、1天、1周、1月，也可以单击自定义“开始时间”和“结束时间”。

表 11-22 日志检索参数

参数名	说明
检索内容	检索的关键字或正则表达式。
服务	选择所需查询日志的服务或模块。
文件	当且仅当选择服务中一个角色时，支持选择指定日志文件进行搜索。
最低日志级别	选择所需查询日志的最低日志级别，选择某一级别后会显示从本级别到更高日志级别的日志。 级别从低到高依次为： TRACE < DEBUG < INFO < WARN < ERROR < FATAL

参数名	说明
主机范围	• 单击 ... 可勾选所需主机。 • 请输入所需查询日志的节点主机名或管理平面的IP地址。 • 各IP地址间用“,”隔开，例如：192.168.10.10,192.168.10.11。 • 如果IP地址连续，用“-”连接。例如：192.168.10.[10-20]。 • 如果IP地址分段连续，连续时用“-”连接，各IP地址段间用“,”隔开，例如：192.168.10.[10-20,30-40]。 **说明** – 如不指定，默认选择所有主机。 – 一次性输入最多10个表达式。 – 所有表达式一次性最多匹配2000个主机。
高级配置	• 最大数量：一次性显示的最大日志条数，如果检索到的日志数量超过设定值，时间较早的将被忽略。不配表示不限制。 • 检索超时：用于限制每个节点上的最大检索时间，超时后会中止搜索，已经搜索到的结果仍会显示。

步骤4 单击“检索”开始搜索，结果包含字段如表11-23所示。

表 11-23 检索结果

参数名	说明
时间	该行日志产生的具体时间点。
来源	产生日志的集群。
主机名称	记录该行日志的日志文件所在节点的主机名。
位置	该行日志所在的日志文件的具体路径。 单击位置信息可进入在线日志浏览页面。默认显示该日志所在行前后各100条日志，可单击页首或页尾的“更多”显示更多日志信息。单击“下载”可以下载该日志文件到本地。
行号	该行日志在日志文件中所在的行数。
级别	该行日志的级别。
日志	日志的具体内容。

说明

在检索过程中可单击“停止”强制停止当前检索进度，并在列表中显示已检索出的结果。

步骤5 单击“过滤”，可以针对界面上已经显示的日志信息进行二次筛选，具体字段如表11-24所示。填写完毕后，单击“过滤”进行检索，单击“重置”可清空已填写信息。

表 11-24 过滤

参数名	说明
关键字	需要检索的日志关键字。
主机名称	需要检索的主机名。
位置	所需检索的日志文件路径。
开始时间	所需检索日志信息的开始时间。
结束时间	所需检索日志信息的结束时间。
来源集群	需要检索的集群。

----结束

11.5.2.2 下载日志

操作场景

FusionInsight Manager支持批量导出各个服务角色所有实例生成的日志，无需手工登录单个节点获取。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 日志 > 下载”。

步骤3 选择日志下载范围。

1. “服务”：单击 ··· 勾选所需服务。
2. “主机”：填写服务所部署主机的IP，也可单击 ··· 勾选所需主机。
3. 单击右上角的 设置日志的起始收集时间“开始时间”和“结束时间”。

步骤4 单击“下载”完成日志下载。

下载的日志压缩包中会包括对应开始时间和结束时间的拓扑信息，方便查看与定位。

拓扑文件以“topo_<拓扑结构变化时间点>.txt”命名。文件内容包括节点IP、节点主机名以及节点所安装的服务实例（OMS节点以“Manager:Manager”标识）。

例如：

```
192.168.204.124|suse-124|
DBService:DBServer;KrbClient:KerberosClient;LdapClient:SlapdClient;LdapServer:SlapdServer;Manager:Manager;meta:meta
```

----结束

11.5.3 健康检查

11.5.3.1 查看健康检查任务

操作场景

管理员可以在健康检查的管理中心查看所有健康检查任务，便于在修改某些配置之后的场景对比修改前后是否对集群产生影响。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 健康检查”。

默认显示所有已保存的检查报告，以列表形式展示，包含如下所示的信息：

表 11-25 健康检查报告记录

项目	说明
检查对象	表示被检查的对象，可打开下拉菜单查看详情。
状态	表示检查的结果状态，包含未发现问题、发现问题和检查中。
检查类型	表示执行检查的主体，包含“系统”、“集群”、“主机”、“服务”和“OMS”五个检查维度。集群维度默认包含所有检查。
启动方式	表示此次检查的属性，是否自动触发或手动执行。
开始时间	表示此次检查的开始时间。
结束时间	表示此次检查的结束时间。
操作	支持“导出报告”和“查看帮助”。

说明

- 在检查记录列表右上方，可以筛选指定的检查对象和结果状态。
- 如果检查类型为集群时，“查看帮助”在“检查对象”的下拉菜单中。
- 系统执行健康检查时，涉及检查对象的监控指标数据时，并非以当前实时的监控数据进行判断，而是收集近期的历史数据，因此存在时间延迟。

----结束

11.5.3.2 管理健康检查报告

操作场景

用户可以在FusionInsight Manager对已保存的所有健康检查报告进行管理，即下载和删除历史健康检查报告。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 健康检查”。

步骤3 在目标健康检查报告所在行，单击“导出报告”，下载报告文件。

----结束

11.5.3.3 修改健康检查配置

操作场景

管理员可以启用自动健康检查减少手工操作时间。自动健康检查默认会对整个集群进行检查。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 健康检查 > 配置”。

“定期健康检查”表示是否启用自动执行健康检查，选择“启用”表示启用，默认“不启用”表示不启用。

启用后根据运维需要选择检查周期为：“每天”、“每周”或“每月”。

步骤3 单击“确定”保存配置。

----结束

11.5.4 备份恢复设置

11.5.4.1 创建备份任务

操作场景

FusionInsight Manager支持在界面上创建备份任务，运行备份任务将对指定的数据进行备份。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 备份恢复 > 备份管理 > 创建”。

步骤3 设置“备份对象”为“OMS”或需要备份数据的集群。

步骤4 在“任务名称”输入参数值。

步骤5 在“备份类型”选择任务执行属性。

表 11-26 备份类型说明

类型	参数	说明
周期备份	开始时间	表示周期备份任务第一次启动的时间
	周期	表示任务下次启动，与上一次运行的时间间隔，支持“小时”或“天”
	备份策略	可以选择下策略： ● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份
手动备份	无	需要手动运行任务才能进行备份

步骤6 在“备份配置”指定需要备份的数据。

- 支持备份元数据和业务数据。
- 各组件不同数据的备份任务操作请参考**备份恢复管理**。

步骤7 单击“确定”保存。

步骤8 在备份任务列表，可以查看刚创建的备份任务。

在指定的备份任务“操作”列，选择“更多 > 即时备份”，可以立即运行备份任务。

----结束

11.5.4.2 创建恢复任务

操作场景

FusionInsight Manager支持在界面上创建恢复任务，运行恢复任务将把指定的备份数据恢复到集群中。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 备份恢复 > 恢复管理 > 创建”。

步骤3 设置“任务名称”。

步骤4 设置“恢复对象”为“OMS”或需要恢复数据的集群。

步骤5 在“恢复配置”指定需要恢复的数据。

- 支持恢复元数据和业务数据。
- 各组件不同数据的恢复任务操作请参考**备份恢复管理**。

步骤6 单击“确定”保存。

步骤7 在恢复任务列表，可以查看刚创建的恢复任务。

在指定的恢复任务“操作”列，单击“执行”，可以立即运行恢复任务。

----结束

11.5.4.3 其他任务管理说明

操作场景

FusionInsight Manager还支持对备份恢复进行不同的维护管理功能。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“运维 > 备份恢复 > 备份管理”或“运维 > 备份恢复 > 恢复管理”。

步骤3 在任务列表指定任务的“操作”列，选择需要执行的操作。

表 11-27 更多维护管理功能

操作入口	说明
“配置”	修改备份任务的参数。
“恢复”	部分业务数据的备份任务执行成功后，可以直接使用此功能快速恢复数据。
“更多 > 即时备份”	立即运行备份任务。
“更多 > 停止”	可以停止正在运行的任务。
“更多 > 删除”或“删除”	删除任务。
“更多 > 挂起”	禁用自动备份任务。
“更多 > 重新执行”	启用自动备份任务。
“更多 > 查询历史”或“查询历史”	打开任务运行日志窗口，查看运行详细情况以及备份路径。
“查看”	检查恢复任务的参数设置。
“执行”	运行恢复任务。

----结束

11.6 审计

11.6.1 审计管理页面概述

操作场景

“审计”页面记录用户对集群Manager页面操作信息。管理员可通过该页面查看用户在Manager上的历史操作记录。审计管理包含的审计内容信息，请参考**审计日志**。

概述

登录FusionInsight Manager，单击“审计”，界面展示FusionInsight Manager审计信息，包括操作类型、安全级别、开始时间、结束时间、用户、来源名称、主机、服务、实例、操作结果等。

- 用户可以在“所有安全级别”中选择“高危”、“危险”、“一般”和“提示”级别的审计日志。
- 在高级搜索中，用户可设置过滤条件来查询审计日志。
  a. 在“操作类型”中，用户可根据用户管理、集群、服务、健康检查等来指定操作类型查询对应的审计日志。
  b. 在“服务”中，用户可选择相应的服务来查询审计日志。

  说明

  在服务中选择“--”，表示除服务以外其他类型的审计日志。
  c. 在“操作结果”中，用户可选择所有、成功、失败和未知来查询审计日志。
- 单击手动刷新当前页面，也可在 修改审计表格显示的列。
- 单击“导出全部”，可一次性导出所有审计信息，可导出“TXT”或者“CSV”格式。

11.6.2 配置审计日志转储

操作场景

Manager的审计日志默认保存在数据库中，如果长期保留可能引起数据目录的磁盘空间不足问题，管理员如果需要将审计日志保存到其他归档服务器，可以在FusionInsight Manager设置转储参数及时自动转储，便于管理审计日志信息。

若用户未配置审计日志转储，当审计日志达到十万条，系统自动将这十万条审计日志保存到文件中。保存路径为主管理节点“${BIGDATA_DATA_HOME}/dbdata_om/dumpData/iam/operatelog”，保存的文件名格式为“OperateLog_store_YY_MM_DD_HH_MM_SS.csv”，保存的审计日志历史文件数最大为50。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“审计 > 配置”。

步骤3 单击“审计日志转储”右侧的开关。

“审计日志转储”默认为不启用，开关显示为 表示启用。

步骤4 根据表11-28填写转储参数。

表 11-28 审计日志转储参数

参数名	参数解释	参数值
SFTP IP 模式	目标IP的IP地址模式，可选择“IPv4”或者“IPv6”。	IPv4

参数名	参数解释	参数值
SFTP IP	指定审计日志转储后存放的SFTP服务器，建议使用基于SSH v2的SFTP服务，否则存在安全风险。	192.168.10.51（举例）
SFTP端口	指定审计日志转储后存放的SFTP服务器连接端口。	22（举例）
保存路径	指定SFTP服务器上保存审计日志的路径。	/opt/omm/oms/auditLog（举例）
SFTP用户名	指定登录SFTP服务器的用户名。	root（举例）
SFTP密码	指定登录SFTP服务器的密码。	*SFTP服务器的密码*
SFTP公共密钥	可选参数，指定SFTP服务器的公共密钥，建议配置SFTP的公共密钥，否则可能存在安全风险。	-
转储模式	指定转储模式 ● “按数量”：日志到达指定条数（默认10万条）时开始转储 ● “按时间”：指定某一日期开始转储，转储频率为一年一次。	● 按数量 ● 按时间
转储日期	当选择“按时间”转储模式时可用。选择一个转储日期后，系统将在此日期开始转储。转储的日志范围为当前年份1月1日0时之前的所有审计日志。	11-06（举例）

说明

SFTP公共密钥为空时，系统将进行安全风险提示，确定安全风险后再保存配置。

步骤5 单击“确定”，设置完成。

说明

审计日志转储文件关键字段参考：

- “USERTYPE”表示用户类型，“0”表示“人机”用户，“1”表示“机机”用户。
- “LOGLEVEL”表示安全级别，“0”表示高危，“1”表示危险，“2”表示一般，“3”表示提示。
- “OPERATERESULT”表示操作结果，“0”表示成功，“1”表示失败。

----结束

11.7 租户资源

11.7.1 多租户介绍

11.7.1.1 简介

定义

多租户是MRS集群中的多个资源集合（每个资源集合是一个租户），具有分配和调度资源的能力。资源包括计算资源和存储资源。

背景

现代企业的数据集群在向集中化和云化方向发展，企业级大数据集群需要满足：

- 不同用户在集群上运行不同类型的应用和作业（分析、查询、流处理等），同时存放不同类型和格式的数据。
- 某些类型的用户（例如银行、政府单位等）对数据安全非常关注，很难容忍将自己的数据与其他用户的放在一起。

这给大数据集群带来了以下挑战：

- 合理地分配和调度资源，以支持多种应用和作业在集群上平稳运行。
- 对不同的用户进行严格的访问控制，以保证数据和业务的安全。

多租户将大数据集群的资源隔离成一个个资源集合，彼此互不干扰，用户通过“租用”需要的资源集合，来运行应用和作业，并存放数据。在大数据集群上可以存在多个资源集合来支持多个用户的不同需求。

对此，MRS企业级大数据集群提供了完整的企业级大数据多租户解决方案。

优势

- 合理配置和隔离资源

  租户之间的资源是隔离的，一个租户对资源的使用不影响其他租户，保证了每个租户根据业务需求去配置相关的资源，可提高资源利用效率。
- 测量和统计资源消费

  系统资源以租户为单位进行计划和分配，租户是系统资源的申请者和消费者，其资源消费能够被测量和统计。
- 保证数据安全和访问安全

  多租户场景下，分开存放不同租户的数据，以保证数据安全；控制用户对租户资源的访问权限，以保证访问安全。

11.7.1.2 技术原理

11.7.1.2.1 多租户管理页面概述

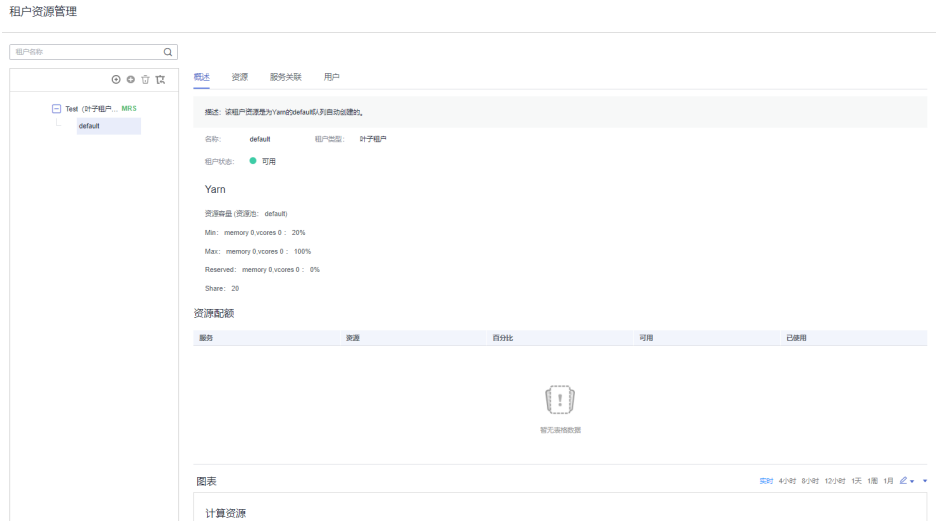
统一的多租户管理

登录FusionInsight Manager，选择“租户资源 > 租户资源管理”，可以查看到FusionInsight Manager作为统一的多租户管理平台，集成了租户生命周期管理、租户资源配置、租户服务关联和租户资源使用统计等功能，为企业提供了成熟的多租户管理模型，实现集中式的租户和业务管理。

图形化的操作界面

FusionInsight Manager实现全图形化的多租户管理界面：通过树形结构实现多级租户的管理和操作，将当前租户的基本信息和资源配额集成在一个界面中，方便运维和管理，如图11-8所示。

图 11-8 多租户管理



层级式的租户管理

FusionInsight Manager支持层级式的租户管理，可以为租户进一步添加子租户，实现资源的再次配置。一级租户下一级的子租户属于二级租户，以此类推。为企业提供了成熟的多租户管理模型，实现集中式的租户和业务管理。

简化的权限管理

FusionInsight Manager对普通用户封闭了租户内部的权限管理细节，对管理员简化了权限管理的操作方法，提升了租户权限管理的易用性和用户体验。

- 使用RBAC方式，在多租户管理时，可根据业务场景为各用户分别配置不同权限。
- 租户的管理员，具有租户的管理权限，包括：查看当前租户的资源和服务、在当前租户中添加/删除子租户并管理子租户资源的权限。支持定义单个租户的管理员，可以将租户的管理权限委托给系统管理员之外的其他用户。
- 租户对应的角色，具有租户的计算资源和存储资源的全部权限。创建租户时，系统自动创建租户对应的角色，可以添加用户并绑定该角色为其他用户授权，以使用该租户的资源。

清晰的资源管理

- **资源自主配置**

  FusionInsight Manager支持在创建租户时配置计算资源和存储资源，和进一步添加、修改、删除租户内资源。

  修改租户的计算资源或存储资源，当前租户对应的角色所关联的权限将自动更新。

- **资源使用统计**

  资源使用统计是管理员获取当前集群应用和服务的运行状态，提高集群运维效率，做出运维决策的重要依据。FusionInsight Manager通过“资源配额”展示租

户的资源统计，包括租户动态计算资源vCores和Memory，HDFS存储资源（Space）的使用统计。

说明

- “资源配额”视图动态计算租户资源使用情况。

服务	资源	百分比	可用	已使用
Yarn	Memory	0.00%	24.00 GB	0 MB
Yarn	CPU	0.00%	12 vCores	0 vCores

  Superior调度器可用资源计算方式分别如下：

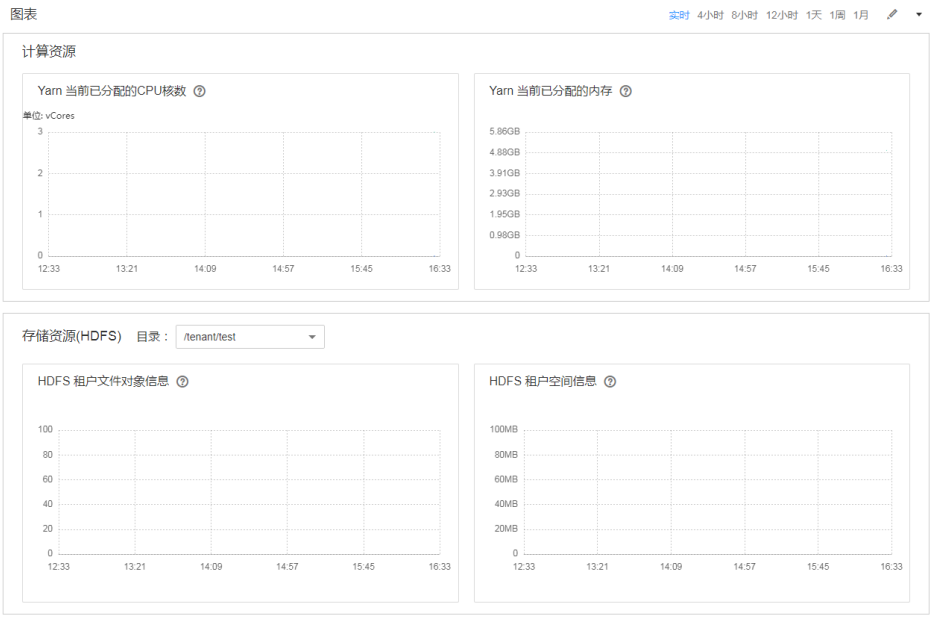
  - Superior

    Yarn可用资源（Memory、CPU）为资源池容量按照队列权重按比例分配。
- 当租户管理员绑定一个租户角色时，租户管理员将拥有该租户的管理权限，以及该租户全部资源的权限。

- **资源图形化监控**

  资源图形化监控支持表11-29中监控项图形化显示，如图11-9所示。

图 11-9 精细化监控



默认显示实时的监控数据，用户可单击自定义时间区间，缺省时间区间包括：4小时、8小时、12小时、1天、1周、1月，单击 ▼ ，在弹出菜单中选择“导出”，导出对应的监控项信息。

表 11-29 监控项

所属服务	监控指标项	说明
HDFS	HDFS租户空间信息 ● 分配的空间大小 ● 已使用的空间大小	HDFS可选择指定的存储目录进行监控。存储目录与当前租户在“资源”中添加的目录一致。

所属服务	监控指标项	说明
	HDFS租户文件对象信息 ● 已使用的文件对象个数	
Yarn	Yarn当前已分配的CPU核数 ● AM分配的最大CPU核数 ● 已分配的CPU核数 ● AM已使用的CPU核数	当前租户的监控信息。如某租户未配置相应子项，则不显示。 监控数据取自Yarn原生WebUI中“Scheduler > Application Queues > Queue:*租户名*”。
	Yarn当前已分配的内存 ● AM分配的最大内存 ● 已分配的内存 ● AM已使用的内存	

11.7.1.2.2 相关模型

多租户相关模型

多租户相关模型如下图所示。

图 11-10 多租户相关模型

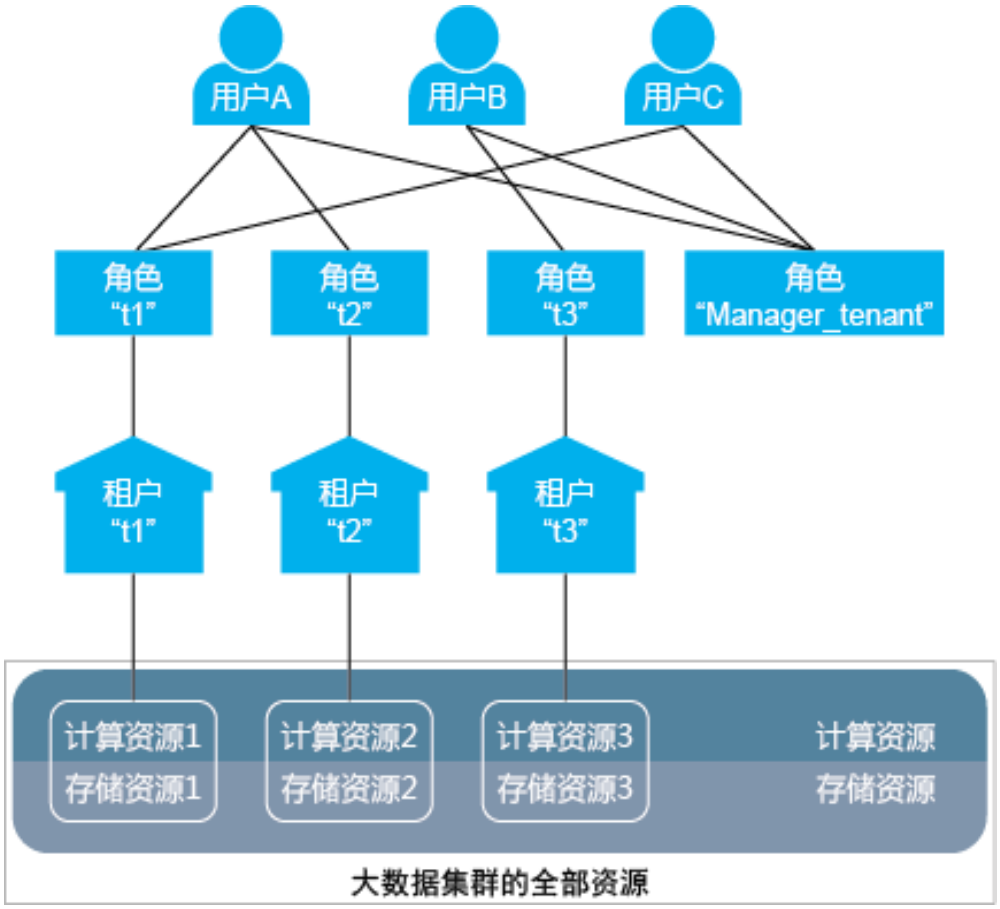


图11-10中涉及的概念如表11-30所示。

表 11-30 相关概念说明

概念	说明
用户	用户是拥有用户名，密码等账号信息的自然人，是大数据集群的使用者。 图11-10中有三个不同的用户：用户A、用户B和用户C。
角色	角色是承载一个或多个权限的载体。权限是限定在具体对象上的，例如对HDFS中的“/tenant”目录的访问权限，这里权限就限定在“/tenant”目录这个具体对象上。 图11-10中有四个不同的角色：角色“t1”、角色“t2”、角色“t3”和角色“Manager_tenant”。 • 角色“t1”、角色“t2”和角色“t3”为创建租户时，集群自动生成的角色，角色名和租户名相同，分别对应租户“t1”、租户“t2”和租户“t3”，不能单独使用。 • 角色“Manager_tenant”为集群中本身存在的角色，不能单独使用。
租户	租户是从大数据集群中划分出的资源集合。多个不同的租户统称为多租户，租户内部进一步划分出的资源集合是子租户。 图11-10中有三个不同的租户：租户“t1”、租户“t2”和租户“t3”。

概念	说明
资源	• 计算资源包括CPU和内存。 租户的计算资源是从集群总计算资源中划分出的，租户之间不可以互占计算资源。 图11-10中：计算资源1、计算资源2和计算资源3分别是租户“t1”、租户“t2”和租户“t3”从集群中划分出的计算资源。 • 存储资源包括磁盘或第三方存储系统。 租户的存储资源是从集群总存储资源中划分出的，租户之间不可以互占存储资源。 图11-10中：存储资源1、存储资源2和存储资源3分别是租户“t1”、租户“t2”和租户“t3”从集群中划分出的存储资源。

若用户想要使用租户资源或为租户添加/删除子租户，则需要同时绑定该租户对应的角色和角色“Manager_tenant”。在图11-10中，各用户绑定的角色如表11-31所示。

表 11-31 各用户绑定的角色

用户	绑定的角色	权限
用户A	• 角色“t1” • 角色“t2” • 角色“Manager_tenant”	• 使用租户“t1”和租户“t2”的资源。 • 为租户“t1”和租户“t2”添加/删除子租户。
用户B	• 角色“t3” • 角色“Manager_tenant”	• 使用租户“t3”的资源。 • 为租户“t3”添加/删除子租户。
用户C	• 角色“t1” • 角色“Manager_tenant”	• 使用租户“t1”的资源。 • 为租户“t1”添加/删除子租户。

用户和角色是多对多的关系，一个用户可以绑定多个角色，一个角色可以被多个用户绑定。用户通过绑定角色和租户建立关系，因此用户和租户也是多对多的关系。一个用户可以使用多个租户的资源，多个用户也可以使用同一个租户的资源，例如图11-10中，用户A使用租户“t1”和租户“t2”的资源，用户A和用户C都使用租户“t1”的资源。

说明

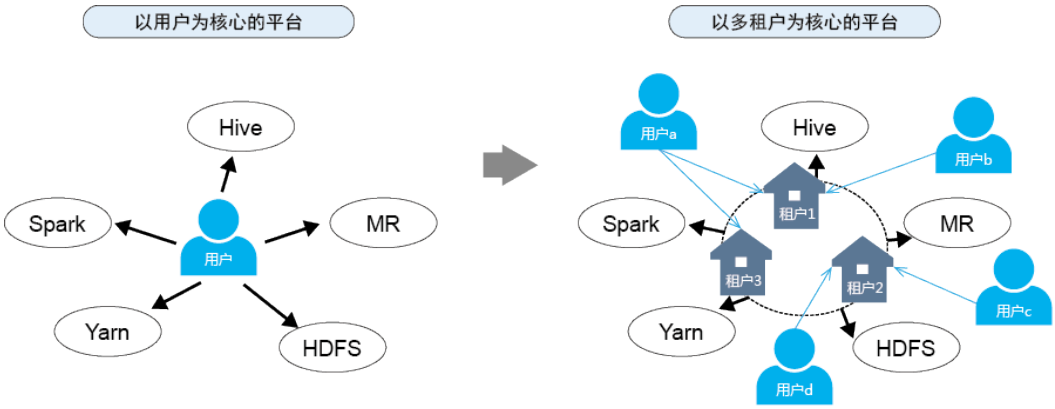
“父租户”、“子租户”、“一级租户”、“二级租户”的概念是针对客户的多租户业务场景设定的，注意与FusionInsight Manager上的“叶子租户资源”和“非叶子租户资源”的概念区别。

- 一级租户：按照租户所在层级确定名称，如最初创建的租户为一级租户，一级租户的子租户为二级租户。
- 父租户与子租户：用来表述租户间上下层级关系的称呼。
- 非叶子租户资源：创建租户时，选择的租户类型，该类型可以创建子租户。
- 叶子租户资源：创建租户时，选择的租户类型，该类型不可以创建子租户。

多租户平台

租户是FusionInsight大数据平台的核心概念，使传统的以用户为核心的大数据平台向以多租户为核心的大数据平台转变，更好的适应现代企业多租户应用环境，如图11-11所示。

图 11-11 以用户为核心的平台和以多租户为核心的平台



对于以用户为核心的大数据平台，用户直接访问并使用全部的资源和服务。

- 用户的应用可能只用到集群的部分资源，资源利用效率低。
- 不同用户的数据可能存放在一起，难以保证数据安全。

对于以租户为核心的大数据平台，用户通过访问租户来使用需要的资源和服务。

- 按照应用需求分配和调度出需要的资源，以租户来统一使用，资源利用效率高。
- 用户通过分配不同的角色获得使用不同租户资源的权限，以保障访问安全。
- 不同的租户之间数据隔离，以保证数据安全。

11.7.1.2.3 资源概述

MRS集群的资源分为计算资源和存储资源。多租户可实现资源的隔离：

- **计算资源**

  计算资源包括CPU和内存。租户之间不可以相互占用计算资源，私有计算资源独立。
- **存储资源**

  存储资源包括磁盘或第三方存储系统。租户之间不可以相互访问数据，私有存储资源独立。

计算资源

计算资源可分为静态服务资源和动态资源：

- **静态服务资源**

  静态服务资源是集群分配给各个服务的计算资源，每个服务的计算资源总量固定，不与其他服务共享，是静态的。这些服务包括Flume、HBase、HDFS和Yarn。

- **动态资源**

  动态资源是分布式资源管理服务Yarn动态调度给任务队列的计算资源。Mapreduce、Spark2x、Flink和Hive的任务队列由Yarn来动态调度资源。

说明

大数据集群为Yarn分配的资源是静态服务资源，可以由Yarn动态分配给任务队列计算使用。

存储资源

存储资源是分布式文件存储服务HDFS中可分配的数据存储空间资源。目录是HDFS存储资源分配的基本单位，租户通过指定HDFS文件系统的目录来获取存储资源。

11.7.1.2.4 动态资源

简介

Yarn是大数据集群中的分布式资源管理服务，大数据集群为Yarn分配资源，资源总量可配置。Yarn内部为任务队列进一步分配和调度计算资源。对于Mapreduce、Spark、Flink和Hive的任务队列，计算资源完全由Yarn来分配和调度。

Yarn任务队列是计算资源分配的基本单位。

对于租户，通过Yarn任务队列申请到的资源是动态资源。用户可以动态创建并修改任务队列的配额，可以查看任务队列的使用状态和使用统计。

资源池

现代企业IT经常会面对纷繁复杂的集群环境和上层需求。例如以下业务场景：

- 集群异构，集群中各个节点的计算速度、存储容量和网络性能存在差异，需要把复杂应用的所有任务按照需求，合理地分配到各个计算节点上。
- 计算分离，多个部门需要数据共享，但是需要把计算完全分离在不同的计算节点上。

这就要求对计算资源的节点进一步分区。

资源池用来指定动态资源的配置。Yarn任务队列和资源池关联，可实现资源的分配和调度。

一个租户只能设置一个默认资源池。用户通过绑定租户相关的角色，来使用该租户资源池的资源。若需要使用多个资源池的资源，可通过绑定多个租户相关的角色实现。

调度机制

Yarn动态资源支持标签调度（Label Based Scheduling）策略，此策略通过为计算节点（Yarn NodeManager）创建标签（Label），将具有相同标签的计算节点添加到同一

个资源池中，Yarn根据任务队列对资源的需求，将任务队列和有相应标签的资源池动态关联。

例如，集群中有40个以上的节点，根据各节点的硬件和网络配置，分别用Normal、HighCPU、HighMEM、HighIO为四类节点创建标签，添加到四个资源池中，资源池中的各节点性能如表11-32所示。

表 11-32 不同资源池中的各节点性能

标签名	节点数	硬件和网络配置	添加到	关联
Normal	10	一般	资源池A	普通的任务队列
HighCPU	10	高性能CPU	资源池B	计算密集型的任务队列
HighMEM	10	大量内存	资源池C	内存密集型的任务队列
HighIO	10	高性能网络	资源池D	IO密集型的任务队列

任务队列只能使用所关联的资源池里的计算节点。

- 普通的任务队列关联资源池A，使用硬件和网络配置一般的Normal节点。
- 计算密集型的任务队列关联资源池B，使用具有高性能CPU的HighCPU节点。
- 内存密集型的任务队列关联资源池C，使用具有大量内存的HighMEM节点。
- IO密集型的任务队列关联资源池D，使用具有高性能网络的HighIO节点。

Yarn任务队列与特定的资源池关联，可以更有效地使用资源，保证节点性能充足且互不影响。

FusionInsight Manager中最多支持添加50个资源池。系统默认包含一个默认资源池。

调度器介绍

MRS集群默认即启用了Superior调度器。

- Superior调度器为增强型，Superior取名源自苏必利尔湖，意指由该调度器管理的数据足够大。

为满足企业需求，克服Yarn社区在调度上遇到的挑战与困难，Superior调度器做了以下增强：

- 增强资源共享策略

  Superior调度器支持队列层级，在同集群集成开源调度器的特性，并基于可配置策略进一步共享资源。针对实例，管理员可通过Superior调度器为队列同时配置绝对值或百分比的资源策略计划。Superior调度器的资源共享策略将Yarn的标签调度增强为资源池特性，Yarn集群中的节点可根据容量或业务类型不同，进行分组以使队列更有效地利用资源。
- 基于租户的资源预留策略

  部分租户可能在某些时间中运行关键任务，租户所需的资源应保证可用。Superior调度器构建了支持资源预留策略的机制，在这些租户队列运行的任务可立即获取到预留资源，以保证计划的关键任务可正常执行。
- 租户和资源池的用户公平共享

Superior调度器提供了队列内用户间共享资源的配置能力。每个租户中可能存在不同权重的用户，高权重用户可能需要更多共享资源。

- 大集群环境下的调度性能优势

  Superior调度器接收到各个NodeManager上报的心跳信息，并将资源信息保存在内存中，使得调度器能够全局掌控集群的资源使用情况。Superior调度器采用了push调度模型，令调度更加精确、高效，大大提高了大集群下的资源使用率。另外，Superior调度器在NodeManager心跳间隔较大的情况下，调度性能依然优异，不牺牲调度性能，也能避免大集群环境下的“心跳风暴”。

- 优先策略

  当某个服务在获取所有可用资源后还无法满足最小资源的要求，则会发生优先抢占。抢占功能默认关闭。

11.7.1.2.5 存储资源

简介

HDFS是大数据集群中的分布式文件存储服务，存放大数据集群上层应用的所有用户数据，例如写入HBase表或Hive表的数据。

目录是HDFS存储资源分配的基本单位。HDFS支持传统的层次型文件组织结构。用户或者应用程序可以创建目录，在目录中创建、删除、移动或重命名文件。租户通过指定HDFS文件系统的目录来获取存储资源。

调度机制

系统支持将HDFS目录存储到指定标签的节点上，或存储到指定硬件类型的磁盘上。例如以下业务场景：

- 实时查询与数据分析共集群时，实时查询只需部署在部分节点上，其数据也应尽可能的只存储在这些节点上。
- 关键数据根据实际业务需要保存在具有高度可靠性的节点中。

管理员可以根据实际业务需要，通过数据特征灵活配置HDFS数据存储策略，将数据保存在指定的节点上。

对于租户，存储资源是各租户所占用的HDFS资源。可以通过将指定目录的数据存储到租户配置的存储路径中，实现存储资源调度，保证租户间的数据隔离。

用户可以添加/删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额来管理存储资源。

11.7.1.3 多租户使用

11.7.1.3.1 使用说明

租户主要用于资源控制、业务隔离的场景。在实际业务中，管理员需要先明确使用集群资源的业务场景，规划租户。

说明

- 新安装集群的Yarn组件默认使用的是Superior调度器，参见**使用Superior调度器的租户业务**。

多租户使用包含三类操作：创建租户、管理租户和管理资源。各操作的具体动作如表11-33所示。

表 11-33 使用租户的各种操作

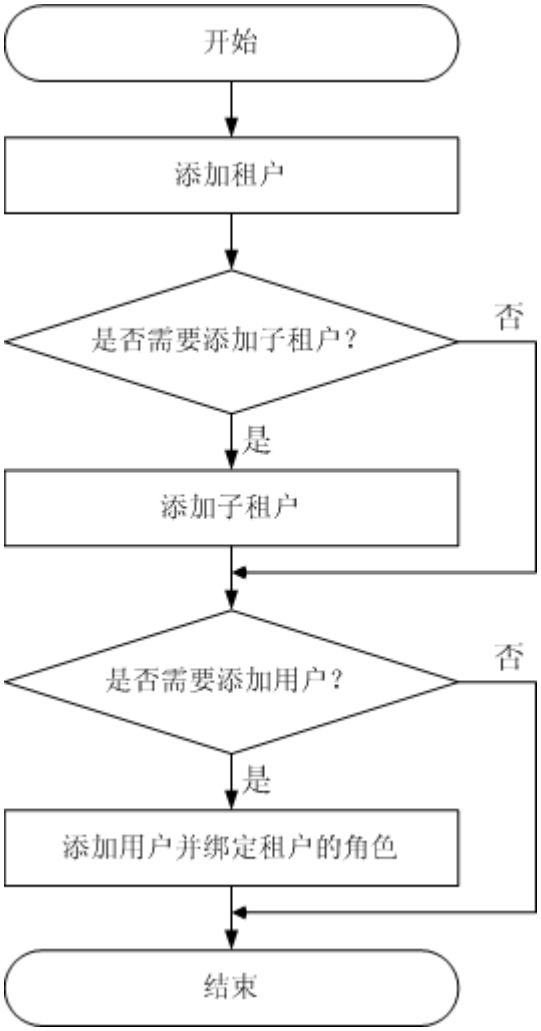
操作	具体动作	说明
创建租户	• 添加租户 • 添加子租户 • 添加用户并绑定租户的角色	创建租户时，便可根据业务需求，为租户配置计算资源、存储资源和关联服务；为租户添加用户，并为用户绑定需要的角色。 创建一级租户的用户，需要绑定“Manager_administrator”或“System_administrator”角色。 创建子租户的用户，至少需要绑定父租户对应的角色。
管理租户	• 管理租户目录 • 恢复租户数据 • 清除租户非关联队列 • 删除租户	管理租户是随着业务变化对租户进行的编辑操作。 管理或删除一级租户的用户，以及恢复租户数据的用户，需要绑定“Manager_administrator”或“System_administrator”角色。 管理或删除子租户的用户，至少需要绑定父租户对应的角色。
管理资源	• 添加资源池 • 修改资源池 • 删除资源池 • 配置队列 • 配置资源池的队列容量策略 • 清除队列配置	管理资源是随着业务变化对租户再次配置资源的操作。 管理资源的用户，需要绑定“Manager_administrator”或“System_administrator”角色。

11.7.1.3.2 流程概述

在实际业务中，管理员需要先明确使用集群资源的业务场景，规划租户。然后在FusionInsight Manager界面添加租户，并配置租户的动态资源、存储资源以及所关联的服务。

创建租户的具体操作流程如流程概述所示。

图 11-12 创建租户流程



创建租户的操作说明如表11-34所示。

表 11-34 创建租户的操作说明

操作	说明
添加租户	可配置待添加租户的计算资源、存储资源和关联服务。
添加子租户	可配置待添加子租户的计算资源、存储资源和关联服务。
添加用户并绑定租户的角色	若一个用户想要使用“tenant1”租户包含的资源，或为“tenant1”租户添加/删除子租户，则需要同时绑定“Manager_tenant”和“tenant1_*集群ID*”两个角色。

11.7.2 使用 Superior 调度器的租户业务

11.7.2.1 创建租户

11.7.2.1.1 添加租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建租户，以满足实际使用场景。

前提条件

- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 单击⊕，打开添加租户的配置页面，参见**表11-35**为租户配置属性。

表 11-35 租户参数一览

参数名	描述
集群	选择要创建租户的集群。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
租户资源类型	指定租户是否是一个叶子租户： • 选择“叶子租户资源”：当前租户为叶子租户，不支持添加子租户。 • 选择“非叶子租户资源”：当前租户为非叶子租户，支持添加子租户。 **说明** MRS 3.2.0及之后版本，如果关联服务需要选择“ClickHouse”，租户资源类型只能选择“叶子租户”。

参数名	描述
计算资源	为当前租户选择动态计算资源。 ● 选择“Yarn”时，系统自动在Yarn中以租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 ● 不选择“Yarn”时，系统不会自动创建任务队列。
配置模式	计算资源参数配置模式。 ● 选择“基础”时，只需配置“默认资源池容量（%）”参数即可。 ● 选择“高级”时，可手动配置资源分配权重，租户的最小/最大/预留资源。
默认资源池容量（%）	配置当前租户在默认资源池中使用的计算资源百分比，取值范围0~100%。
权重	资源分配权重，取值范围从0到100。
最小资源	保证租户资源能获得的资源（有抢占支持）。取值可以是父租户资源的百分比或绝对值。当租户资源作业量比较少时，资源会自动借给其他租户资源，当租户资源能使用的资源不满足最小资源时，可以通过抢占来要回之前借出的资源。
最大资源	租户资源最多能使用的资源，租户资源不能得到比最大资源设定更多的资源。取值可以是父租户资源的百分比或绝对值。
预留资源	租户资源预留资源。即使租户资源内没有作业，预留的资源也不能给别的租户资源使用。取值可以是父租户资源的百分比或绝对值。
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将分配存储资源。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。

参数名	描述
存储空间配额	配置当前租户使用的HDFS存储空间配额。 • 取值范围：当存储空间配额单位设置为MB时，范围为1～8796093022208。当存储空间配额单位设置为GB时，范围为1～8589934592。 • 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 • 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。
存储路径	配置租户在HDFS中的存储目录。 • 系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“/tenant/ta1”。 • 第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
服务	是否需要关联使用其他服务的资源，参见**步骤4**。
描述	配置当前租户的描述信息。

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 创建租户时系统会自动创建一个Yarn任务队列，并自动以租户名称命名该队列。如果已经存在同名队列，新队列命名为“租户名称-N”。“N”表示从1开始的自然数，存在同名队列的时候N会自动累加以区别已有队列。例如“saletenant”、“saletenant-1”和“saletenant-2”。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源，单击“确定”。

- “服务”选择“HBase”，在“关联类型”可以选择“独占”或者“共享”。

  说明

  - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
  - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

- MRS 3.2.0及之后版本：“服务”选择“ClickHouse”。
  - 关联类型”：当“服务”选择“ClickHouse”时，关联类型”只支持“共享”。
  - “关联逻辑集群”：若ClickHouse没有开启逻辑集群，则默认关联default_cluster，若已经开启逻辑集群，则按需选择需要关联的逻辑集群。

- “CPU优先级”：CPU优先级取值范围为-20~19，该值关联OS的NICE值，取值越小，CPU优先级越高。
- “内存”：内存限制为百分比，如该值设置为80，则当前租户可使用的内存总额为：服务可使用内存 * 80%。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

11.7.2.1.2 添加子租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建子租户，将当前租户的资源进一步分配以满足实际使用场景。

前提条件

- 已添加父租户，且属于非叶子租户。
- 已根据业务需求规划子租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择父租户节点然后单击，打开添加子租户的配置页面，参见**表11-36**为子租户配置属性。

表 11-36 子租户参数一览

参数名	描述
集群	显示上级父租户所在集群。
父租户资源	显示上级父租户的名称。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划子租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。

参数名	描述
租户资源类型	指定租户是否是一个叶子租户： ● 选择“叶子租户资源”：当前租户为叶子租户，不支持添加子租户。 ● 选择“非叶子租户资源”：当前租户为非叶子租户，支持添加子租户，但租户层级不能超过5层。
计算资源	为当前租户选择动态计算资源。 ● 选择“Yarn”时，系统自动在Yarn中以子租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 ● 不选择“Yarn”时，系统不会自动创建任务队列。
配置模式	计算资源参数配置模式。 ● 选择“基础”时，只需配置“默认资源池容量（%）”参数即可。 ● 选择“高级”时，可手动配置资源分配权重，租户的最小/最大/预留资源。
默认资源池容量 （%）	配置当前租户使用的计算资源百分比，基数为父租户的资源总量。
权重	资源分配权重，取值范围从0到100。
最小资源	保证租户资源能获得的资源（有抢占支持）。取值可以是父租户资源的百分比或绝对值。当租户资源作业量比较少时，资源会自动借给其他租户资源，当租户资源能使用的资源不满足最小资源时，可以通过抢占来要回之前借出的资源。
最大资源	租户资源最多能使用的资源，租户资源不能得到比最大资源设定更多的资源。取值可以是父租户资源的百分比或绝对值。
预留资源	租户资源预留资源。即使租户资源内没有作业，预留的资源也不能给别的租户资源使用。取值可以是父租户资源的百分比或绝对值。
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。

参数名	描述
存储空间配额	配置当前租户使用的HDFS存储空间配额。 ● 当存储空间配额单位设置为MB时，范围为1～8796093022208，当“存储空间配额单位”设置为GB时，范围为1～8589934592。 ● 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 ● 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 ● 如果此配额大于父租户的配额，实际存储量不超过父租户配额。
存储路径	配置租户在HDFS中的存储目录。 ● 系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“/tenant/ta1”，系统默认自动配置此参数值为“/tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。 ● 支持在父目录中自定义存储路径。
服务	是否需要关联使用其他服务的资源，参见**步骤4**。
描述	配置当前租户的描述信息。

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 子租户可以将当前租户的资源进一步分配。每一级别父租户下，直接子租户的资源百分比之和不能超过100%。所有一级租户的计算资源百分比之和也不能超过100%。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：
   - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
   - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

11.7.2.1.3 添加用户并绑定租户的角色

操作场景

创建好的租户不能直接登录集群访问资源，管理员需要通过FusionInsight Manager为已有租户创建新用户，通过绑定租户的角色继承其操作权限，以满足业务使用。

前提条件

管理员已明确业务需求，并已创建了租户。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 用户”。

步骤2 若在系统中添加新的用户，请单击“添加用户”，打开添加用户的配置页面。

若为系统中已有的用户绑定租户权限，请单击该用户所在行的“修改”，打开修改用户的配置页面。

参见**表11-37**为用户配置属性。

表 11-37 用户参数一览

参数名	描述
用户名	指定当前的用户名，长度为3~32个字符，可包含数字、字母、下划线（_）、中划线（-）和空格。 • “用户名”不能与集群各节点所有操作系统用户名相同，否则此用户无法正常使用。 • 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。使用“User1”时请输入正确的用户名。

参数名	描述
用户类型	可选值包括“人机”和“机机”。 ● “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。 ● “机机”用户：用于应用开发的场景。选择该值用户密码随机生成，无需填写。
密码	选择“人机”用户需填写“密码”。 密码必须包含8~64个字符，至少包含以下类型字符中的四种：大写字母、小写字母、数字、特殊字符和空格。不能与用户名或倒序的用户名相同。
确认密码	再次输入密码。
用户组	单击“添加”，选择对应用户组将用户添加进去。 ● 如果用户组添加了角色，则用户可获得对应角色中的权限。 ● 例如，为新用户分配Hive的权限，请将用户加入Hive组。
主组	选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。
角色	单击“添加”为用户绑定租户的角色。 **说明** ● 若一个用户想要获取使用“tenant1”租户包含的资源，且能够为“tenant1”租户添加/删除子租户，则需要同时绑定“Manager_tenant”和“tenant1_*集群ID*”两个角色。 ● 如果租户关联了HBase服务且当前集群启用了Ranger鉴权，用户需要通过Ranger界面配置HBase相关执行权限。
描述	配置当前用户的描述信息。

步骤3 单击“确定”完成用户创建。

----结束

11.7.2.2 管理租户

11.7.2.2.1 管理租户目录

操作场景

管理员通过FusionInsight Manager管理指定租户使用的HDFS存储目录，能根据业务需求对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

已添加具有HDFS存储资源的租户。

查看租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击目标的租户。

步骤3 单击“资源”页签。

步骤4 查看“HDFS存储”表格。

- 指定租户目录的“文件目录数上限”列表示文件和目录数量配额。
- 指定租户目录的“存储空间配额 ”列表示租户目录的存储空间大小。

----结束

添加租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，单击“添加目录”。

- “父目录”，表示当前租户对应父租户的存储目录。

  说明

  当前租户不是子租户则不显示此参数。

- “路径”，填写租户目录的路径。

  说明

  当前租户不是子租户则新路径将在HDFS的根目录下创建。

- “文件\目录数上限”填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  当前已使用的文件数的数据采集周期为1个小时，因此超过文件数阈值的告警上报会存在延迟。

- “存储空间配额”，填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  已使用的存储空间的数据采集周期为1个小时，因此超过存储空间阈值的告警上报会存在延迟。

步骤5 单击“确定”完成租户目录添加。

----结束

修改租户目录属性

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。

- “文件\目录数上限”，填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。
- “存储空间配额”填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

步骤5 单击“确定”完成租户目录修改。

----结束

删除租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“删除”。

说明

不支持删除创建租户时系统创建的租户目录。

步骤5 单击“确定”完成租户目录删除。

----结束

11.7.2.2.2 恢复租户数据

操作场景

租户默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，管理员需要通过FusionInsight Manager手动恢复配置数据。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“租户资源状态”，绿色表示租户可用，灰色表示租户不可用。

2. 单击“资源”，查看“Yarn”或者“HDFS存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击，在弹出的确认窗中输入当前登录的用户密码确认身份，单击“确定”。

步骤5 在“恢复租户资源数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

11.7.2.2.3 删除租户

操作场景

根据业务需求，对于当前不再使用的租户，管理员可以通过FusionInsight Manager删除租户，释放租户占用的资源。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择待删除的租户，单击。

说明

- 根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户资源的数据。”，否则将自动删除租户对应的存储空间。

步骤3 单击“确定”，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

租户删除后，Yarn中对应的租户任务队列不会被删除。同时Yarn角色管理中，此租户任务队列不再显示。

----结束

11.7.2.3 管理资源

11.7.2.3.1 添加资源池

操作场景

在集群中，管理员可从逻辑上对所有Yarn的节点进行分区，使多个NodeManager形成一个Yarn资源池。每个NodeManager只能属于一个资源池。管理员通过FusionInsight Manager添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池，便于指定的队列利用这些计算资源。

系统中默认包含了一个名为“default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 单击“添加资源池”。

步骤4 设置资源池的属性。

- “集群”：选择待添加资源池的集群名称。
- “名称”：填写资源池的名称。长度为1~50个字符，可包含数字、字母或下划线（_），且不能以下划线（_）开头。
- “资源标签”：配置资源池的资源标签，包括数字、字母、下划线（_）或减号（-），长度为1~50个字符，且只能以数字或者字母开头。
- “资源”：在界面左边可用主机列表中，勾选指定的主机，单击 >> ，将选中的主机加入已选主机列表。只支持选择本集群中的主机。资源池中的主机列表可以为空。

  说明

  根据业务需求，可以通过主机名称 、核数、内存、操作系统、平台类型，筛选需要选取的资源主机。

步骤5 单击“确定”保存。

完成资源池创建后，管理员可以在资源池的列表中查看资源池的名称、成员、类型。已加入自定义资源池的主机，不再是“default”资源池的成员。

----结束

11.7.2.3.2 修改资源池

操作场景

根据业务需要，资源池的主机需要调整时，管理员可以通过FusionInsight Manager修改已有资源池中的成员。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“编辑”。

步骤4 在“资源”修改“主机”。

- 增加主机：在界面左边主机列表，选择指定的主机名称，单击 >> ，将选中的主机加入资源池。
- 删除主机：在界面右边主机列表，选择指定的主机名称，单击 << ，将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

11.7.2.3.3 删除资源池

操作场景

根据业务需要，资源池不再使用时，管理员可以通过FusionInsight Manager进行删除资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列容量配置**。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

步骤4 在弹出窗口中单击“确定”。

----结束

11.7.2.3.4 配置队列

操作场景

根据业务需求，管理员可以在FusionInsight Manager修改指定租户的队列配置。

前提条件

已添加使用Superior调度器的租户。

操作步骤

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 单击“动态资源计划”页签。

步骤3 单击“队列配置”页签。

步骤4 “集群”参数选择待操作的集群名称，然后指定租户资源名的“操作”列，单击“修改”。

说明

- 在“租户资源管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击“队列配置（*队列名*）”后面的也能打开修改队列配置页面。
- 一个队列只能绑定一个非default资源池。
- 对于“每个YARN容器最多分配核数”、“每个YARN容器最大分配内存（MB）”、“最多运行任务数”、“每个用户最多运行任务数”和“最多挂起任务数”等参数，为便于操作，当子租户值为-1时，父租户值可以设置为具体限制值；当父租户设置为具体限制值时，子租户可以设置为-1。
- “每个YARN容器最多分配核数”和“每个YARN容器最大分配内存（MB）”需要同时修改为非-1的值才会生效。

表 11-38 队列配置参数

参数名	描述
AM最多占有资源（%）	表示当前队列内所有Application Master所占的最大资源百分比。
每个YARN容器最多分配核数	表示当前队列内单个YARN容器可分配的最多核数，默认为-1，表示取值范围内不限制。
每个YARN容器最大分配内存（MB）	表示当前队列内单个YARN容器可分配的最大内存，默认为-1，表示取值范围内不限制。
最多运行任务数	表示当前队列最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
每个用户最多运行任务数	表示每个用户在当前队列中最多同时可执行任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可执行任务。取值范围为-1～2147483647。
最多挂起任务数	表示当前队列最多同时可挂起任务的数目，默认为-1，表示取值范围内不限制（为空意义相同），为0表示不可挂起任务。取值范围为-1～2147483647。
资源分配规则	表示单个用户任务间的资源分配规则，包括FIFO和FAIR。 一个用户若在当前队列上提交了多个任务，FIFO规则代表一个任务完成后再执行其他任务，按顺序执行。FAIR规则代表各个任务同时获取到资源并平均分配资源。
默认资源标签	表示在指定资源标签（Label）的节点上执行任务。
Active状态	• ACTIVE表示当前队列可接受并执行任务。 • INACTIVE表示当前队列可接受但不执行任务，若提交任务，任务将处于挂起状态。

参数名	描述
Open状态	● OPEN表示当前队列处于打开状态。 ● CLOSED表示当前队列处于关闭状态，若提交任务，任务直接会被拒绝。
故障时是否队列迁移	集群开启单集群跨AZ高可用时，如果AZ故障后，需要当该租户正在运行的队列重新提交至其他AZ，可设置“故障时是否队列迁移”参数为“是”。

步骤5 单击“确定”完成配置。

----结束

11.7.2.3.5 配置资源池的队列容量策略

操作场景

添加资源池后，需要为Yarn任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。

该任务指导系统管理员通过FusionInsight Manager配置队列策略。使用Superior调度器的租户队列，可以配置使用不同资源池的资源。

前提条件

- 已登录FusionInsight Manager。
- 已添加资源池。
- 任务队列不与其他队列相关联资源池，除了默认资源池。

操作步骤

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 单击“动态资源计划”页签。

步骤3 单击“资源分布策略”页签。

步骤4 “集群”参数选择待操作的集群名称，然后在“资源池”选择指定的资源池。

步骤5 在“资源分配”列表指定队列的“操作”列，单击“修改”。

步骤6 在“修改资源分配”窗口的“资源配置策略”页签设置任务队列在此资源池中的资源配置策略。

- “权重”：在发生资源抢占时，权重越大则会优先抢占资源。其初始值与最小资源百分比值一致。
- “最小资源”：表示租户能获得的最少资源。
- “最大资源”：表示租户能获得的最多资源。
- “预留资源”：表示保留给租户自身队列，且不能借用给其他租户队列的资源。

步骤7 在“修改资源分配”窗口的“用户策略”页签设置用户策略。

说明

defaultUser(built-in)表示如果一个用户未配置策略，则默认使用defaultUser所指定的策略。该策略不可删除。

- 单击“添加用户策略”添加用户策略。
  - “用户名”：表示用户的名称。
  - “权重”：在发生资源抢占时，权重越大则会优先抢占资源。
  - “最多核数”：表示用户最多可以使用的虚拟核数。
  - “最大内存”：表示用户最大可以使用的内存。
- 单击“操作”列的“修改”修改现有用户策略。
- 单击“操作”列的“清除”删除现有用户策略。

步骤8 单击“确定”保存配置。

----结束

11.7.2.3.6 清除队列容量配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，管理员可以在FusionInsight Manager清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

操作步骤

步骤1 登录FusionInsight Manager界面。

步骤2 选择“租户资源 > 动态资源计划”。

步骤3 “集群”参数选择待操作的集群名称，然后在“资源池”，选择待操作的资源池。

步骤4 在资源分配表格，指定租户资源名的“操作”列，单击“清除”。

步骤5 在弹出的对话框中单击“确定”，清除队列在当前资源池的配置。

----结束

11.7.2.4 管理全局用户策略

操作场景

如果租户配置使用Superior调度器，那么系统可以控制具体用户使用资源调度器的行为，包含：

- 最大运行任务数
- 最大挂起任务数

- 默认队列

操作步骤

- 添加策略
  a. 在FusionInsight Manager，单击“租户资源”。
  b. 单击“动态资源计划”页签。
  c. 单击“全局用户策略”页签。

     说明

     defaults(default setting)表示如果一个用户未配置全局用户策略，则默认使用defaults所指定的策略。该策略不可删除。

  d. 单击“添加全局用户策略”，在弹出窗口中填写以下参数。
     - 集群：选择需要操作的集群。
     - 用户名：表示需要控制资源调度的用户，请输入当前集群中已存在用户的名称。
     - 最大运行任务数：表示该用户在当前集群中能运行的最大任务数量。
     - 最大挂起任务数：表示该用户在当前集群中能挂起的最大任务数量。
     - 默认队列：表示用户的队列，请输入当前集群中已存在队列的名称。
- 修改策略
  a. 在FusionInsight Manager，单击“租户资源”。
  b. 单击“动态资源计划”页签。
  c. 单击“全局用户策略”页签。
  d. 在指定用户策略所在行，单击“操作”列中的“修改”。
  e. 调整相关参数后，单击“确定”。
- 删除策略
  a. 在FusionInsight Manager，单击“租户资源”。
  b. 单击“动态资源计划”页签。
  c. 单击“全局用户策略”页签。
  d. 在指定用户策略所在行，单击“操作”列中的“删除”。
     在弹出窗口单击“确定”。

11.7.3 使用 Capacity 调度器的租户业务

11.7.3.1 创建租户

11.7.3.1.1 添加租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建租户，以满足实际使用场景。

前提条件

- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 单击，打开添加租户的配置页面，参见表11-39为租户配置属性。

表 11-39 租户参数一览

参数名	描述
集群	选择要创建租户的集群。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
租户资源类型	指定租户是否是一个叶子租户： • 选择“叶子租户资源”：当前租户为叶子租户，不支持添加子租户。 • 选择“非叶子租户资源”：当前租户为非叶子租户，支持添加子租户。 **说明** MRS 3.2.0及之后版本，如果关联服务需要选择“ClickHouse”，租户资源类型只能选择“叶子租户”。
计算资源	为当前租户选择动态计算资源。 • 选择“Yarn”时，系统自动在Yarn中以租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 • 不选择“Yarn”时，系统不会自动创建任务队列。
配置模式	计算资源参数配置模式。 • 选择“基础”时，只需配置“默认资源池容量（%）”参数即可。 • 选择“高级”时，可手动配置资源分配权重，租户的最小/最大/预留资源。

参数名	描述
默认资源池容量 （%）	配置当前租户在默认资源池中使用的计算资源百分比，取值范围0~100%。
权重	资源分配权重，取值范围从0到100。
最小资源	保证租户资源能获得的资源（有抢占支持）。取值可以是父租户资源的百分比或绝对值。当租户资源作业量比较少时，资源会自动借给其他租户资源，当租户资源能使用的资源不满足最小资源时，可以通过抢占来要回之前借出的资源。
最大资源	租户资源最多能使用的资源，租户资源不能得到比最大资源设定更多的资源。取值可以是父租户资源的百分比或绝对值。
预留资源	租户资源预留资源。即使租户资源内没有作业，预留的资源也不能给别的租户资源使用。取值可以是父租户资源的百分比或绝对值。
存储资源	为当前租户选择存储资源。 ● 选择“HDFS”时，系统将分配存储资源。 ● 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。
存储空间配额	配置当前租户使用的HDFS存储空间配额。 ● 取值范围：当存储空间配额单位设置为MB时，范围为1～8796093022208。当存储空间配额单位设置为GB时，范围为1～8589934592。 ● 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 ● 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。
存储路径	配置租户在HDFS中的存储目录。 ● 系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“/tenant/ta1”。 ● 第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
服务	是否需要关联使用其他服务的资源，参见**步骤4**。
描述	配置当前租户的描述信息。

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 创建租户时系统会自动创建一个Yarn任务队列，并自动以租户名称命名该队列。如果已经存在同名队列，新队列命名为“租户名称-N”。“N”表示从1开始的自然数，存在同名队列的时候N会自动累加以区别已有队列。例如“saletenant”、“saletenant-1”和“saletenant-2”。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

- “服务”选择“HBase”，在“关联类型”可以选择“独占”或者“共享”。

  说明

  - “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
  - “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

- MRS 3.2.0及之后版本：“服务”选择“ClickHouse”。
  - 关联类型”：当“服务”选择“ClickHouse”时，关联类型”只支持“共享”。
  - “关联逻辑集群”：若ClickHouse没有开启逻辑集群，则默认关联default_cluster，若已经开启逻辑集群，则按需选择需要关联的逻辑集群。
  - “CPU优先级”：CPU优先级取值范围为-20~19，该值关联OS的NICE值，取值越小，CPU优先级越高。
  - “内存”：内存限制为百分比，如该值设置为80，则当前租户可使用的内存总额为：服务可使用内存 * 80%。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

1. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

11.7.3.1.2 添加子租户

操作场景

根据业务对资源消耗以及隔离的规划与需求，管理员可以通过FusionInsight Manager创建子租户，将当前租户的资源进一步分配以满足实际使用场景。

前提条件

- 已添加父租户，且属于非叶子租户。
- 已根据业务需求规划租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
- 已规划当前租户可分配的资源，确保每一级别租户下，直接子租户的资源之和不超过当前租户。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择父租户节点然后单击，打开添加子租户的配置页面，参见表11-40为子租户配置属性。

表 11-40 子租户参数一览

参数名	描述
集群	显示上级父租户所在集群。
父租户资源	显示上级父租户的名称。
名称	• 指定当前租户的名称，长度为3~50个字符，可包含数字、字母或下划线（_）。 • 根据业务需求规划子租户的名称，不得与当前集群中已有的角色、HDFS目录或者Yarn队列重名。
租户类型	指定租户是否是一个叶子租户： • 选择“叶子租户”：当前租户为叶子租户，不支持添加子租户。 • 选择“非叶子租户”：当前租户为非叶子租户，支持添加子租户，但租户层级不能超过5层。
计算资源	为当前租户选择动态计算资源。 • 选择“Yarn”时，系统自动在Yarn中以子租户名称创建任务队列。 – 如果是叶子租户，叶子租户可直接提交到任务队列中。 – 如果是非叶子租户，非叶子租户不能直接将任务提交到队列中。但是，Yarn会额外为非叶子租户增加一个任务队列（隐含），队列默认命名为“default”，用于统计当前租户剩余的资源容量，实际任务不会分配在此队列中运行。 • 不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量（%）	配置当前租户使用的计算资源百分比，基数为父租户的资源总量。
默认资源池最大容量（%）	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。

参数名	描述
存储资源	为当前租户选择存储资源。 • 选择“HDFS”时，系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。 • 不选择“HDFS”时，系统不会分配存储资源。
文件\目录数上限	配置文件和目录数量配额。
存储空间配额	配置当前租户使用的HDFS存储空间配额。 • 当存储空间配额单位设置为MB时，范围为1～8796093022208，当“存储空间配额单位”设置为GB时，范围为1～8589934592。 • 此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。 • 如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 • 如果此配额大于父租户的配额，实际存储量不超过父租户配额。
存储路径	配置租户在HDFS中的存储目录。 • 系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“/tenant/ta1”，系统默认自动配置此参数值为“/tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。 • 支持在父目录中自定义存储路径。
描述	配置当前租户的描述信息

说明

创建租户时将自动创建租户对应的角色、计算资源和存储资源。

- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“系统 > 权限> 角色”进行手动管理，角色名称为“*租户名称_集群ID*”。首个集群的集群ID默认不显示。
- 使用此租户时，请创建一个系统用户，并绑定租户对应的角色。具体操作请参见**添加用户并绑定租户的角色**。
- 子租户可以将当前租户的资源进一步分配。每一级别父租户下，直接子租户的资源百分比之和不能超过100%。所有一级租户的计算资源百分比之和也不能超过100%。

步骤3 当前租户是否需要关联使用其他服务的资源?

- 是，执行**步骤4**。
- 否，执行**步骤5**。

步骤4 单击“关联服务”，配置当前租户关联使用的其他服务资源。

1. 在“服务”选择“HBase”。
2. 在“关联类型”选择：

- “独占”表示该租户独占服务资源，其他租户不能再关联此服务。
- “共享”表示共享服务资源，可与其他租户共享使用此服务资源。

说明

- 创建租户时，租户可以关联的服务资源只有HBase。为已有的租户关联服务时，可以关联的服务资源包含：HDFS、HBase和Yarn。
- 若为已有的租户关联服务资源：在租户列表单击目标租户，切换到“服务关联”页签，单击“关联服务”单独配置当前租户关联资源。
- 若为已有的租户取消关联服务资源：在租户列表单击目标的租户，切换到“服务关联”页签，单击“删除”，并勾选“我已阅读此信息并了解其影响。”，再单击“确定”删除与服务资源的关联。

3. 单击“确定”。

步骤5 单击“确定”，等待界面提示租户创建成功。

----结束

11.7.3.1.3 添加用户并绑定租户的角色

操作场景

创建好的租户不能直接登录集群访问资源，管理员需要通过FusionInsight Manager为已有租户创建新用户，通过绑定租户的角色继承其操作权限，以满足业务使用。

前提条件

管理员已明确业务需求，并已创建了租户。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 用户”。

步骤2 若在系统中添加新的用户，请单击“添加用户”，打开添加用户的配置页面。

若为系统中已有的用户绑定租户权限，请单击该用户所在行的“修改”，打开修改用户的配置页面。

参见**表11-41**为用户配置属性。

表 11-41 用户参数一览

参数名	描述
用户名	指定当前的用户名，长度为3~32个字符，可包含数字、字母、下划线（_）、中划线（-）或空格。 • “用户名”不能与集群各节点所有操作系统用户名相同，否则此用户无法正常使用。 • 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。使用“User1”时请输入正确的用户名。

参数名	描述
用户类型	可选值包括“人机”和“机机”。 ● “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。 ● “机机”用户：用于应用开发的场景。选择该值用户密码随机生成，无需填写。
密码	选择“人机”用户需填写“密码”。 密码必须包含8~64个字符，至少包含以下类型字符中的四种：大写字母、小写字母、数字、特殊字符和空格。不能与用户名或倒序的用户名相同。
确认密码	再次输入密码。
用户组	单击“添加”，选择对应用户组将用户添加进去。 ● 如果用户组添加了角色，则用户可获得对应角色中的权限。 ● 例如，为新用户分配Hive的权限，请将用户加入Hive组。
主组	选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。
角色	单击“添加”为用户绑定租户的角色。 **说明** 若一个用户想要获取使用“tenant1”租户包含的资源，且能够为“tenant1”租户添加/删除子租户，则需要同时绑定“Manager_tenant”和“tenant1_*集群ID*”两个角色。
描述	配置当前用户的描述信息。

步骤3 单击“确定”完成用户创建。

----结束

11.7.3.2 管理租户

11.7.3.2.1 管理租户目录

操作场景

管理员通过FusionInsight Manager管理指定租户使用的HDFS存储目录，能根据业务需求对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

已添加具有HDFS存储资源的租户。

查看租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击目标的租户。

步骤3 单击“资源”页签。

步骤4 查看“HDFS存储”表格。

- 指定租户目录的“文件目录数上限”列表示文件和目录数量配额。
- 指定租户目录的“存储空间配额 ”列表示租户目录的存储空间大小。

----结束

添加租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，单击“添加目录”。

- “父目录”，表示当前租户对应父租户的存储目录。

  说明

  当前租户不是子租户则不显示此参数。

- “路径”，填写租户目录的路径。

  说明

  当前租户不是子租户则新路径将在HDFS的根目录下创建。

- “文件\目录数上限”填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  当前已使用的文件数的数据采集周期为1个小时，因此超过文件数阈值的告警上报会存在延迟。

- “存储空间配额”，填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

  说明

  已使用的存储空间的数据采集周期为1个小时，因此超过存储空间阈值的告警上报会存在延迟。

步骤5 单击“确定”完成租户目录添加。

----结束

修改租户目录属性

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要修改HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。

- “文件\目录数上限”，填写文件和目录数量配额。
- 文件数阈值配置（%），只有设置了“文件\目录数上限”才会生效。表示当已使用的文件数超过了设置的“文件\目录数上限”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。
- “存储空间配额”填写租户目录的存储空间大小。
- 存储空间阈值配置（%），表示已使用存储空间超过了设置的“存储空间配额”的百分数后将会产生告警。不设置则不会根据实际使用情况上报告警。

步骤5 单击“确定”完成租户目录修改。

----结束

删除租户目录

步骤1 在FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击需要删除HDFS存储目录的租户。

步骤3 单击“资源”页签。

步骤4 在“HDFS存储”表格，指定租户目录的“操作”列，单击“删除”。

说明

不支持删除创建租户时系统创建的租户目录。

步骤5 单击“确定”完成租户目录删除。

----结束

11.7.3.2.2 恢复租户数据

操作场景

租户默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，管理员需要通过FusionInsight Manager手动恢复配置数据。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“租户资源状态”，绿色表示租户可用，灰色表示租户不可用。

2. 单击“资源”，查看“Yarn”和“HDFS存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击，在弹出的确认窗中输入当前登录的用户密码确认身份，单击“确定”。

步骤5 在“恢复租户资源数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

11.7.3.2.3 删除租户

操作场景

根据业务需求，对于当前不再使用的租户，管理员可以通过FusionInsight Manager删除租户，释放租户占用的资源。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。

操作步骤

步骤1 登录FusionInsight Manager，单击“租户资源”。

步骤2 在左侧租户列表，选择待删除的租户，单击。

说明

- 根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户的数据。”，否则将自动删除租户对应的存储空间。
- 如果使用不属于supergroup组的用户执行删除租户操作，并且不保留租户数据，需要使用属于supergroup组的用户登录HDFS客户端，手动清理租户对应的存储空间，以免数据残留。

步骤3 单击“确定”，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

租户删除后，Yarn中对应的租户任务队列不会被删除。同时Yarn角色管理中，此租户任务队列不再显示。

----结束

11.7.3.2.4 Capacity Scheduler 模式下清除租户非关联队列

操作场景

在Yarn Capacity Scheduler模式下，删除租户的时候，只是把租户队列的容量设置为0，并且把状态设为“STOPPED”，但是队列在Yarn的服务里面仍然残留。由于Yarn的机制，无法动态删除队列，管理员可以执行命令手动清除残留的队列。

对系统的影响

- 脚本运行过程中会重启controller服务，同步Yarn的配置，并重启主备ResourceManager实例。
- 重启controller服务时，无法登录和操作FusionInsight Manager。
- 重启主备ResourceManager实例后，Yarn组件以及依赖Yarn的组件会出现短暂的服务不可用告警。

前提条件

已删除某个租户，但该租户对应的队列依然存在。

操作步骤

步骤1 确定该租户对应的队列依然存在。

1. 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 服务 > Yarn”。通过“ResourceManager(主)”链接进入ResourceManager WebUI界面。
2. 单击左侧“Scheduler”界面，可以查看租户对应的队列依然存在，且状态为“STOPPED”，“Configured Capacity”值为0。

步骤2 以**omm**用户登录主管理节点。

步骤3 执行以下目录，执行“cleanQueuesAndRestartRM.sh”脚本。

cd ${BIGDATA_HOME}/om-server/om/sbin

./cleanQueuesAndRestartRM.sh -c *集群ID*

说明

“集群ID”为需执行操作集群ID号，可在FusionInsight Manager的“集群 > *待操作集群的名称* > 集群属性”中查看。

在脚本运行过程中，需输入**yes**及管理员密码。

```
Running the script will restart Controller and restart ResourceManager.
Are you sure you want to continue connecting (yes/no)?yes
Please input admin password:
Begin to backup queues ...
...
```

步骤4 脚本运行成功后，在FusionInsight Manager界面，选择“集群 > *待操作集群名称* > 服务 > Yarn”。通过“ResourceManager(主)”链接进入ResourceManager WebUI界面。

步骤5 单击左侧“Scheduler”界面，确认被删除租户的队列已经清除。

----结束

11.7.3.3 管理资源

11.7.3.3.1 添加资源池

操作场景

在集群中，管理员可从逻辑上对所有Yarn的节点进行分区，使多个NodeManager形成一个Yarn资源池。每个NodeManager只能属于一个资源池。管理员通过FusionInsight Manager添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池，便于指定的队列利用这些计算资源。

系统中默认包含了一个名为“default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 单击“添加资源池”。

步骤4 设置资源池的属性。

- “集群”：选择待添加资源池的集群名称。
- “名称”：填写资源池的名称。长度为1~50个字符，可包含数字、字母或下划线（_），且不能以下划线（_）开头。
- “资源标签”：配置资源池的资源标签，包括数字、字母、下划线（_）或减号（-），长度为1~50个字符，且只能以数字或者字母开头。
- “资源”：在界面左边可用主机列表中，勾选指定的主机，单击 >> ，将选中的主机加入已选主机列表。只支持选择本集群中的主机。资源池中的主机列表可以为空。

  说明

  根据业务需求，可以通过主机名称 、核数、内存、操作系统、平台类型，筛选需要选取的资源主机。

步骤5 单击“确定”保存。

完成资源池创建后，管理员可以在资源池的列表中查看资源池的名称、成员、类型。已加入自定义资源池的主机，不再是“default”资源池的成员。

----结束

11.7.3.3.2 修改资源池

操作场景

根据业务需要，资源池的主机需要调整时，管理员可以通过FusionInsight Manager修改已有资源池中的成员。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“编辑”。

步骤4 在“资源”修改“主机”。

- 增加主机：在界面左边主机列表，选择指定的主机名称，单击 >> ，将选中的主机加入资源池。
- 删除主机：在界面右边主机列表，选择指定的主机名称，单击 << ，将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

11.7.3.3.3 删除资源池

操作场景

根据业务需要，资源池不再使用时，管理员可以通过FusionInsight Manager进行删除资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列容量配置**。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 资源池”。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

步骤4 在弹出窗口中单击“确定”。

----结束

11.7.3.3.4 配置队列

操作场景

根据业务需要，管理员可以通过FusionInsight Manager修改指定租户的队列配置。

前提条件

已添加使用Capacity调度器的租户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 动态资源计划”。

默认显示“资源分布策略”。

步骤3 单击“队列配置”页签。

步骤4 “集群”参数选择待操作的集群名称，然后在指定租户资源名的“操作”列，单击“修改”。

说明

- 在“租户资源管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击“队列配置（*队列名*）”名后面的也能打开修改队列配置窗口。
- 一个队列只能绑定一个非default资源池，即新添加的资源池只能绑定一个队列，作为这个队列的默认资源池。

表 11-42 队列配置参数

参数名	描述
租户资源名（队列）	租户及队列名称。
最大应用数量	表示最大应用程序数量。
AM最大资源百分比	表示集群中可用于运行application master的最大资源占比。
用户资源最小上限百分比（%）	表示每个用户最低资源保障（百分比）。任何时刻，一个队列中每个用户可使用的资源量均有一定的限制。当一个队列中同时运行多个用户的应用程序时，每个用户的使用资源量在一个最小值和最大值之间浮动，其中，最小值取决于正在运行的应用程序数目，而最大值则由此参数决定。 比如，假设此参数的值设置为25。当两个用户向该队列提交应用程序时，每个用户可使用资源量不能超过50%，如果三个用户提交应用程序，则每个用户可使用资源量不能超多33%，如果四个或者更多用户提交应用程序，则每个用户可用资源量不能超过25%。
用户资源上限因子	表示用户使用的最大资源限制因子，与当前租户在集群中实际资源百分比相乘，可计算出用户使用的最大资源百分比。
状态	表示资源计划当前的状态，“运行”为运行状态，“停止”为停止状态。
默认资源池	表示队列使用的资源池，默认为“default”。 如果需要修改为其他资源池，需要先配置队列容量，请参见**配置资源池的队列容量策略**。

步骤5 单击“确定”完成配置。

----结束

11.7.3.3.5 配置资源池的队列容量策略

操作场景

添加资源池后，需要为Yarn任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。每个队列只能配置一个资源池的队列容量策略。

管理员可以在任何一个资源池中查看队列并配置队列容量策略。配置队列策略后，Yarn任务队列与资源池形成关联关系。

前提条件

已添加队列，即已创建关联了计算资源的租户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“租户资源 > 动态资源计划”。

默认显示“资源分布策略”。

步骤3 “集群”参数选择待操作的集群名称，然后在“资源池”，选择待操作的资源池。

步骤4 在“资源分配”表格，指定租户资源名队列的“操作”列，单击“修改”。

步骤5 在“修改资源分配”窗口设置任务队列在此资源池中的资源容量策略。

- “资源容量（%）”：表示当前租户计算资源使用的资源百分比。
- “最大资源容量（%）”：表示当前租户计算资源使用的最大资源百分比。

步骤6 单击“确定”保存配置。

说明

删除队列的资源容量值并保存，可以取消队列在此资源池中的资源容量策略，表示解除队列与资源池的关联关系。需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

----结束

11.7.3.3.6 清除队列容量配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，管理员可以在FusionInsight Manager清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

操作步骤

步骤1 登录FusionInsight Manager界面。

步骤2 选择“租户资源 > 动态资源计划”。

步骤3 “集群”参数选择待操作的集群名称，然后在“资源池”，选择待操作的资源池。

步骤4 在资源分配表格，指定租户资源名的“操作”列，单击“清除”。

步骤5 在弹出的对话框中单击“确定”，清除队列在当前资源池的配置。

----结束

11.7.4 切换调度器

操作场景

新安装的MRS集群默认即使用了Superior调度器，如果是历史版本升级的集群，集群管理员可以根据以下指导，将Yarn的调度器从Capacity调度器一键式切换到Superior调度器。

前提条件

- 确保集群网络通畅，网络环境安全，Yarn服务状态正常。
- 在切换调度器期间，不允许做添加、删除、修改租户，以及启停服务等操作。

对系统的影响

- 调度器切换过程中，由于要重启Resource Manager，因此切换期间向Yarn提交任务会失败。
- 调度器切换过程中，正在Yarn上面执行的Job的Task任务会继续执行，但不会启动新的Task。
- 调度器切换完成后，在Yarn上面执行的任务有可能会失败进而导致业务中断。
- 调度器切换完成后，在租户管理中将使用Superior的相关参数。
- 调度器切换完成后，Capacity调度器中“资源容量”为“0”的租户队列在Superior调度器中分配不到资源，提交到该租户队列的任务会执行失败。建议在Capacity调度器中不要将租户队列的“资源容量”配置为“0”。
- 调度器切换完成后，在观察期内，不允许对资源池、Yarn节点标签（Label）和租户做添加、删除的操作。若添加或者删除了资源池、Yarn节点标签（Label）或租户，将不支持回退到Capacity调度器。

  说明

  - 切换调度器观察期建议为一周，如果对资源池、Yarn节点标签（Label）或租户做了添加、删除的操作，将视为观察期结束。
- 回退可能会丢失部分或者所有的Yarn任务信息。

从 Capacity 调度器切换到 Superior 调度器

步骤1 修改Yarn服务参数，并确保Yarn服务状态正常。

1. 使用管理员账号，登录FusionInsight Manager系统。

2. 选择“集群 > 服务 > Yarn > 配置 > 全部配置”，搜索并查看“yarn.resourcemanager.webapp.pagination.enable”参数值是否为“true”。
   - 是，执行**步骤1.3**。
   - 否，将参数值设置为“true”，单击“保存”保存配置。在Yarn服务“概览”页面，选择“更多 > 重启服务”，验证身份后，单击“确定”。等待服务重启成功后，执行**步骤1.3**。
3. 选择“集群 > *待操作的集群名称* > 服务”，查看Yarn服务的状态是否正常。

步骤2 使用**omm**用户登录主管理节点。

步骤3 执行调度器切换。

调度器切换分为三种模式：

0：将Capacity调度器配置转换到Superior，然后将Capacity调度器切换到Superior。

1：只将Capacity调度器配置转换到Superior。

2：只将Capacity调度器切换到Superior。

- 集群环境相对简单，租户数小于20的情况下，建议执行模式0，将Capacity调度器配置转换到Superior的同时切换调度器。

  执行以下命令。

  sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 0**

  说明

  “集群ID”为需执行操作集群ID号，可在FusionInsight Manager的“集群 > *待操作集群的名称* > 集群属性”中查看。

  ```
  Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
  Start to convert Capacity scheduler configurations to Superior. Please wait...
  Convert configurations successfully.
  Start to switch the Yarn scheduler to Superior. Please wait...
  Switch the Yarn scheduler to Superior successfully.
  ```

- 集群环境相对复杂，租户信息复杂，且要求将capacity调度器队列配置信息保留到Superior调度器，建议先执行模式1，将Capacity调度器配置信息转化成Superior配置信息，对转换过来的配置信息做检查后，再执行模式2，将Capacity调度器切换到Superior。

  a. 执行以下命令，将Capacity调度器配置信息转化成Superior配置信息。

  sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 1**

  ```
  Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
  Start to convert Capacity scheduler configurations to Superior. Please wait...
  Convert configurations successfully.
  ```

  b. 执行以下命令，将Capacity调度器切换到Superior。

  sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 2**

  ```
  Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
  Start to switch the Yarn scheduler to Superior. Please wait...
  Switch the Yarn scheduler to Superior successfully.
  ```

- 不保存Capacity调度器队列配置，建议直接执行模式2，只切换调度器，不转换配置。

  a. 登录FusionInsight Manager，删除除了default租户的所有租户。

b. 登录FusionInsight Manager，删除除了default资源池的所有资源池。

执行以下命令，将Capacity调度器切换到Superior。

sh ${BIGDATA_HOME}/om-server/om/sbin/switchScheduler.sh -c *集群ID* **-m 2**

```
Start to convert Capacity scheduler to Superior Scheduler, clusterId=1
Start to switch the Yarn scheduler to Superior. Please wait...
Switch the Yarn scheduler to Superior successfully.
```

说明

登录主管理节点，可查看调度器切换的日志信息。

- ${BIGDATA_LOG_HOME}/controller/aos/switch_scheduler.log
- ${BIGDATA_LOG_HOME}/controller/aos/aos.log

----结束

11.8 系统设置

11.8.1 权限设置

11.8.1.1 用户管理

11.8.1.1.1 创建用户

操作场景

FusionInsight Manager最大支持50000个用户（包括系统内置用户）。默认情况下，系统只有一个用户“admin”具有FusionInsight Manager最高操作权限。管理员应根据实际业务场景需要，通过FusionInsight Manager创建新用户并指定其操作权限以满足业务使用。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在用户列表上方，单击“添加用户”。

步骤4 填写“用户名”。用户名由数字、字母、下划线、中划线（-）或空格组成，不区分大小写，不能与系统或操作系统中已有的用户名相同。

步骤5 设置“用户类型”，可选值包括“人机”和“机机”。

- “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时选择“密码策略”并填写“密码”和“确认密码”。
- “机机”用户：用于组件应用开发的场景。选择该值则用户密码随机生成，无需填写。

步骤6 根据业务实际需要，在“用户组”，单击“添加”，选择一个或多个用户组添加到列表中。

说明

- 如果选中的用户组绑定了角色或者在Ranger中配置了权限策略，用户将获得对应的权限。
- 安装FusionInsight Manager后默认生成的部分用户组包含特殊权限，请根据界面上用户组描述信息选择正确的用户组。
- 如果已有的用户组无法满足使用，可以单击“创建新用户组”先创建用户组，参见**添加用户组**。

步骤7 根据业务实际需要，在“用户组”添加的所有组中选择一个组作为用户创建目录和文件的主组。

下拉列表包含“用户组”中添加的全部组。

说明

由于一个用户可以属于多个组（包括主组和附属组，主组只有一个，附属组可以有多个），设置用户的主组是为便于维护以及遵循hadoop社区的权限机制。此外用户的主组和其他组在权限控制方面，作用一致。

步骤8 根据业务实际需要，在“角色”，单击“添加”，为单个用户绑定角色。

说明

- 创建用户时添加角色可细化用户的权限。
- 创建用户时，如果用户从用户组获得的权限还不满足业务需要，则可以再分配其他已创建的角色。也可以单击“创建新角色”先创建角色，参见**添加角色**。

  为新用户分配角色授权，最长可能需要3分钟时间生效，如果从用户组获得的权限已满足使用，则无需再添加角色。
- 组件启用Ranger鉴权后，除系统默认用户组或角色的权限外，其他权限需要通过配置Ranger策略为用户赋权。
- 若用户既没有加入用户组也没有设置角色，通过此用户登录FusionInsight Manager后，用户将无权查看或操作。

步骤9 根据业务实际需要填写“描述”。

步骤10 单击“确定”完成用户创建。

“人机”用户创建成功后，通常需要修改初始密码后才可以正常使用，可以使用该用户登录FusionInsight Manager，按照界面提示重置密码即可。

----结束

11.8.1.1.2 修改用户信息

操作场景

管理员可以在FusionInsight Manager修改已创建的用户信息，包括修改用户组、主组、角色分配权限和描述。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要修改信息的用户所在行，单击“修改”。

根据实际情况，修改对应参数。

说明

修改用户的用户组，或者修改用户的角色权限，最长可能需要3分钟时间生效。

MRS 3.1.2及之后版本：

- 用户不能修改自己的密码策略绑定关系（**admin**用户除外）。
- 被锁定的用户不能修改密码策略绑定关系。
- 修改用户密码策略绑定关系后，用户下次修改密码时将会生效。
- 修改用户绑定的密码策略后，若用户密码剩余有效天数大于新绑定密码策略密码有效期时，将该用户密码有效期置为新密码策略有效期；若用户密码剩余有效天数小于新绑定密码策略密码有效期时，该用户密码有效期保持不变。

步骤4 单击“确定”完成修改操作。

----结束

11.8.1.1.3 导出用户信息

操作场景

管理员可以在FusionInsight Manager导出所有已创建的用户信息。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 单击“导出全部”，可一次性导出所有用户信息。

用户信息包含以下几个字段：用户名、创建时间、描述、用户类型（0表示人机账号，1表示机机账号）、主组、用户组列表、绑定的角色列表。

步骤4 在“保存类型”选择“TXT”或“CSV”。单击“确定”开始导出。

----结束

11.8.1.1.4 锁定用户

操作场景

由于业务变化，用户可能长期暂停使用，为了保证安全，管理员可以锁定用户。

锁定用户的方法包含以下两种方式：

- 自动锁定：通过设置密码策略中的“密码连续错误次数”，将超过登录失败次数的用户自动锁定。具体操作请参见**配置密码策略**。
- 手动锁定：由管理员通过手动的方式将用户锁定。

以下将具体介绍手动锁定。不支持锁定“机机”用户。

对系统的影响

用户被锁定后，不能在FusionInsight Manager重新登录或在集群中重新进行身份认证。锁定后的用户需要管理员手动解锁或者等待锁定时间结束才能恢复使用。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要锁定用户所在行，单击“锁定”。

步骤4 在弹出的窗口勾选“我已阅读此信息并了解其影响。”，单击“确定”完成锁定操作。

----结束

11.8.1.1.5 解锁用户

操作场景

在用户输入错误密码次数大于允许输入错误次数，造成用户被锁定的场景下，管理员可以通过FusionInsight Manager为锁定的用户解锁。仅支持解锁使用FusionInsight Manager创建的用户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要解锁用户所在行，单击“解锁”。

步骤4 在弹出的窗口勾选“我已阅读此信息并了解其影响。”，单击“确定”完成解锁操作。

----结束

11.8.1.1.6 删除用户

操作场景

根据业务需要，管理员应在FusionInsight Manager删除不再使用的系统用户。

说明

- 用户删除后，已经发放的TGT在24小时内仍然有效，用户可以使用该TGT继续进行安全认证并访问系统。
- 如新建用户与已删除用户同名，则会继承已删除用户的拥有的所有Owner权限。建议根据实际业务需求决定是否删除该用户持有的资源。例如HDFS上的文件。
- 默认的**admin**用户无法删除。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要删除用户所在行，选择“更多 > 删除”。

说明

如果需要批量删除多个用户，勾选需要删除的用户后直接单击“删除”即可。

步骤4 在弹出的窗口单击“确定”完成删除操作。

----结束

11.8.1.1.7 修改用户密码

操作场景

出于安全的考虑，“人机”类型系统用户密码必须定期修改。

如果用户具备使用FusionInsight Manager的权限时，可以通过FusionInsight Manager完成修改自身密码工作。

如果用户不具备使用FusionInsight Manager的权限时，可以通过客户端修改自身密码。

前提条件

- 已获取当前的密码策略。
- 已在集群内的任一节点安装了客户端，并获取此节点IP地址。请联系管理员获取客户端安装用户密码。

使用 FusionInsight Manager 修改密码

步骤1 登录FusionInsight Manager。

步骤2 移动鼠标到界面右上角的用户名。

在弹出菜单中单击“修改密码”。

步骤3 在“密码修改界面”分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];',<.>/\?。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。

----结束

使用客户端修改密码

步骤1 以客户端安装用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，修改系统用户密码。此操作对所有服务器生效。

kpasswd *系统用户名称*

例如，修改系统用户“test1”，执行**kpasswd test1**。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];',<.>/\?。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。

说明

如果kpasswd命令运行出错，可以尝试：

- 关闭ssh会话再重新打开。
- 执行**kdestroy**命令后再执行**kpasswd**。

----结束

11.8.1.1.8 初始化用户密码

操作场景

用户如果遗忘密码或公共账号密码需要定期修改时，管理员可通过FusionInsight Manager初始化密码。初始化密码后系统用户首次使用账号需要修改密码。

说明

此操作仅支持“人机”用户。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在要初始化密码用户所在行，选择“更多 > 初始化密码”。在弹出窗口中输入当前登录的管理员用户密码确认身份，单击“确定”，在确认对话框单击“确定”。

步骤4 填写“新密码”和“确认新密码”，单击“确定”。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];',<.>/\?。
- 不可和用户名相同或用户名的倒序字符相同。

- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。

----结束

11.8.1.1.9 导出认证凭据文件

操作场景

用户为安全模式集群进行应用开发的场景下，需要获取用户keytab文件用于安全认证。管理员可以通过FusionInsight Manager导出keytab文件。

说明

修改用户密码后，之前导出的keytab将失效，需要重新导出。

前提条件

下载“人机”用户的认证凭据文件前，需要使用Manager界面或者客户端修改过一次此用户的密码，否则下载获取的keytab文件无法使用。请参见**修改用户密码**。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在需导出文件的用户所在行，选择“更多 > 下载认证凭据”，待文件自动生成后指定保存位置，并妥善保管该文件。

认证凭据中会携带kerberos服务的“krb5.conf”文件。

解压认证凭据文件后可以获取两个文件：

- “krb5.conf”文件包含认证服务连接信息。
- “user.keytab”文件包含用户认证信息。

----结束

11.8.1.2 用户组管理

操作场景

FusionInsight Manager最大支持5000个用户组（包括系统内置用户组）。根据不同业务场景需要，管理员使用FusionInsight Manager创建并管理不同用户组。用户组通过绑定角色获取操作权限，用户加入用户组后，可获得用户组具有的操作权限。用户组同时可以达到对用户进行分类并统一管理多个用户。

前提条件

- 管理员已明确业务需求，并已创建业务场景需要的角色。
- 已登录FusionInsight Manager。

添加用户组

步骤1 选择“系统 > 权限 > 用户组”。

步骤2 在组列表上方，单击“添加用户组”。

步骤3 填写“组名”和“描述”。

“组名”由数字、字母、或下划线、中划线（-）或空格组成，不区分大小写，长度为1～64位，不能与系统中已有的用户组名相同。

步骤4 在“角色”，单击“添加”选择指定的角色并添加。

说明

- 对于已启用Ranger授权的组件（HDFS与Yarn除外），Manager上非系统默认角色的权限将无法生效，需要通过配置Ranger策略为用户组赋权。
- HDFS与Yarn的资源请求在Ranger中的策略条件未能覆盖的情况下，组件ACL规则仍将生效。

步骤5 在“用户”，单击“添加”选择指定的用户并添加。

步骤6 单击“确定”完成用户组创建。

----结束

查看用户组信息

用户组列表默认显示所有用户组。单击指定用户组名称左侧的箭头展开详细信息，可以查看此用户组中的用户数、用户以及绑定的角色。

修改用户组信息

在要修改信息用户组所在的行，单击“修改”，修改用户组信息。

导出用户组信息

单击“导出全部”，可一次性导出所有用户组信息，可导出“TXT”或者“CSV”格式。

用户组信息包含以下几个字段：用户组名、描述、用户列表、角色列表。

删除用户组

在要删除用户组所在行，单击“删除”。如果需要批量删除多个用户组，勾选需要删除的用户组后再单击列表上方“删除”即可。用户组中包含用户时，不允许删除。如需删除，请先通过修改用户组删除其包含的所有用户，再删除该用户组。

11.8.1.3 角色管理

操作场景

FusionInsight Manager最大支持5000个角色（包括系统内置角色，不包括租户自动创建的角色）。根据不同业务场景需要，管理员使用FusionInsight Manager创建并管理不同角色，通过角色对Manager和组件进行授权管理。

前提条件

- 管理员已明确业务需求。
- 登录FusionInsight Manager。

添加角色

步骤1 选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

“角色名称”由数字、字母、或下划线组成，长度为3～50位，不能与系统中已有的角色名相同。角色名称不能以Manager、System、default开头，例如角色名称不能为“Manager_test”等。

步骤3 在“配置资源权限”列表，选择待增加权限的集群，为角色选择服务权限。

在设置组件的权限时，可通过右上角的“搜索”框输入资源名称，然后单击搜索图标显示搜索结果。

搜索范围仅包含当前权限目录，无法搜索子目录。搜索关键字支持模糊搜索，不区分大小写。

说明

- 对于已启用Ranger授权的组件（HDFS与Yarn除外），Manager上非系统默认角色的权限将无法生效，需要通过配置Ranger策略为用户组赋权。
- HDFS与Yarn的资源请求在Ranger中的策略条件未能覆盖的情况下，组件ACL规则仍将生效。
- 设置组件的权限时，每次最大支持1000条权限。

步骤4 单击“确定”完成。

----结束

修改角色信息

在要修改信息角色所在的行，单击“修改”。

导出角色信息

单击“导出全部”，可一次性导出所有角色信息，可导出“TXT”或者“CSV”格式文件。

角色信息包含以下几个字段：角色名、描述、是否默认角色。

删除角色

在要删除角色所在行，单击“删除”。如果需要批量删除多个角色，勾选需要删除的角色后单击列表上方“删除”即可。角色被用户绑定时不可删除；如需删除，请先通过修改用户解除角色和用户之间的关联，再删除该角色。

任务示例（创建 Manager 角色）

步骤1 选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，在“角色名称”和“描述”输入角色名字与描述。

步骤3 在“配置资源权限”区域选择“Manager”，按照以下说明设置角色“权限”。

Manager权限：

- Cluster：
  - 查看权限：“集群”页面查看权限、“运维 > 告警”页面下“告警”、“事件”的查看权限。
  - 管理权限：“集群”、“运维”页面的管理权限。
- User：
  - 查看权限：“系统”页面下“权限”区域中内容的查看权限。
  - 管理权限：“系统”页面下“权限”区域中内容的管理权限。
- Audit ：

  管理权限：“审计”页面信息的管理权限。
- Tenant：

  管理权限：“租户”页面管理权限；“运维 > 告警”页面下“告警”、“事件”的查看权限。
- System：

  管理权限：“系统”页面除“权限”区域外，其他区域的管理权限；“运维 > 告警”页面下“告警”、“事件”的查看权限。

步骤4 单击“确定”完成。

----结束

11.8.1.4 安全策略

11.8.1.4.1 配置密码策略

操作场景

根据业务安全需要，管理员可以在FusionInsight Manager设置密码安全规则、用户登录安全规则及用户锁定规则。

须知

- 密码策略涉及用户管理的安全性，请根据企业安全要求谨慎修改，否则会有安全性风险。
- 修改密码策略之后，再修改用户密码，此时新的密码策略才会生效。

修改密码策略（MRS 3.1.2 之前版本）

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 密码策略”。

步骤3 在“操作”列单击“修改”，根据界面提示，修改密码策略。

具体参数参见表11-43。

表 11-43 密码策略参数说明

参数名称	描述
最小密码长度	密码包含的最小字符个数，取值范围为8～32。默认值为“8”。
字符类型的数目	密码字符包含大写字母、小写字母、数字、空格和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=）空格的最小种类。可选择数值为“4”和“5”。默认值“4”表示可使用大写字母、小写字母、数字、特殊符号，选择“5”表示可使用全部。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。默认值为“5”。
用户锁定时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。默认值为“5”。
密码有效期（天）	密码有效使用天数：取值范围0～90，0表示永久有效，默认值为“90”。
重复使用规则	修改密码时，不允许使用最近N次使用过的密码，N=1～5，默认为“1”。此策略只影响“人机”用户。
密码失效提前提醒天数	密码失效提前提醒天数：表示提醒密码失效到密码真正失效的天数。提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录FusionInsight Manager界面时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒，默认值为“5”。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔，取值范围为0～1440。“0”表示永远有效，“1440”表示1天。默认值为“5”。

步骤4 单击“确定”保存配置。修改密码策略之后，再修改用户密码，此时新的密码策略才会生效。

----结束

添加密码策略（MRS 3.1.2 及之后版本）

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 密码策略”。

步骤3 单击“新增密码策略”，根据界面提示，修改密码策略。

具体参数参见表11-44。

表 11-44 密码策略参数说明

参数名称	描述
密码策略名	由字母、数字、下划线（ _ ）、中划线（ - ）组成，且不能以中划线（ - ）开头，不区分大小写，长度3-32位。
最小密码长度	密码包含的最小字符个数，取值范围为8～32。
字符类型的数目	密码字符包含大写字母、小写字母、数字、空格和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=）的最小种类。可选择数值为“4”和“5”。默认值“4”表示可使用大写字母、小写字母、数字、特殊符号，选择“5”表示可使用全部。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。
用户锁定时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。
密码有效期（天）	密码有效使用天数：取值范围0～90，0表示永久有效。
重复使用规则	修改密码时，不允许使用最近N次使用过的密码，N=1～5，默认为“1”。 此策略只影响“人机”用户。
密码失效提前提醒天数	提醒密码失效到密码真正失效的天数。提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录FusionInsight Manager界面时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔，取值范围为0～1440。“0”表示永远有效，“1440”表示1天。

步骤4 单击“确定”保存配置。

系统新建用户默认使用default密码策略，新的密码策略创建成功后，用户在新创建用户时，可以手动选择该用户对应的密码策略。对于已经创建好的用户，也可以修改已有用户的密码策略，请参考**修改用户信息**。

----结束

说明

密码策略最多支持创建32条。

修改密码策略（MRS 3.1.2 及之后版本）

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 密码策略”。

步骤3 单击待修改的密码策略所在行的“修改”，进入“修改密码策略”页面，根据界面提示，修改密码策略。

具体参数参见**表11-44**。

步骤4 单击“确定”保存配置。

----结束

说明

- 用户不能修改自己所绑定的密码策略（**admin**用户除外）。
- 修改密码策略后，若该密码策略所绑定的用户密码剩余有效天数大于当前新密码策略有效期时，统一将这些用户的密码有效期设置为新密码策略的有效期；若用户密码剩余有效天数小于新密码策略有效期时，密码过期时间保持不变。

删除密码策略（MRS 3.1.2 及之后版本）

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 密码策略”。

步骤3 单击待删除的密码策略所在行的“删除”，在弹出窗口中单击“确定”。

----结束

说明

default密码策略和已经被用户绑定的密码策略不能被删除。

11.8.1.4.2 配置私有属性

操作场景

admin用户或绑定Manager_administrator角色的管理员用户，可以在FusionInsight Manager配置私有属性功能开关，用于支持用户（集群中所有业务用户）设置或取消自己的私有（Independent）属性。

开启私有属性开关后，需要业务用户登录后设置Independent属性，完成用户私有属性配置。

限制约束

- 管理员不能设置或取消业务用户的Independent属性。
- 管理员不能获取私有用户的认证凭据。

前提条件

已获取要求权限的管理员用户和密码。

操作步骤

配置私有属性开关

步骤1 以admin用户或绑定Manager_administrator角色的用户登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 安全策略 > 配置Independent”。

步骤3 打开或关闭Independent属性，根据提示输入密码，单击“确认”完成身份验证。

步骤4 身份验证通过后，等待修改OMS配置完成，单击“完成”结束操作。

说明

关闭Independent属性功能后：

- 已拥有这个属性的业务用户可以在右上角用户名下取消Independent属性，取消后无法重新设置。取消后已创建的私有表继续保持私有属性，取消后无法继续创建私有表。
- 没有这个属性的业务用户无法在右上角用户名下进行设置和取消操作。

配置用户私有属性

步骤5 以业务用户登录FusionInsight Manager。

须知

设置Independent属性后，管理员不能初始化私有用户（业务用户设置了Independent属性后，即为私有用户）的密码；如果忘记此用户密码，密码将无法找回。

admin用户无法设置Independent属性。

步骤6 移动鼠标到界面右上角的用户名。

步骤7 在弹出的菜单栏中单击“设置Independent”或“取消Independent”。

说明

- 私有属性功能开关已开启，业务用户当前已设置私有属性时，菜单栏显示“取消Independent”。
- 私有属性功能开关已开启，业务用户当前已取消私有属性时，菜单栏显示“设置Independent”。
- 私有属性功能开关已关闭，业务用户当前已设置私有属性时，菜单栏显示“取消Independent”。
- 私有属性功能开关已关闭，业务用户当前已取消私有属性时，菜单栏不显示。

步骤8 根据界面提示，输入密码，单击“确定”完成身份验证。

步骤9 身份验证通过后，在确认对话框中单击“确定”。

----结束

11.8.2 对接设置

11.8.2.1 配置 SNMP 北向参数

操作场景

如果用户需要在统一的运维网管平台查看集群的告警、监控数据，管理员可以在FusionInsight Manager使用SNMP服务将相关数据上报到网管平台。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 对接 > SNMP”。

步骤3 单击“SNMP服务”右侧的开关。

“SNMP服务”默认为不启用，开关显示为 表示启用。

步骤4 根据表11-45所示的说明填写对接参数。

表 11-45 对接参数

参数名称	参数说明
版本	SNMP协议版本号，取值范围： ● V2C：低版本，安全性较低。 ● V3：高版本，安全性更高。 推荐使用V3版本。
本地端口	本地端口，默认值“20000”，取值范围“1025”到“65535”。
读团体名	该参数仅在设置“版本”为v2c时可用，用于设置只读团体名。
写团体名	该参数仅在设置“版本”为v2c时可用，用于设置可写团体名。
安全用户名	该参数仅在设置“版本”为v3时可用，用于设置协议安全用户名。
认证协议	该参数仅在设置“版本”为v3时可用，用于设置认证协议，推荐选择SHA。
认证密码	该参数仅在设置“版本”为v3时可用，用于设置认证密钥。
确认认证密码	该参数仅在设置“版本”为v3时可用，用于确认认证密钥。
加密协议	该参数仅在设置“版本”为v3时可用，用于设置加密协议，推荐选择AES256。
加密密码	该参数仅在设置“版本”为v3时可用，用于设置加密密钥。
确认加密密码	该参数仅在设置“版本”为v3时可用，用于确认加密密钥。

说明

- “安全用户名”中禁止出现以64的公因子（1、2、4、8等）为单位长度的重复字符串，例如abab，abcdabcd。
- “认证密码”和“加密密码”密码长度为8到16位，至少需要包含大写字母、小写字母、数字、特殊字符中的3种类型字符。两个密码不能相同。两个密码不可和安全用户名相同或安全用户名的倒序字符相同。
- 使用SNMP协议从安全方面考虑，需要定期修改“认证密码”和“加密密码”密码。
- 使用SNMP v3版本时，安全用户在5分钟之内连续鉴权失败5次将被锁定，5分钟后自动解锁。

步骤5 单击“添加Trap目标”，在弹出的“添加Trap目标”对话框中填写以下参数：

- 目标标识：Trap目标标识，一般指接收Trap的网管或主机标识。长度限制1～255字节，一般由字母或数字组成。
- 目标IP模式：目标IP的IP地址模式，可选择“IPV4”或者“IPV6”。
- 目标IP：目标IP，要求可与管理节点的管理平面IP地址互通。
- 目标端口：接收Trap的端口，要求与对端保持一致，取值范围“0～65535”。
- Trap团体名：该参数仅在设置版本为V2C时可用，用于设置主动上报团体名。

单击“确定”，设置完成，退出“添加Trap目标”对话框。

步骤6 单击“确定”，设置完成。

----结束

11.8.2.2 配置 Syslog 北向参数

操作场景

如果用户需要在统一的告警平台查看集群的告警和事件，管理员可以在FusionInsight Manager使用Syslog协议将相关数据上报到告警平台。

须知

Syslog协议未做加密，传输数据容易被窃取，存在安全风险。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 对接 > Syslog”。

步骤3 单击“Syslog服务”右侧的开关。

“Syslog服务”默认为不启用，开关显示为 表示启用。

步骤4 根据表11-46所示的说明填写北向参数。

表 11-46 Syslog 对接参数

参数区域	参数名称	参数说明
Syslog协议	服务IP模式	设置对接服务器IP地址模式，可选择“IPV4”或者“IPV6”。
	服务IP	设置对接服务器IP地址。
	服务端口	设置对接端口。
	协议	设置协议类型，可选值： • TCP • UDP
	安全级别	设置上报消息的严重程度，取值范围： • Emergency • Alert • Critical • Error • Warning • Notice • Informational（默认值） • Debug **说明** “安全级别”和“Facility”共同组成发出消息的优先级（**Priority**）。 优先级（**Priority**）=“Facility” × 8 +“安全级别” “安全级别”和“Facility”各项对应的数值请参考**表11-47**。
	Facility	设置产生日志的模块。可选项参考**表11-47**，推荐使用默认值“local use 0 (local0)”。
	标识符	设置产品标识，默认为“FusionInsight Manager”。 标识符可以包含字母、数字、下划线、 空格、|、 $、{、 }、 点、中划线，并且不能超过256个字符。
报告信息	报文格式	设置告警报告的消息格式，具体要求请参考界面帮助。 报文格式可以包含字母、数字、下划线、 空格、|、 $、{、 }、 点、中划线，并且不能超过1024个字符。 **说明** 报文格式中信息域的说明请参考**表11-48**。
	报告信息类型	设置需要上报的告警类型。
	上报消息级别	设置需要上报的告警级别。

参数区域	参数名称	参数说明
未恢复告警上报设置	周期上报未恢复告警	设置是否按指定周期上报未清除的告警。打开开关表示启用此功能，关闭开关表示不启用。开关默认为关闭。
	间隔时间（分钟）	设置周期上报告警的时间间隔，当“周期上报未恢复告警”开关设置为打开时启用。单位为分钟，默认值为“15”，支持范围为“5”到“1440”（1天）。
心跳设置	上报心跳	设置是否开启周期上报Syslog心跳消息。打开开关表示开启此功能，关闭开关表示不启用。开关默认为关闭。
	心跳周期（分钟）	设置周期上报心跳的时间间隔，当“上报心跳”开关设置为打开时启用。单位为分钟，默认值为“15”，支持范围为“1”到“60”。
	心跳报文	设置心跳上报的内容，当“上报心跳”开关设置为打开时启用，不能为空。支持数字、字母、下划线、竖线、冒号、空格、英文逗号和句号字符，长度小于等于256。

说明

设置周期上报心跳报文后，在某些集群容错自动恢复的场景下（例如主备OMS倒换）可能会出现报文上报中断的现象，此时等待自动恢复即可。

步骤5 单击“确定”，设置完成。

----结束

参考信息

表 11-47 “安全级别”和“Facility”**字段数值编码**

安全级别	Facility	数值编码
Emergency	kernel messages	0
Alert	user-level messages	1
Critical	mail system	2
Error	system daemons	3
Warning	security/authorization messages (note 1)	4
Notice	messages generated internally by syslog	5
Informational	line printer subsystem	6
Debug	network news subsystem	7
-	UUCP subsystem	8

安全级别	Facility	数值编码
-	clock daemon (note 2)	9
-	security/authorization messages (note 1)	10
-	FTP daemon	11
-	NTP subsystem	12
-	log audit (note 1)	13
-	log alert (note 1)	14
-	clock daemon (note 2)	15
-	local use 0~7 (local0 ~ local7)	16~23

表 11-48 报文格式信息域表

信息域	描述
dn	集群名称
id	告警ID
name	告警名称
serialNo	告警序列号 **说明** 故障告警及其对应的恢复告警的告警序列号相同。
category	告警类型，取值范围： ● 0：故障告警 ● 1：恢复告警 ● 2：事件
occurTime	告警产生时间
clearTime	告警清除时间
isAutoClear	告警是否自动清除，取值范围： ● 1：是 ● 0：否
locationInfo	告警位置信息
clearType	告警清除类型，取值范围： ● -1：未清除 ● 0：自动清除 ● 2：手动清除

信息域	描述
level	告警级别，取值范围： • 1：紧急告警 • 2：重要告警 • 3：次要告警 • 4：提示告警
cause	告警原因
additionalInfo	附加信息
object	告警对象

11.8.2.3 配置监控指标数据转储

操作场景

监控数据上报功能可以将系统中采集到的监控数据写入到文本文件，并以FTP或SFTP的形式上传到指定的服务器中。

使用该功能前，管理员需要在FusionInsight Manager页面进行相关配置。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 对接 > 监控数据上传”。

步骤3 单击“监控数据上传”右边的开关。

“监控数据上传”默认为不启用，开关显示为 表示启用。

步骤4 根据表11-49所示的说明填写上传参数：

表 11-49 上传配置参数

参数名称	参数说明
FTP IP地址模式	必选参数，指定服务器IP地址模式，可选择“IPV4”或“IPV6”。
FTP IP地址	必选参数，指定监控指标数据对接后存放监控文件的FTP服务器。
FTP端口	必选参数，指定连接FTP服务器的端口。
FTP用户名	必选参数，指定登录FTP服务器的用户名。
FTP密码	必选参数，指定登录FTP服务器的密码。
保存路径	必选参数，指定监控文件在FTP服务器保存的路径。

参数名称	参数说明
转储时间间隔（秒）	必选参数，指定监控文件在FTP服务器保存的周期，单位为秒。
转储模式	必选参数，指定监控文件发送时使用的协议。可选协议为“SFTP”和“FTP”。建议使用基于SSH v2的SFTP模式，否则可能存在安全风险。
SFTP服务公钥	可选参数，指定FTP服务器的公共密钥，“转储模式”选择“SFTP”时此参数生效。

步骤5 单击“确定”，设置完成。

说明

选择转储模式为SFTP，当SFTP服务公钥为空时，先进行安全风险提示，确定安全风险后再保存配置。

----结束

数据格式

配置完成后，监控上报功能会将集群中监控数据周期性地写入到文本文件中，并根据用户配置的上报周期，将这些文件上报到对应的FTP/SFTP服务中。

- 监控文件产生规则
  - 按照指标的采集周期，监控指标会被分别写入到每30s，60s，以及300s产生的文件

    30s周期：默认采集周期为30s的实时指标。

    60s周期：默认采集周期为60s的实时指标。

    300s周期：非30s、60s采集的所有指标。
  - 文件名格式：*metirc_{周期}_{文件创建时间YYYYMMDDHHMMSS }.log*

    例如：metric_60_20160908085915.log

    metric_300_20160908085613.log
- 监控文件内容
  - 监控写入文件格式：

    “集群ID|集群名称|显示名称|服务名称|指标ID|采集时间|采集主机@m@子指标|单位|指标值”，其中：各字段间以“|”分隔，例如：

    ```
    1|xx1|Host|Host|10000413|2019/06/18 10:05:00|189-66-254-146|KB/s|309.910
    1|xx1|Host|Host|10000413|2019/06/18 10:05:00|189-66-254-152|KB/s|72.870
    2|xx2|Host|Host|10000413|2019/06/18 10:05:00|189-66-254-163|KB/s|100.650
    ```

    说明：实际的文件中不存在对应的文件格式标题。
  - 监控文件上传间隔：

    监控文件上传时间间隔可以在页面通过“转储时间间隔（秒）”配置，目前支持30s-300s之间均可。配置完成后，系统会按照指定的时间间隔，将文件定期上传到对应的FTP/SFTP服务器。
- 监控指标说明文件
  - 指标全集文件

指标全集文件all-shown-metric-zh_CN包括了所有指标的详细信息。第三方系统从上报的文件内容中解析出指标id后，可以通过查询指标全集文件获取指标详细信息。

指标全集文件位置：

主备OMS节点：{FusionInsight安装路径}/om-server/om/etc/om/all-shown-metric-zh_CN

指标全集文件内容参考：

```
实时指标ID,5分钟指标ID,指标名称,指标采集周期(秒),是否默认采集,指标所属服务,指标所属角色
00101,10000101,JobHistoryServer非堆内存使用量,30,false,Mapreduce,JobHistoryServer
00102,10000102,JobHistoryServer非堆内存分配量,30,false,Mapreduce,JobHistoryServer
00103,10000103,JobHistoryServer堆内存使用量,30,false,Mapreduce,JobHistoryServer
00104,10000104,JobHistoryServer堆内存分配量,30,false,Mapreduce,JobHistoryServer
00105,10000105,阻塞线程数,30,false,Mapreduce,JobHistoryServer
00106,10000106,运行线程数,30,false,Mapreduce,JobHistoryServer
00107,10000107,GC时间,30,false,Mapreduce,JobHistoryServer
00110,10000110,JobHistoryServer的CPU使用率,30,false,Mapreduce,JobHistoryServer
...
```

- 重要指标字段说明

  实时指标ID：指标的采集周期为30s/60s的指标ID，一个独立的指标项只可能存在30s或者60s的实时指标项。

  5分钟指标ID：指标对应的5分钟（300s）的指标ID。

  指标采集周期(秒)：主要是针对实时指标的采集周期，可选值为30或60。

  指标所属服务：指标所属的服务名名称，标明指标所属的服务类型，如HDFS、HBase等。

  指标所属角色：指标所属的角色名名称，标明指标所属的实际角色类型，如JobServer、RegionServer等。

- 解析说明

  针对采集周期为30s/60s的指标，参考该指标说明文件的是第1列，即**实时指标ID**即可找到对应的指标说明。

  针对采集周期为300s的指标，参考该指标说明文件对应的第2列，即**5分钟指标ID**即可找到对应的指标说明。

11.8.3 导入证书

操作场景

CA证书用于FusionInsight Manager各个模块、集群的组件客户端与服务端在通信过程中加密数据，实现安全通信。FusionInsight Manager支持快速导入CA证书，以确保产品安全使用。适用于以下场景：

- 首次安装好集群以后，需要更换企业证书。
- 企业证书有效时间已过期或安全性加强，需要更换为新的证书。

对系统的影响

- 更换证书过程中集群需要重启，此时系统无法访问且无法提供服务。
- 更换证书以后，所有组件和Manager的模块使用的证书将自动更新。
- 更换证书以后，还未信任该证书的本地环境，需要重新安装证书。

前提条件

- 证书文件和密钥文件可向企业证书管理员申请或由管理员生成。
- 获取需要导入到集群的CA证书文件（*.crt）、密钥文件（*.key）以及保存访问密钥文件密码的文件（password.property）。证书名称和密钥名称支持大小写字母和数字。以上文件在生成以后需要打包成tar格式压缩包。
- 准备一个访问密钥文件的密码用于访问密钥文件。

  密码复杂度要求如下，如果密码复杂度不满足如下要求，可能存在安全风险：
  - 密码字符长度最小为8位。
  - 至少需要包含大写字母、小写字母、数字、特殊字符~`!?,.;-_'(){}[]/<>@#$%^&*+|\=中的4种类型字符。
- 向证书管理员申请证书时，请提供访问密钥文件的密码并申请crt、cer、cert和pem格式证书文件，以及key和pem格式密钥文件。申请的证书需要有签发功能。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 证书”。

步骤2 在“上传证书”右侧单击 ···，在文件窗口中浏览已获取的证书文件tar压缩包并确认选择此文件。

步骤3 单击上传文件，Manager将上传压缩包并自动执行导入操作。

步骤4 导入完成后提示同步集群配置并重启WEB服务使新证书生效，单击“确定”。

步骤5 在弹出窗口输入当前登录用户密码验证身份，单击“确定”自动同步集群配置并重启WEB服务。

步骤6 重启完成后在浏览器地址栏中，输入并访问FusionInsight Manager的网络地址，验证能否正常打开页面。

步骤7 登录FusionInsight Manager。

步骤8 选择“集群 > *待操作集群的名称* > 概览 > 更多 > 重启”。

步骤9 输入当前登录的用户密码确认身份，单击“确定”。

----结束

11.8.4 OMS 管理

11.8.4.1 OMS 维护页面概述

总览

登录FusionInsight Manager以后，选择“系统 > OMS”后，打开OMS维护页面，管理员可以在此页面对OMS进行维护操作，包含查看基本信息、查看OMS业务模块的服务状态，也可以手工触发健康检查。

说明

OMS为操作维护系统的管理节点，OMS节点一般有两个，互为主备。

基本信息

FusionInsight Manager支持显示当前OMS的关联信息，包含如表11-50所示内容：

表 11-50 OMS 信息说明

项目	说明
版本	表示OMS版本，与FusionInsight Manager版本相同。
IP模式	表示当前集群网络的IP地址模式。
HA模式	表示OMS工作模式，由安装FusionInsight Manager时的配置文件指定。
当前主用	表示OMS主进程节点主机名，即主管理节点主机名。单击主机名可进入对应的主机详情页面。
当前备用	表示OMS备进程节点主机名，即备管理节点主机名。单击主机名可进入对应的主机详情页面。
持续时间	表示OMS进程启动持续的时间。

OMS 服务状态

FusionInsight Manager支持显示OMS所有业务模块的运行状态，每个业务模块的状态显示为●表示运行正常。

健康检查

管理员可以在OMS维护页面单击“健康检查”开始为OMS的状态进行检查。如果某些检查项存在问题，可直接打开检查说明进行处理。

进入/退出维护模式

配置OMS进入或退出维护模式。

系统参数

在大集群场景下对接DMPS集群。

11.8.4.2 修改 OMS 服务配置参数

操作场景

根据用户环境的安全要求，管理员可以在FusionInsight Manager修改OMS中Kerberos与LDAP配置。

对系统的影响

修改OMS的服务配置参数后，需要重启对应的OMS模块，此时FusionInsight Manager将无法正常使用。

操作步骤

修改okerberos配置

步骤1 登录FusionInsight Manager，选择“系统 > OMS”。

步骤2 在okerberos所在行，单击“修改配置”。

步骤3 根据表11-51所示的说明修改参数。

表 11-51 okerberos 参数配置一览表

参数名	说明
连接KDC最大时延（毫秒）	应用连接到Kerberos的超时时间，单位为毫秒，请填写整数值。
最大尝试次数	应用连接到Kerberos的最大重试次数，请填写整数值。
操作Ldap最大时延（毫秒）	Kerberos连接LDAP的超时时间，单位为毫秒。
搜索Ldap最大时延（毫秒）	Kerberos在LDAP查询用户信息的超时时间，单位为毫秒。
Kadmin监听端口	kadmin服务的端口。
KDC监听端口	kinit服务的端口。
Kpasswd监听端口	kpasswd服务的端口。
重置LDAP账户密码	Kerberos访问LDAP使用的机机账户（cn=krbadmin,ou=Users,dc=hadoop,dc=com和cn=krbkdc,ou=Users,dc=hadoop,dc=com）。 如果被勾选，账户的密码会使用随机生成的新密码进行重置。 **说明** 仅MRS 3.1.2及之后版本支持该参数。

步骤4 单击“确定”。

在弹出窗口输入当前登录用户密码验证身份，单击“确定”，在确认重启的对话框中单击“确定”。

修改oldap配置

步骤5 在oldap所在行，单击“修改配置”。

步骤6 根据表11-52所示的说明修改参数。

表 11-52 oldap 参数配置一览表

参数名	说明
Ldap服务监听端口	LDAP服务端口号。
重置LDAP账户密码	LDAP用于数据管理、同步和状态检测的机机账户（cn=root,dc=hadoop,dc=com和cn=pg_search_dn,ou=Users,dc=hadoop,dc=com）。 如果被勾选，账户的密码会使用随机生成的新密码进行重置。 **说明** 仅MRS 3.1.2及之后版本支持该参数。

步骤7 单击“确定”。

在弹出窗口输入当前登录用户密码验证身份，单击“确定”，在确认重启的对话框中单击“确定”。

说明

如果重置LDAP账户密码需要重启ACS，操作步骤如下：

1. 使用PuTTY，以**omm**用户登录主管理节点，执行以下命令更新域配置：
   sh ${BIGDATA_HOME}/om-server/om/sbin/restart-RealmConfig.sh
   提示以下信息表示命令执行成功：
   Modify realm successfully. Use the new password to log into FusionInsight again.
2. 执行**sh $CONTROLLER_HOME/sbin/acs_cmd.sh stop**，停止ACS。
3. 执行**sh $CONTROLLER_HOME/sbin/acs_cmd.sh start**，启动ACS。

重启集群

步骤8 登录FusionInsight Manager，参考**滚动重启集群**章节，重启集群。

----结束

11.8.5 部件管理

11.8.5.1 查看部件包

操作场景

完整的MRS集群由多个部件包组成，FusionInsight Manager单独安装某些服务前需要检查此服务对应的部件包是否已安装。

操作步骤

步骤1 登录FusionInsight Manager，选择“系统 > 部件”。

步骤2 在“已安装部件”查看所有部件列表。

说明

在“平台类型”列可查看部件已注册的OS及平台类型。

步骤3 单击部件名称左侧的，可查看部件包含的服务及其版本号。

----结束

11.9 集群管理

11.9.1 配置客户端

11.9.1.1 安装客户端

操作场景

该操作指导安装工程师安装MRS集群所有服务（不包含Flume）的客户端。MRS针对不同服务提供了Shell脚本，供开发维护人员在不同场景下登录其对应的服务维护客户端完成对应的维护任务。

说明

- 通过Manager界面修改服务端配置或系统升级后，建议重新安装客户端，否则客户端与服务端版本将不一致。

前提条件

- 安装目录可以不存在，会自动创建。但如果存在，则必须为空。目录路径不能包含空格。
- 客户端节点为集群外部服务器时，必须能够与集群业务平面网络互通，否则安装会失败。
- 客户端必须启用NTP服务，并保持与服务端时间一致，否则安装会失败。
- 对于下载所有组件客户端的情况，HDFS与Mapreduce是合一目录（“*客户端目录*/HDFS/”）。
- 安装和使用客户端可以使用任意用户进行操作，用户名和密码请从系统管理员处获取，本章节以“user_client”进行举例。要求“user_client”用户为服务器文件目录（如“/opt/Bigdata/hadoopclient”）和客户端安装目录（如“/opt/client”）的“owner”，两个目录的权限为“755”。
- 使用客户端需要已从系统管理员处获取“组件业务用户”（默认用户或新增用户）和“密码”。
- 使用**omm**和**root**以外的用户安装客户端时，若“/var/tmp/patch”目录已存在，需将此目录权限修改为“777”，将此目录内的日志权限修改为“666”。

操作步骤

步骤1 获取软件包。

登录FusionInsight Manager，在“集群”下拉列表中单击需要操作的集群名称。

选择“更多 > 下载客户端”，弹出“下载集群客户端”信息提示框。

说明

在只安装单个服务的客户端的场景中，选择“集群 > *待操作集群的名称* > 服务 > *服务名称* > 更多 > 下载客户端”，弹出“下载客户端”信息提示框。

步骤2 “选择客户端类型”中选择“完整客户端”。

“仅配置文件”下载的客户端配置文件，适用于应用开发任务中，完整客户端已下载并安装后，系统管理员通过Manager界面修改了服务端配置，开发人员需要更新客户端配置文件的场景。

平台类型包括x86_64和aarch64两种：

- x86_64：可以部署在x86平台的客户端软件包。
- aarch64：可以部署在TaiShan服务器的客户端软件包。

说明

集群支持下载x86_64和aarch64两种类型客户端，但是客户端类型必须和待安装节点的架构匹配，否则客户端会安装失败。

步骤3 是否在集群的节点中生成客户端文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。单击“确定”，等待下载完成后，使用**omm**用户或**root**用户将获取的软件包复制到将要安装客户端的服务器文件目录，例如“/opt/Bigdata/hadoopclient”。然后执行**步骤5**。

  说明

  当用户无法获取**root**用户权限，需要用**omm**用户操作。

- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，执行**步骤4**。

步骤4 上传软件包。

使用WinSCP工具，以准备安装客户端的用户（如“user_client”），将获取的软件包上传到将要安装客户端的服务器文件目录，例如“/opt/Bigdata/hadoopclient”。

客户端软件包名称格式为：“FusionInsight_Cluster_*<集群ID>*_Services_Client.tar”。

后续步骤及章节以FusionInsight_Cluster_1_Services_Client.tar进行举例。

说明

客户端所在主机可以是集群内节点，也可以是集群外节点。当该节点为集群外部服务器时，必须能够与集群网络互通，并启用NTP服务以保持与服务端时间一致。

例如可以为外部服务器配置与集群一样的NTP时钟源，配置之后可以执行**ntpq -np**命令检查时间是否同步。

- 如果显示结果的NTP时钟源IP地址前有“*”号，表示同步正常，如下：

```
remote refid st t when poll reach delay offset jitter
==============================================================================
*10.10.10.162 .LOCL. 1 u 1 16 377 0.270 -1.562 0.014
```

- 如果显示结果的NTP时钟源IP前无“*”号，且“refid”项内容为“.INIT.”，或者回显异常，表示同步不正常，请联系技术支持。

```
remote refid st t when poll reach delay offset jitter
==============================================================================
10.10.10.162 .INIT. 1 u 1 16 377 0.270 -1.562 0.014
```

也可以为外部服务器配置与集群一样的chrony时钟源，配置之后可以执行**chronyc sources**命令检查时间是否同步。

- 如果显示结果的主OMS节点chrony服务IP地址前有“*”号，表示同步正常，如下：

```
MS Name/IP address         Stratum Poll Reach LastRx Last sample
===============================================================================
^* 10.10.10.162              10  10   377   626    +16us[  +15us] +/-  308us
```

- 如果显示结果的主OMS节点NTP服务IP前无“*”号，且“**Reach**”项内容为“0”，表示同步不正常。

```
MS Name/IP address         Stratum Poll Reach LastRx Last sample
===============================================================================
^? 10.1.1.1                     0  10     0     -     +0ns[   +0ns] +/-    0ns
```

步骤5 以**user_client**用户登录将要安装客户端的服务器。

步骤6 解压软件包。

进入安装包所在目录，例如“/opt/Bigdata/hadoopclient”。执行如下命令解压安装包到本地目录。

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤7 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig.tar: OK
```

步骤8 解压获取的安装文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

步骤9 配置客户端网络连接。

1. 确保客户端所在主机能与解压目录下“hosts”文件（例如“/opt/Bigdata/hadoopclient/FusionInsight_Cluster_*<集群ID>*_Services_ClientConfig/hosts”）中所列出的各主机在网络上互通。
2. 当客户端所在主机不是集群中的节点时，需要在客户端所在节点的“/etc/hosts”文件（更改此文件需要**root**用户权限）中设置集群所有节点主机名和业务平面IP地址映射，主机名和IP地址请保持一一对应，可执行以下步骤在hosts文件中导入集群的域名映射关系。

a. 切换至**root**用户或者其他具有修改hosts文件权限的用户。

su - root

b. 进入客户端解压目录。

cd /opt/Bigdata/hadoopclient/FusionInsight_Cluster_1_Services_ClientConfig

c. 执行**cat realm.ini >> /etc/hosts**，将域名映射关系导入到hosts文件中。

说明

- 当客户端所在主机不是集群中的节点时，配置客户端网络连接，可避免执行客户端命令时出现错误。
- 如果采用yarn-client模式运行Spark任务，请在“*客户端安装目录*/Spark/spark/conf/spark-defaults.conf”文件中添加参数“spark.driver.host”，并将参数值设置为客户端的IP地址。
- 当采用yarn-client模式时，为了Spark WebUI能够正常显示，需要在Yarn的主备节点（即集群中的ResourceManager节点）的hosts文件中，配置客户端的IP地址及主机名对应关系。

步骤10 进入安装包所在目录，执行如下命令安装客户端到指定目录（绝对路径），例如安装到“**/opt/client**”目录。

cd /opt/Bigdata/hadoopclient/FusionInsight_Cluster_1_Services_ClientConfig

执行**./install.sh /opt/client**命令，等待客户端安装完成（以下只显示部分屏显结果）。

```
The component client is installed successfully
```

说明

- 如果已经安装的全部服务或某个服务的客户端使用了“/opt/hadoopclient”目录，再安装其他服务的客户端时，需要使用不同的目录。
- 卸载客户端请删除客户端安装目录。
- 如果要求安装后的客户端仅能被该安装用户（如“user_client”）使用，请在安装时加“-o”参数，即执行**./install.sh /opt/hadoopclient -o**命令安装客户端。
- 如果安装NTP服务器为chrony模式，请在安装时加“chrony”参数，即执行**./install.sh /opt/client -o chrony**命令安装客户端。
- 由于HBase使用的Ruby语法限制，如果安装的客户端中包含了HBase客户端，建议客户端安装目录路径只包含大写字母、小写字母、数字以及_-?.@+=字符。
- 客户端节点为集群外部服务器且此节点无法与主oms节点的业务平面IP互通时或者无法访问主节点的20029端口时，客户端可以正常安装成功，但无法注册到集群中，无法在界面上进行展示。

步骤11 检查客户端是否安装成功，请登录客户端。

1. 执行**cd /opt/client**命令进入客户端安装目录。
2. 执行**source bigdata_env**命令配置客户端环境变量。
3. 如果集群为安全模式，执行以下命令，设置**kinit**认证，输入客户端用户登录密码；普通模式集群无需执行用户认证。

   kinit admin

   ```
   Password for xxx@HADOOP.COM: #输入admin用户登录密码（与登录集群的用户密码一致）
   ```

4. 输入**klist**命令查询并确认权限内容。

   ```
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: xxx@HADOOP.COM

   Valid starting       Expires              Service principal
   04/09/2021 18:22:35  04/10/2021 18:22:29  krbtgt/HADOOP.COM@HADOOP.COM
   ```

说明

- 使用kinit认证时，票据默认会存放到“/tmp/krb5cc_*uid*”目录中。

  *uid*表示当前登录操作系统的用户id，例如root用户的uid为0，那么root用户登录系统后使用kinit认证的票据会默认存放在“/tmp/krb5cc_0”。

  若当前用户对于“/tmp”目录没有读写权限，则会将票据缓存路径修改为“客户端安装目录/tmp/krb5cc_uid”，例如客户端安装目录为“/opt/hadoopclient”，则kinit认证的票据会存放在“/opt/hadoopclient/tmp/krb5cc_uid”。
- 使用kinit认证时，如果使用相同的用户登录操作系统，则存在票据相互覆盖的风险。可使用**-c** *cache_name*参数指定票据缓存位置，或者通过设置KRB5CCNAME环境变量避免该问题。

步骤12 集群重装后，之前安装的客户端将不再可用，需要重新部署客户端。

1. 以**root**用户登录客户端所在节点。
2. 使用以下命令查看客户端所在目录（下例中“/opt/hadoopclient”为客户端所在目录）。

   ll /opt

   ```
   drwxr-x---. 6 root root       4096 Dec 11 19:00 hadoopclient
   drwxr-xr-x. 3 root root       4096 Dec  9 02:04 godi
   drwx------. 2 root root      16384 Nov  6 01:03 lost+found
   drwxr-xr-x. 2 root root       4096 Nov  7 09:49 rh
   ```
3. 使用**mv**命令移除所有客户端程序所在文件夹内的文件（例如移除“**/opt/client**”文件夹）。

   mv /opt/client */tmp/clientbackup*
4. 重新安装客户端。

----结束

11.9.1.2 使用客户端

操作场景

客户端安装后，用户可以通过客户端在运维场景或业务场景中使用shell命令，也可以在应用程序开发场景中使用客户端中的样例工程。

该任务指导用户在运维场景或业务场景中使用客户端。

前提条件

- 已安装客户端。

  例如安装目录为“/opt/client”。
- 各组件业务用户由系统管理员根据业务需要创建。

  “机机”用户需要下载keytab文件，“人机”用户第一次登录时需修改密码。

操作步骤

步骤1 以客户端安装用户登录安装客户端的节点。

步骤2 执行以下命令，切换到客户端安装目录。

cd /opt/client

步骤3 执行以下命令配置环境变量。

source bigdata_env

步骤4 如果集群为安全模式，执行以下命令进行用户认证。普通模式集群无需执行用户认证。

kinit *组件业务用户*

步骤5 根据实际业务需要，执行shell命令。

----结束

11.9.1.3 更新已安装客户端的配置

操作场景

集群提供了客户端，可以在连接服务端、查看任务结果或管理数据的场景中使用。用户如果在Manager修改了服务配置参数并重启了服务，已安装的客户端需要重新下载并安装，或者使用配置文件更新客户端。

前提条件

已安装客户端。

操作步骤

方法一：

步骤1 登录FusionInsight Manager，在“集群”下拉列表中单击需要操作的集群名称。

步骤2 选择“更多 > 下载客户端 > 仅配置文件”。

此时生成的压缩文件包含所有服务的配置文件。

步骤3 是否在集群的节点中生成配置文件?

- 是，勾选“仅保存到如下路径”，单击“确定”开始生成客户端文件，文件生成后默认保存在主管理节点“/tmp/FusionInsight-Client”。支持自定义其他目录且**omm**用户拥有目录的读、写与执行权限。然后执行**步骤4**。
- 否，单击“确定”指定本地的保存位置，开始下载完整客户端，等待下载完成，然后执行**步骤4**。

步骤4 使用WinSCP工具，以客户端安装用户将压缩文件保存到客户端安装的目录，例如“/opt/hadoopclient”。

步骤5 解压软件包。

例如下载的客户端文件为“FusionInsight_Cluster_1_Services_Client.tar”执行如下命令进入客户端所在目录，解压文件到本地目录。

cd /opt/hadoopclient

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

步骤6 校验软件包。

执行**sha256sum**命令校验解压得到的文件，检查回显信息与sha256文件里面的内容是否一致，例如：

sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar.sha256

```
FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar: OK
```

步骤7 解压获取配置文件。

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar

步骤8 在客户端安装目录下执行如下命令，使用配置文件更新客户端。

sh refreshConfig.sh *客户端安装目录 配置文件所在目录*

例如，执行以下命令：

sh refreshConfig.sh /opt/hadoopclient /opt/hadoopclient/FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles

界面显示以下信息表示配置更新成功：

```
Succeed to refresh components client config.
```

----结束

方法二：

步骤1 以**root**用户登录客户端安装节点。

步骤2 进入客户端安装的目录，例如“/opt/client”，执行以下命令更新配置文件：

cd /opt/client

sh autoRefreshConfig.sh

步骤3 按照提示输入FusionInsight Manager管理员用户名，密码以及FusionInsight Manager界面浮动IP。

步骤4 输入需要更新配置的组件名，组件名之间使用“,”分隔。如需更新所有组件配置，可直接单击回车键。

界面显示以下信息表示配置更新成功：

```
Succeed to refresh components client config.
```

----结束

11.9.2 集群互信管理

11.9.2.1 集群互信概述

功能介绍

默认情况下，安全模式下的大数据集群用户只能访问本集群中的资源，无法在其他安全模式集群中进行身份认证并访问资源。

特性描述

- **域**
  每个系统用户安全使用的范围定义为“域”，不同的Manager系统需要定义唯一的域名。跨Manager访问实际上就是用户跨域使用。

- **用户加密**

  配置跨Manager互信，当前Kerberos服务端仅支持并使用“aes256-cts-hmac-sha1-96:normal”和“aes128-cts-hmac-sha1-96:normal”加密类型加密跨域使用的用户，不支持修改。

- **用户认证**

  配置跨Manager集群互信后，两个系统中只要存在同名用户，且对端系统的同名用户拥有访问自身系统中某个资源的对应权限，则可以使用当前系统用户访问远程资源。

- **直接互信**

  系统在配置互信的两个集群分别保存对端系统的互信票据，通过互信票据访问对端系统。

11.9.2.2 修改 Manager 系统域名

操作场景

每个系统用户安全使用的范围定义为“域”，不同的系统需要定义唯一的域名。FusionInsight Manager的域名在安装过程中生成，如果需要修改为特定域名，系统管理员可通过FusionInsight Manager进行配置。

须知

- 修改系统域名为高危操作，在执行本章节操作前，请确认已参考**备份Manager数据**章节成功备份了OMS数据。

对系统的影响

- 修改Manager系统域名时，需要重启所有集群，集群在重启期间无法使用。
- 修改域名后，Kerberos管理员与OMS Kerberos管理员的密码将重新初始化，请使用默认密码并重新修改。组件运行用户的密码是系统随机生成的，如果用于身份认证，请参见**导出认证凭据文件**，重新下载keytab文件。
- 修改域名后，“admin”用户、组件运行用户和系统管理员在修改域名以前添加的“人机”用户，密码会重置为相同密码，请重新修改。重置后的密码由两部分组成：系统生成部分和用户设置部分，系统生成部分为Admin@123，用户设置部分规则参照**表11-54**中“密码后缀”参数的说明，默认值为Admin@123。例如：系统生成部分为Admin@123，用户设置部分为Test#$%@123，则此时重置后的密码为Admin@123Test#$%@123。
- 重置后的密码必需满足当前用户密码策略，使用**omm**用户登录主OMS节点后，执行如下工具脚本可以获取到修改域名后的“人机”用户密码。

  sh ${BIGDATA_HOME}/om-server/om/sbin/get_reset_pwd.sh *密码后缀 user_name*

  - *密码后缀*为用户设置的参数，如果不指定，则填充为默认值，默认值为“Admin@123”。
  - *user_name*为可选参数，默认取值为“admin”。

  例如：

  sh ${BIGDATA_HOME}/om-server/om/sbin/get_reset_pwd.sh Test#$%@123

```
To get the reset password after changing cluster domain name.
 pwd_min_len     : 8
 pwd_char_types  : 4
The password reset after changing cluster domain name is: "Admin@123Test#$%@123"
```

“pwd_min_len”和“pwd_char_types”分别表示当前用户密码策略“最小密码长度”和“密码字符类型数目”，“Admin@123Test#$%@123”为修改系统域名后的“人机”用户密码。

- 修改系统域名后，重置后的密码由系统生成部分和用户设置部分组成，且必需满足当前用户密码策略，长度不足时在Admin@123和用户设置部分中间，使用一个或多个@补全；字符种类为5时，在Admin@123后补充一个空格。

  当用户设置部分为Test@123，使用默认用户密码策略时，新密码为“Admin@123Test@123”，长度为17字符种类为4。需满足当前用户密码策略时，新密码处理如**表11-53**所示。

  表 11-53 满足不同密码策略时的新密码

最小密码长度	字符种类	对比用户密码策略结果	重置后的密码
8到17位	4	已满足用户密码策略	Admin@123Test@123
18位	4	需补充一个@	Admin@123@Test@123
19位	4	需补充两个@	Admin@123@@Test@123
8到18位	5	需补充一个空格	Admin@123 Test@123
19位	5	需补充一个空格和一个@	Admin@123 @Test@123
20位	5	需补充一个空格和两个@	Admin@123 @@Test@123

- 修改系统域名后，系统管理员在修改域名以前添加的“机机”用户，请重新下载keytab文件。
- 修改系统域名后，请重新下载并安装集群客户端。

前提条件

- 系统管理员已明确业务需求，并规划好不同系统的域名。

  域名只能包含大写字母、数字、圆点（.）及下划线（_），且只能以字母或数字开头。
- Manager内所有集群全部组件的运行状态均为“良好”。
- Manager内所有集群的ZooKeeper服务的“acl.compare.shortName”参数需确保为默认值“true”。否则请修改该参数为“true”后重启ZooKeeper服务。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 域和互信”。

步骤3 修改相关参数。

表 11-54 相关参数

参数名	描述
本端域	填写本系统规划好的域名。
密码后缀	“人机”用户密码重置后的用户设置密码部分，必填，默认值为Admin@123。 **说明** 该参数只有在修改了“本端域”参数后，才会生效。且需满足以下条件： • 密码字符长度为8到16位。 • 至少需要包含大写字母、小写字母、数字、特殊字符中的三种类型字符。支持的特殊字符为`~!@#$%^&*()-_=+|[{}];:',<.>/?和空格。

步骤4 单击“确定”，等待修改配置完成后再继续执行后续步骤，完成前请勿提前执行后续步骤。

步骤5 以**omm**用户登录主管理节点。

步骤6 执行以下命令，重启更新域配置。

sh ${BIGDATA_HOME}/om-server/om/sbin/restart-RealmConfig.sh

提示以下信息表示命令执行成功。

```
Modify realm successfully. Use the new password to log into FusionInsight again.
```

说明

重启后部分主机与服务可能无法访问并触发告警，执行“restart-RealmConfig.sh”后大约需要1分钟自动恢复。

步骤7 使用重置后的admin用户及密码（例如Admin@123Admin@123）登录FusionInsight Manager，单击主页上待操作集群名称后的 ，单击“重启”，重启集群。

在弹出窗口中输入当前登录的用户密码确认身份，然后单击“确定”。

在确认重启集群的对话框中单击“确定”，等待界面提示“操作成功。”，单击“完成”。

步骤8 退出FusionInsight Manager，重新登录正常表示配置已成功。

步骤9 使用**omm**用户登录主管理节点，执行以下命令刷新作业提交客户端配置：

sh /opt/executor/bin/refresh-client-config.sh

步骤10 如果有运行中的HetuEngine计算实例，需重启该计算实例。

1. 使用用于访问HetuEngine WebUI界面的用户登录FusionInsight Manager。

2. 选择“集群 > 服务 > HetuEngine”，进入HetuEngine服务页面。
3. 在“概览”页签下的“基本信息”区域，单击“HSConsole WebUI”后的链接，进入HSConsole界面。
4. 对于运行中的计算实例，在实例的“操作”列单击“停止”，待计算实例处于“已停止”状态后，再单击“启动”重启计算实例。

----结束

11.9.2.3 配置跨 Manager 集群互信

操作场景

当不同的两个Manager系统下安全模式的集群需要互相访问对方的资源时，系统管理员可以设置互信的系统，使外部系统的用户可以在本系统中使用。

每个系统用户安全使用的范围定义为“域”，不同的Manager系统需要定义唯一的域名。跨Manager访问实际上就是用户跨域使用。

说明

最多支持配置500个互信集群。

对系统的影响

- 配置跨集群互信后，外部系统的用户可以在本系统中使用，请系统管理员根据企业业务与安全要求，定期检视Manager系统中用户的权限。
- 配置跨集群互信时需要停止所有集群，会造成业务中断。
- 配置跨集群互信后，互信的集群中均会增加Kerberos内部用户“krbtgt/*本集群域名@外部集群域名*”、“krbtgt/*外部集群域名@本集群域名*”，用户不能删除。请系统管理员根据企业安全要求，及时且定期修改密码，需同时修改互信系统中4个用户且密码保持一致。具体请参见**修改组件运行用户密码**。修改密码期间可能影响跨系统业务应用的连接。
- 配置跨集群互信后，各个集群都需要重新下载并安装客户端。
- 配置跨集群互信后，验证配置后是否可以正常工作，且如何使用本系统用户访问对端系统资源，请参见**配置跨集群互信后的用户权限**。

前提条件

- 系统管理员已明确业务需求，并规划好不同系统的域名。域名只能包含大写字母、数字、圆点（.）及下划线（_），且只能以字母或数字开头。
- 配置跨集群互信前，两个Manager系统的域名必须不同。MRS创建ECS/BMS集群时会随机生成唯一系统域名，通常无需修改。
- 配置跨集群互信前，两个集群中不能存在有相同的主机名，也不能存在相同的IP地址。
- 配置互信的两个集群系统时间必须一致，且系统上的NTP服务必须使用同一个时间源。
- 配置互信的两个集群系统内所有集群全部组件的运行状态均为“良好”。
- Manager内所有集群的ZooKeeper服务的“acl.compare.shortName”参数需确保为默认值“true”。否则请修改该参数为“true”后重启ZooKeeper服务。

操作步骤

步骤1 登录其中一个FusionInsight Manager。

步骤2 在主页中停止所有集群。

单击主页上待操作集群名称后的 ，单击“停止”，输入集群管理员密码后在弹出的“停止集群”窗口中单击“确定”，等待集群停止成功。

步骤3 选择“系统 > 权限 > 域和互信”。

步骤4 修改配置参数“互信对端域”。

表 11-55 相关参数

参数名	描述
realm_name	填写对端系统的域名。
ip_port	填写对端系统的KDC地址。 参数值格式为：*对端系统内要配置互信集群的Kerberos服务部署的节点IP地址:端口*。 ● 如果是双平面组网，需填写业务平面IP地址。 ● 采用IPv6地址时，IP地址应写在中括号“[]”中。 ● 部署主备Kerberos服务或者对端系统内有多个集群需要与本端建立互信时，多个KDC地址使用逗号分隔。 ● 端口值可通过查看KrbServer服务的“kdc_ports”参数获取，默认值为“21732”。部署服务的节点IP可通过在KrbServer服务页面选择“实例”页签，查看KerberosServer角色的“业务IP”获取。 例如，Kerberos服务部署在10.0.0.1和10.0.0.2上，与本端系统建立互信，则对应参数值为“10.0.0.1:21732,10.0.0.2:21732”。

说明

如果需要配置与多个Manager系统的互信关系，请单击 添加新项目，并填写参数值。最多支持16个系统。删除多余的配置请单击 。

步骤5 单击“确定”。

步骤6 以**omm**用户登录主管理节点，执行以下命令更新域配置。

sh ${BIGDATA_HOME}/om-server/om/sbin/restart-RealmConfig.sh

提示以下信息表示命令执行成功。

```
Modify realm successfully. Use the new password to log into FusionInsight again.
```

重启后部分主机与服务可能无法访问并触发告警，执行“restart-RealmConfig.sh”后大约需要1分钟自动恢复。

步骤7 登录FusionInsight Manager，启动所有集群。

单击主页上待操作集群名称后的 ，单击“启动”，在“启动集群”窗口单击“确定”，等待集群启动成功。

步骤8 登录另外一个系统的FusionInsight Manager，重复以上操作。

----结束

11.9.2.4 配置跨集群互信后的用户权限

操作场景

配置完跨Manager集群互信后，需要在互信的系统上设置访问用户的权限，这样指定的用户才能在互信系统上进行对应的业务操作。

前提条件

两个系统已完成互信配置。

操作步骤

步骤1 登录本端系统的FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”，检查本次业务操作的用户是否在已存在：

- 是，执行**步骤3**。
- 否，执行**步骤4**。

步骤3 单击指定用户左侧的，检查该用户所在的用户组和角色分配的权限是否满足本次业务需求。若不满足，参见**权限设置**创建新角色并绑定用户，也可以直接修改用户的用户组或角色权限。

步骤4 参见**创建用户**，创建本次业务所需要的用户，同时关联业务所需要的用户组或者角色信息。

步骤5 登录互信系统的FusionInsight Manager，重复**步骤2**～**步骤4**，创建相同名字的用户并设置权限。

----结束

11.9.3 配置定时备份告警与审计信息

操作场景

用户通过修改配置文件，实现定时备份FusionInsight Manager的告警信息、Manager审计信息以及所有服务的审计信息到指定的存储位置。

备份支持使用SFTP协议或FTP协议，FTP协议未加密数据可能存在安全风险，建议使用SFTP。

操作步骤

步骤1 以**omm**用户登录主管理节点。

说明

用户只需在主管理节点执行此操作，不支持在备管理节点上配置定时备份。

步骤2 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/sbin

步骤3 执行以下命令，配置定时备份Manager告警、审计或者服务审计信息。

./setNorthBound.sh -t *信息类型* **-i** *远程服务器IP* **-p** *服务器使用的SFTP或FTP端口* **-u** *用户名* **-d** *保存信息的路径* **-c** *时间间隔（分钟）* **-m** *每个保存文件的信息记录数* **-s** *备份启停开关* **-e** *指定的协议*

例如：

./setNorthBound.sh -t alarm -i 10.0.0.10 -p 22 -u sftpuser -d /tmp/ -c 10 -m 100 -s true -e sftp

此脚本将修改告警信息备份配置文件“alarm_collect_upload.properties”。文件存储路径为“${BIGDATA_HOME}/om-server/tomcat/webapps/web/WEB-INF/classes/config”。

./setNorthBound.sh -t audit -i 10.0.0.10 -p 22 -u sftpuser -d /tmp/ -c 10 -m 100 -s true -e sftp

此脚本将修改审计信息备份配置文件“audit_collect_upload.properties”。文件存储路径为“${BIGDATA_HOME}/om-server/tomcat/webapps/web/WEB-INF/classes/config”。

./setNorthBound.sh -t service_audit -i 10.0.0.10 -p 22 -u sftpuser -d /tmp/ -c 10 -m 100 -s true -e sftp

此脚本将修改服务审计信息备份配置文件“service_audit_collect_upload.properties”。文件存储路径为“${BIGDATA_HOME}/om-server/tomcat/webapps/web/WEB-INF/classes/config”。

步骤4 根据界面提示输入用户的密码。密码将加密保存在配置文件中。

```
Please input sftp/ftp server password:
```

步骤5 显示如下结果，说明修改成功。备管理节点将自动同步配置文件。

```
execute command syncfile successfully.
Config Succeed.
```

----结束

11.9.4 修改 FusionInsight Manager 添加的路由表

操作场景

安装FusionInsight Manager时系统会自动在主管理节点上创建2条路由信息，执行**ip rule list**可以查看，如下示例：

```
0:from all lookup local
32764:from all to 10.10.100.100 lookup ntp_rt  #FusionInsight Manager创建的ntp路由信息（未配置外部NTP时钟源时无此信息）
32765:from 192.168.0.117 lookup om_rt  #FusionInsight Manager创建的om路由信息
32766:from all lookup main
32767:from all lookup default
```

说明

没有配置ntp外部服务器时只会有一条om路由信息“om_rt”。

如果FusionInsight Manager创建的路由信息与企业网络规划配置的路由信息发生冲突时，集群管理员可以使用“autoroute.sh”工具禁用或启用Manager创建的路由信息。

对系统的影响

禁用Manager创建的路由信息后，在设置新的路由信息之前，FusionInsight Manager页面无法登录，集群运行不受影响。

前提条件

已经成功安装Manager。

已获取待创建的WS浮动IP路由的相关信息。

禁用系统创建的路由信息

步骤1 以**omm**用户登录到主管理节点。执行以下命令，禁用系统创建的路由信息。

cd ${BIGDATA_HOME}/om-server/om/sbin

./autoroute.sh disable

```
Deactivating Route.
Route operation (disable) successful.
```

步骤2 执行以下命令，查看运行结果。如下例

ip rule list

```
0:from all lookup local
32766:from all lookup main
32767:from all lookup default
```

步骤3 执行以下命令，输入**root**用户密码，切换到**root**用户下。

su - root

步骤4 分别执行以下命令，手动创建新的WS浮动IP路由信息。

ip route add *WS浮动IP网段号/WS浮动IP子网掩码* **scope link src** *WS浮动IP* **dev** *WS浮动IP对应网卡* **table om_rt**

ip route add default via *WS浮动IP网关* **dev** *WS浮动IP对应网卡* **table om_rt**

ip rule add from *WS浮动IP* **table om_rt**

例如：

ip route add 192.168.0.0/255.255.255.0 scope link src 192.168.0.117 dev eth0:ws table om_rt

ip route add default via 192.168.0.254 dev eth0:ws table om_rt

ip rule add from 192.168.0.117 table om_rt

说明

当前网络的IP地址模式为IPv6时，应执行**ip -6 route add**命令。

步骤5 分别执行以下命令，手动创建新的ntp服务路由信息。未配置外部NTP时钟源时，跳过此步骤。

ip route add default via *NtpIP网关* **dev** *本机IP对应网卡* **table ntp_rt**

ip rule add to *ntpIP* **table ntp_rt**

*本机IP对应网卡*是指可与NTP服务器所在网段互通的网卡。

例如：

ip route add default via 10.10.100.254 dev eth0 table ntp_rt

ip rule add to 10.10.100.100 table ntp_rt

步骤6 执行以下命令，查看运行结果。

如下例，如产生路由表名为“om_rt”和“ntp_rt”的路由信息，则操作成功。

ip rule list

```
0:from all lookup local
32764:from all to 10.10.100.100 lookup ntp_rt  #未配置外部NTP时钟源时无此信息
32765:from 192.168.0.117 lookup om_rt
32766:from all lookup main
32767:from all lookup default
```

----结束

启用系统创建的路由信息

步骤1 以**omm**用户登录到主管理节点。

步骤2 执行以下命令，启用系统创建的路由信息。

cd ${BIGDATA_HOME}/om-server/om/sbin

./autoroute.sh enable

```
Activating Route.
Route operation (enable) successful.
```

步骤3 执行以下命令，查看运行结果。

如下例，如产生路由表名为“ntp_rt”和“om_rt”的两条路由信息，则操作成功。

ip rule list

```
0:from all lookup local
32764:from all to 10.10.100.100 lookup ntp_rt  #未配置外部NTP时钟源时无此信息
32765:from 192.168.0.117 lookup om_rt
32766:from all lookup main
32767:from all lookup default
```

----结束

11.9.5 切换维护模式

操作场景

FusionInsight Manager支持将集群、服务、主机或者OMS配置为维护模式，进入维护模式的对象将不再上报告警，避免在升级等维护变更期间系统产生大量无意义的告警，影响运维人员对集群状态的判断。

- 集群维护模式

  集群未正式上线或暂时离线进行运维操作时（例如非滚动方式的升级），可将整个集群配置为维护模式。
- 服务维护模式

  对特定服务进行维护操作时（例如对该服务的实例进行批量重启等可能影响业务的调试操作、对该服务相关的节点进行直接上下电或修复服务等），可仅将涉及的服务配置为维护模式。
- 主机维护模式

  对主机进行维护操作时（例如节点上下电、隔离主机、重装主机、升级操作系统、替换节点等），可仅将涉及的主机配置为维护模式。
- OMS维护模式

  对OMS节点进行重启、替换、修复等操作时，可将OMS配置为维护模式。

对系统影响

设置维护模式后，非维护操作引起的告警也将被抑制无法上报，直至退出维护模式后，仍然存在的故障才能上报告警，请谨慎操作。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 配置维护模式。

根据实际操作场景，确认需要配置维护模式的对象，参考**表11-56**进行操作。

表 11-56 切换维护模式

场景	步骤
配置集群进入维护模式	1. 在管理界面主页，选择待操作集群名称后的“ ··· > 进入维护模式”。 2. 在弹出的窗口中单击“确定”。 集群进入维护状态后，集群名称后的状态显示为 。维护操作完成后，单击“退出维护模式”，集群将退出维护模式。

场景	步骤
配置服务进入维护模式	1. 在管理界面选择“集群 > *待操作的集群名称* > 服务 > *服务名称*”。 2. 在服务详情页面选择“更多 > 进入维护模式”。 3. 在弹出的窗口中单击“确定”。 服务进入维护状态后，服务列表的对应服务名称后的状态显示为 。维护操作完成后，单击“退出维护模式”，服务将退出维护模式。 **说明** 配置某服务进入维护模式时，建议将依赖该服务的其他上层服务也都设置为维护模式。
配置主机进入维护模式	1. 在管理界面单击“主机”。 2. 在主机页面勾选待操作的主机，选择“更多 > 进入维护模式”。 3. 在弹出的窗口中单击“确定”。 主机进入维护状态后，主机列表的对应主机名称后的状态显示为 。维护操作完成后，单击“退出维护模式”，主机将退出维护模式。
配置OMS进入维护模式	1. 在管理界面选择“系统 > OMS > 进入维护模式”。 2. 在弹出的窗口中单击“确定”。 OMS进入维护状态后，OMS状态显示为 。维护操作完成后，单击“退出维护模式”，OMS将退出维护模式。

步骤3 查看集群维护视图。

在管理界面主页，选择待操作集群名称后的“ ··· > 维护模式视图”，在弹出的窗口中可查看当前集群内处于维护模式的服务及主机，方便查询。

维护操作完成后，可在维护模式视图中批量勾选服务与主机，然后单击“退出维护模式”，即可批量退出维护模式。

----结束

11.9.6 例行维护

为了保证系统长期正常、稳定的运行，管理员或维护工程师需要定期对**表11-57**所示的项目进行检查，并根据检查出的异常结果排除故障。建议检查人员根据企业管理规范，记录每个任务场景的结果并签名确认。

表 11-57 项目一览表

例行维护周期	任务场景	例行维护内容
每天	检查集群服务状态	• 检查每个服务的运行状态和配置状态是否正常，是否为绿色。 • 检查每个服务中，角色实例的运行状态和配置状态是否正常，是否为绿色。 • 检查每个服务中，角色实例的主备状态是否可以正常显示。 • 检查服务与角色实例的“概览”显示结果是否正常。
	检查集群主机状态	• 检查每个主机当前的运行状态是否正常，是否为绿色。 • 检查每个主机当前的磁盘使用率、内存使用率和CPU使用率。关注当前内存与CPU使用率是否处于上升趋势。
	检查集群告警信息	检查前一天是否生成了未处理异常告警，包含已自动恢复的告警。
	检查集群审计信息	检查前一天是否有“高危”和“危险”的操作，是否已确认操作的合法性。
	检查集群备份情况	检查前一天OMS、LDAP、DBService和NameNodeOMS、LDAP和DBServiceOMS、LDAP是否自动备份。
	检查健康检查结果	在FusionInsight Manager执行健康检查，下载健康检查报告确认当前集群是否存在异常状态。建议启用自动健康检查，并及时导出最新的集群健康检查结果，根据检查结果修复不健康项目。
	检查网络通讯	检查集群网络工作状态，节点之间的网络通讯是否存在延时。
	检查存储状态	检查集群总体数据存储量是否出现了突然的增长： • 磁盘使用量是否已接近阈值，排查原因，例如是否有业务留下的垃圾数据或冷数据需要清理。 • 业务是否有增长需求，磁盘分区是否需要扩容。
	检查日志情况	• 检查是否有失败、失去响应的MapReduce、Spark任务，查看HDFS中“/tmp/logs/${username}/logs/${application id}”日志文件并排除问题。 • 检查Yarn的任务日志，查看失败、失去响应的任务日志，并删除重复数据。 • 检查Storm的worker日志。 • 备份日志到存储服务器。
每周	用户管理	检查用户密码是否将要过期并通知修改。“机机用户”用户修改密码需要重新下载keytab文件。
	分析告警	导出指定周期内产生的告警并分析。

例行维护周期	任务场景	例行维护内容
	扫描磁盘	对磁盘健康状态进行检查，建议使用专门的磁盘检查工具。
	统计存储	分批次排查集群节点磁盘数据是否均匀存储，筛选出明显数据增加或不足的硬盘，并确认硬盘是否正常。
	记录变更	安排并记录对集群配置参数和文件实施的操作，为故障分析处理场景提供依据。
每月	分析日志	• 收集集群节点服务器的硬件日志，例如BMC系统日志，并进行分析。 • 收集集群节点服务器的操作系统日志，并进行分析。 • 收集集群日志，并进行分析。
	诊断网络	对集群的网络健康状态进行分析。
	管理硬件	检查设备运行的机房环境，安排清洁设备。

11.10 日志管理

11.10.1 关于日志

日志描述

MRS集群的日志保存路径为“/var/log/Bigdata”。日志分类见下表：

表 11-58 日志分类一览表

日志类型	日志描述
安装日志	安装日志记录了Manager、集群和服务安装的程序信息，可用于定位安装出错的问题。
运行日志	运行日志记录了集群各服务运行产生的运行轨迹信息及调试信息、状态变迁、未产生影响的潜在问题和直接的错误信息。
审计日志	审计日志中记录了用户活动信息和用户操作指令信息，可用于安全事件中定位问题原因及划分事故责任。

MRS日志目录清单见下表：

表 11-59 日志目录一览表

文件目录	日志内容
/var/log/Bigdata/audit	组件审计日志。
/var/log/Bigdata/controller	日志采集脚本日志。 controller进程日志。 controller监控日志。
/var/log/Bigdata/dbservice	DBService日志。
/var/log/Bigdata/flume	Flume日志。
/var/log/Bigdata/hbase	HBase日志。
/var/log/Bigdata/hdfs	HDFS日志。
/var/log/Bigdata/hive	Hive日志。
/var/log/Bigdata/httpd	httpd日志。
/var/log/Bigdata/hue	Hue日志。
/var/log/Bigdata/kerberos	Kerberos日志。
/var/log/Bigdata/ldapclient	LDAP客户端日志。
/var/log/Bigdata/ldapserver	LDAP服务端日志。
/var/log/Bigdata/loader	Loader日志。
/var/log/Bigdata/logman	logman脚本日志管理日志。
/var/log/Bigdata/mapreduce	MapReduce日志。
/var/log/Bigdata/nodeagent	NodeAgent日志。
/var/log/Bigdata/okerberos	OMS Kerberos日志。
/var/log/Bigdata/oldapserver	OMS LDAP日志。
/var/log/Bigdata/ metric_agent	MetricAgent运行日志。
/var/log/Bigdata/omm	oms：“omm”服务端的复杂事件处理日志、告警服务日志、HA日志、认证与授权管理日志和监控服务运行日志。 oma：“omm”代理端的安装运行日志。 core：“omm”代理端与“HA”进程失去响应的dump日志。
/var/log/Bigdata/spark2x	Spark2x日志。
/var/log/Bigdata/sudo	omm执行sudo命令产生的日志。
/var/log/Bigdata/timestamp	时间同步管理日志。
/var/log/Bigdata/tomcat	Tomcat日志。

文件目录	日志内容
/var/log/Bigdata/watchdog	Watchdog日志。
/var/log/Bigdata/yarn	Yarn日志。
/var/log/Bigdata/zookeeper	ZooKeeper日志。
/var/log/Bigdata/oozie	Oozie日志。
/var/log/Bigdata/kafka	Kafka日志。
/var/log/Bigdata/storm	Storm日志。
/var/log/Bigdata/iotdb	IoTDB日志。
/var/log/Bigdata/cdl	CDL日志。
/var/log/Bigdata/upgrade	升级OMS日志。
/var/log/Bigdata/update-service	升级服务日志。

说明

启用多实例功能后，如果系统管理员添加了多个HBase、Hive和Spark服务的实例，新增加服务实例的日志描述、日志级别和日志格式，与原服务日志相同。服务实例的日志将独立保存在名为“/var/log/Bigdata/servicenameN”的目录中，HBase和Hive服务实例的审计日志保存在名为“/var/log/Bigdata/audit/servicenameN”的目录中。以HBase1为例，对应日志分别保存在“/var/log/Bigdata/hbase1”和“/var/log/Bigdata/audit/hbase1”。

安装日志

表 11-60 安装信息一览表

安装日志	日志描述
安装配置日志	记录了安装前配置过程的信息。
安装Manager日志	记录了安装双机Manager操作的信息。
安装集群日志	记录了安装集群步骤的信息。

运行日志

运行日志记录的运行信息描述如表11-61所示。

表 11-61 运行信息一览表

运行日志	日志描述
服务安装前的准备日志	记录服务安装前的准备工作，如检测、配置和反馈操作的信息。
进程启动日志	记录进程启动过程中执行的命令信息。
进程启动异常日志	记录进程启动失败时产生异常的信息，如依赖服务错误、资源不足等
进程运行日志	记录进程运行轨迹信息及调试信息，如函数入口和出口打印、模块间接口消息等。
进程运行异常日志	记录导致进程运行时错误的错误信息，如输入对象为空、编解码失败等错误。
进程运行环境信息日志	记录进程运行环境的信息，如资源状态、环境变量等。
脚本日志	记录脚本执行的过程信息。
资源回收日志	记录资源回收的过程信息。
服务卸载时的清理日志	记录卸载服务时执行的步骤操作信息，如清除目录数据、执行时间等

审计日志

审计日志记录的审计信息包含Manager审计信息和组件审计信息。

表 11-62 Manager 审计信息一览表

操作类型	操作
用户管理	创建用户 修改用户 删除用户 创建组 修改组 删除组 添加角色 修改角色 删除角色 密码策略修改 修改密码 密码重置 用户登录 用户注销 屏幕解锁 下载认证凭据 用户越权操作 用户账号解锁 用户账号锁定 屏幕锁定 导出用户 导出用户组 导出角色

操作类型	操作
集群	启动集群 停止集群 重启集群 滚动重启集群 重启所有过期实例 保存配置 同步集群配置 定制集群监控指标 配置监控转储 保存监控阈值 下载客户端配置 北向Syslog接口配置 北向SNMP接口配置 SNMP清除告警 SNMP添加trap目标 SNMP删除trap目标 SNMP检查告警 SNMP同步告警 创建阈值模板 删除阈值模板 应用阈值模板 保存集群监控配置数据 导出配置数据 导入集群配置数据 导出安装模板 修改阈值模板 取消阈值模板应用 屏蔽告警 发送告警 修改OMS数据库密码 重置组件数据库密码 重启OMM和Controller 启动集群的健康检查 导入证书文件 配置SSO信息 删除健康检查历史报告 修改集群属性 同步维护命令

操作类型	操作
	异步维护命令 定制报表监控指标 导出报表监控数据 SNMP执行异步命令 重启WEB服务 定制静态资源池监控指标 导出静态资源池监控数据 定制主页监控指标 中止任务 还原配置 修改域和互信的配置 修改系统参数 集群进入维护模式 集群退出维护模式 OMS进入维护模式 OMS退出维护模式 批量退出维护模式 修改OMS配置 启用阈值告警 同步所有集群配置

操作类型	操作
服务	启动服务 停止服务 同步服务配置 刷新服务队列 定制服务监控指标 重启服务 滚动重启服务 导出服务监控数据 导入服务配置数据 启动服务的健康检查 服务配置 上传配置文件 下载配置文件 同步实例配置 实例入服 实例退服 启动实例 停止实例 定制实例监控指标 重启实例 滚动重启实例 导出实例监控数据 导入实例配置数据 创建实例组 修改实例组 删除实例组 移动到另一个实例组 服务进入维护模式 服务退出维护模式 修改服务显示名称 修改服务关联关系 下载监控数据 屏蔽告警 取消屏蔽告警 导出服务的报表数据 添加报表的自定义参数 修改报表的自定义参数 删除报表的自定义参数

操作类型	操作
	倒换控制节点 新增挂载表 修改挂载表
主机	设置节点机架 启动所有角色 停止所有角色 隔离主机 取消隔离主机 定制主机监控指标 导出主机监控数据 主机进入维护模式 主机退出维护模式 导出主机基本信息 导出主机分布的报表数据 导出主机趋势的报表数据 导出主机集群的报表数据 导出服务的报表数据 定制主机集群监控指标 定制主机趋势监控指标
告警	导出告警 清除告警 导出事件 批量清除告警
采集日志	采集日志文件 下载日志文件 采集服务堆栈信息 采集实例堆栈信息 准备服务堆栈信息 准备实例堆栈信息 清理服务堆栈信息 清理实例堆栈信息
审计日志	修改审计转储配置 导出审计日志

操作类型	操作
备份恢复	创建备份任务 执行备份任务 批量执行备份任务 停止备份任务 删除备份任务 修改备份任务 锁定备份任务 解锁备份任务 创建恢复任务 执行恢复任务 停止恢复任务 重试恢复任务 删除恢复任务
多租户	保存静态配置 添加租户 删除租户 关联租户服务 删除租户服务 配置资源 创建资源 删除资源 增加资源池 修改资源池 删除资源池 恢复租户数据 修改租户全局配置 修改容量调度器队列配置 修改超级调度器队列配置 修改容量调度器资源分布 清除容量调度器资源分布 修改超级调度器资源分布 清除超级调度器资源分布 添加资源目录 修改资源目录 删除资源目录 定制租户监控指标

操作类型	操作
健康检查	启动集群的健康检查 启动服务的健康检查 启动主机的健康检查 启动oms健康检查 启动系统的健康检查 更新健康检查的配置 导出健康检查报告 导出集群健康检查的结果 导出服务健康检查的结果 导出主机健康检查的结果 删除健康检查历史报告 导出健康检查历史报告 下载健康检查报告

表 11-63 组件审计信息一览表

审计日志	操作类型	操作
CDL审计日志	业务操作	创建Link 删除Link 创建Job 启动Job 删除Job
IoTDB审计日志	维护管理	授权 收回权限 认证和登录信息
	业务操作	删除时间序列/分区/函数/索引 修改时间序列操作
ClickHouse审计日志	维护管理	授权 收回权限 认证和登录信息
	业务操作	创建数据库/表 插入、删除、查询、执行数据迁移任务
DBService审计日志	维护管理	备份恢复操作

审计日志	操作类型	操作
HBase审计日志	DDL（数据定义）语句	创建表 删除表 修改表 增加列族 修改列族 删除列族 启用表 禁用表 用户信息修改 修改密码 用户登录
	DML（数据操作）语句	put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表）
	权限控制	给用户授权 取消用户授权
HDFS审计日志	权限管理	文件/文件夹访问权限 文件/文件夹owner信息
	文件操作	创建文件夹 创建文件 打开文件 追加文件内容 修改文件名称 删除文件/文件夹 设置文件时间属性 设置文件副本个数 多文件合并 文件系统检查 文件链接
Hive审计日志	元数据操作	元数据定义，如创建数据库、表等 元数据删除，如删除数据库、表等 元数据修改，如增加列、重命名表等 元数据导入/导出

审计日志	操作类型	操作
	数据维护	向表中加载数据 向表中插入数据
	权限管理	创建/删除角色 授予/回收角色 授予/回收权限
Hue审计日志	服务启动	启动Hue
	用户操作	用户登录 用户退出
	任务操作	创建任务 修改任务 删除任务 提交任务 保存任务 任务状态更新
KrbServer审计日志	维护管理	修改kerberos账号密码 添加kerberos账号 删除kerberos账号 用户认证
LdapServer审计日志	维护管理	添加操作系统用户 添加组 添加用户到组 删除用户 删除组
Loader审计日志	安全管理	用户登录
	元数据管理	查询connector 查询framework 查询step
	数据源连接管理	查询数据源连接 增加数据源连接 更新数据源连接 删除数据源连接 激活数据源连接 禁用数据源连接

审计日志	操作类型	操作
	作业管理	查询作业 创建作业 更新作业 删除作业 激活作业 禁用作业 查询作业所有执行记录 查询作业最近执行记录 提交作业 停止作业
Mapreduce审计日志	程序运行	启动Container请求 停止Container请求 Container结束，状态为成功 Container结束，状态为失败 Container结束，状态为中止 提交任务 结束任务
Oozie审计日志	任务管理	提交任务 启动任务 kill任务 暂停任务 恢复任务 重新运行任务
Spark2x审计日志	元数据操作	元数据定义，如创建数据库、表等 元数据删除，如删除数据库、表等 元数据修改，如增加列、重命名表等 元数据导入/导出
	数据维护	向表中加载数据 向表中插入数据
Storm审计日志	Nimbus	提交拓扑 中止拓扑 重分配拓扑 去激活拓扑 激活拓扑

审计日志	操作类型	操作
	UI	中止拓扑 重分配拓扑 去激活拓扑 激活拓扑
Yarn审计日志	任务提交	提交作业到队列相关的操作
Zookeeper审计日志	权限管理	设置ZNODE访问权限
	ZNODE操作	创建ZNODE 删除ZNODE 设置ZNODE数据

FusionInsight Manager的审计日志保存在数据库中，可通过“审计”页面查看及导出审计日志。

组件审计日志的文件信息见下表。部分组件审计日志文件保存在“/var/log/Bigdata/audit”，例如HDFS、HBase、Mapreduce、Hive、Hue、Yarn、Storm和ZooKeeper。每天凌晨3点自动将组件审计日志压缩备份到“/var/log/Bigdata/audit/bk”，最多保留最近的90个压缩备份文件，不支持修改备份时间。配置保留个数，请参见**配置审计日志本地备份数**。

其他组件审计日志文件保存在组件日志目录中。

表 11-64 组件审计日志目录

组件名称	审计日志目录
DBService	/var/log/Bigdata/audit/dbservice/dbservice_audit.log
HBase	/var/log/Bigdata/audit/hbase/hm/hbase-audit-hmaster.log /var/log/Bigdata/audit/hbase/hm/hbase-ranger-audit-hmaster.log /var/log/Bigdata/audit/hbase/rs/hbase-audit-regionserver.log /var/log/Bigdata/audit/hbase/rs/hbase-ranger-audit-regionserver.log /var/log/Bigdata/audit/hbase/rt/hbase-audit-restserver.log /var/log/Bigdata/audit/hbase/ts/hbase-audit-thriftserver.log

组件名称	审计日志目录
HDFS	/var/log/Bigdata/audit/hdfs/nn/hdfs-audit-namenode.log /var/log/Bigdata/audit/hdfs/nn/ranger-plugin-audit.log /var/log/Bigdata/audit/hdfs/dn/hdfs-audit-datanode.log /var/log/Bigdata/audit/hdfs/jn/hdfs-audit-journalnode.log /var/log/Bigdata/audit/hdfs/zkfc/hdfs-audit-zkfc.log /var/log/Bigdata/audit/hdfs/httpfs/hdfs-audit-httpfs.log /var/log/Bigdata/audit/hdfs/router/hdfs-audit-router.log
Hive	/var/log/Bigdata/audit/hive/hiveserver/hive-audit.log /var/log/Bigdata/audit/hive/hiveserver/hive-rangeraudit.log /var/log/Bigdata/audit/hive/metastore/metastore-audit.log /var/log/Bigdata/audit/hive/webhcat/webhcat-audit.log
Hue	/var/log/Bigdata/audit/hue/hue-audits.log
Kafka	/var/log/Bigdata/audit/kafka/audit.log
Loader	/var/log/Bigdata/loader/audit/default.audit
CDL	/var/log/Bigdata/audit/cdl/service/cdl-audit.log
Mapreduce	/var/log/Bigdata/audit/mapreduce/jobhistory/mapred-audit-jobhistory.log
Oozie	/var/log/Bigdata/audit/oozie/oozie-audit.log
Spark2x	/var/log/Bigdata/audit/spark2x/jdbcserver/jdbcserver-audit.log /var/log/Bigdata/audit/spark2x/jdbcserver/ranger-audit.log /var/log/Bigdata/audit/spark2x/jobhistory/jobhistory-audit.log
Storm	/var/log/Bigdata/audit/storm/logviewer/audit.log /var/log/Bigdata/audit/storm/nimbus/audit.log /var/log/Bigdata/audit/storm/supervisor/audit.log /var/log/Bigdata/audit/storm/ui/audit.log
Yarn	/var/log/Bigdata/audit/yarn/rm/yarn-audit-resourcemanager.log /var/log/Bigdata/audit/yarn/rm/ranger-plugin-audit.log /var/log/Bigdata/audit/yarn/nm/yarn-audit-nodemanager.log
ZooKeeper	/var/log/Bigdata/audit/zookeeper/quorumpeer/zk-audit-quorumpeer.log
IoTDB	/var/log/Bigdata/audit/iotdb/iotdbserver/log_audit.log

11.10.2 Manager 日志清单

日志描述

日志存储路径：Manager相关日志的默认存储路径为“/var/log/Bigdata/*Manager组件*”。

- ControllerService：/var/log/Bigdata/controller/（OMS安装、运行日志）
- Httpd：/var/log/Bigdata/httpd（httpd安装、运行日志）
- logman：/var/log/Bigdata/logman（日志打包工具日志）
- NodeAgent：/var/log/Bigdata/nodeagent（NodeAgent安装、运行日志）
- okerberos：/var/log/Bigdata/okerberos（okerberos安装、运行日志）
- oldapserver：/var/log/Bigdata/oldapserver（oldapserver安装、运行日志）
- MetricAgent：/var/log/Bigdata/metric_agent（MetricAgent运行日志）
- omm：/var/log/Bigdata/omm（omm安装、运行日志）
- timestamp：/var/log/Bigdata/timestamp（NodeAgent启动时间日志）
- tomcat：/var/log/Bigdata/tomcat（Web进程日志）
- watchdog：/var/log/Bigdata/watchdog（watchdog日志）
- upgrade：/var/log/Bigdata/upgrade（升级OMS日志）
- UpdateService：/var/log/Bigdata/update-service（升级服务日志）
- Sudo：/var/log/Bigdata/sudo（sudo脚本执行日志）
- OS：/var/log/*message文件*（OS系统日志）
- OS Performance：/var/log/osperf（OS性能统计日志）
- OS Statistics：/var/log/osinfo/statistics（OS参数配置信息日志）

日志归档规则：

Manager的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过10MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件。

表 11-65 Manager 日志列表

日志类型	日志文件名	描述
Controller运行日志	controller.log	记录组件安装、升级、配置、监控、告警和日常运维操作日志。
	controller_client.log	Rest接口运行日志。
	acs.log	Acs运行日志。
	acs_spnego.log	acs中spnego用户日志
	aos.log	Aos运行日志。
	plugin.log	Aos插件日志

日志类型	日志文件名	描述
	backupplugin.log	备份恢复进程运行日志
	controller_config.log	配置运行日志
	controller_nodesetup.log	Controller加载任务日志
	controller_root.log	Controller进程系统日志
	controller_trace.log	Controller与NodeAgent之间RPC通信日志
	controller_monitor.log	监控日志
	controller_fsm.log	状态机日志
	controller_alarm.log	Controller发送告警日志
	controller_backup.log	Controller备份恢复日志
	install.log，restore_package.log，installPack.log，distributeAdapterFiles.log，install_os_optimization.log	oms安装日志
	oms_ctl.log	oms启停日志
	preInstall_client.log	客户端安装前预处理日志
	installntp.log	ntp安装日志
	modify_manager_param.log	修改Manager参数日志
	backup.log	OMS备份脚本运行日志
	supressionAlarm.log	告警脚本运行日志
	om.log	生成om证书日志
	backupplugin_ctl.log	备份恢复插件进程启动日志
	getLogs.log	采集日志脚本运行日志
	backupAuditLogs.log	审计日志备份脚本运行日志
	certStatus.log	证书定期检查日志
	distribute.log	证书分发日志
	ficertgenetrate.log	证书替换日志，包括生成二级证书、cas证书、httpd证书的日志。

日志类型	日志文件名	描述
	genPwFile.log	生成证书密码文件日志
	modifyproxyconf.log	修改HTTPD代理配置的日志
	importTar.log	证书导入信任库日志
Httpd	install.log	Httpd安装日志
	access_log, error_log	Httpd运行日志
logman	logman.log	日志打包工具日志。
NodeAgent	install.log，install_os_optimization.log	NodeAgent安装日志
	installntp.log	ntp安装日志
	start_ntp.log	ntp启动日志
	ntpChecker.log	ntp检查日志
	ntpMonitor.log	ntp监控日志
	heartbeat_trace.log	NodeAgent与Controller心跳日志
	alarm.log	告警日志
	monitor.log	监控日志
	nodeagent_ctl.log，start-agent.log	NodeAgent启动日志
	agent.log	NodeAgent运行日志
	cert.log	证书日志
	agentplugin.log	监控agent侧插件运行日志
	omaplugin.log	OMA插件运行日志
	diskhealth.log	磁盘健康检查日志
	supressionAlarm.log	告警脚本运行日志
	updateHostFile.log	更新主机列表日志
	collectLog.log	节点日志采集脚本运行日志
	host_metric_collect.log	主机指标采集运行日志
	checkfileconfig.log	文件权限配置检查运行日志

日志类型	日志文件名	描述
	entropycheck.log	熵值检查运行日志
	timer.log	节点定时调度日志
	pluginmonitor.log	组件监控插件日志
	agent_alarm_py.log	NodeAgent检查文件权限发送告警日志
okerberos	addRealm.log，modifyKerberosRealm.log	切域日志
	checkservice_detail.log	Okerberos健康检查日志
	genKeytab.log	生成keytab日志
	KerberosAdmin_genConfigDetail.log	启动kadmin进程时，生成kadmin.conf的运行日志
	KerberosServer_genConfigDetail.log	启动krb5kdc进程时，生成krb5kdc.conf的运行日志
	oms-kadmind.log	kadmin进程的运行日志
	oms_kerberos_install.log，postinstall_detail.log	okerberos安装日志
	oms-krb5kdc.log	krbkdc运行日志
	start_detail.log	okerberos启动日志
	realmDataConfigProcess.log	切域失败，回滚日志
	stop_detail.log	okerberos停止日志
oldapserver	ldapserver_backup.log	Oldapserver备份日志
	ldapserver_chk_service.log	Oldapserver健康检查日志
	ldapserver_install.log	Oldapserver安装日志
	ldapserver_start.log	Oldapserver启动日志
	ldapserver_status.log	Oldapserver进程状态检查日志。
	ldapserver_stop.log	Oldapserver停止日志
	ldapserver_wrap.log	Oldapserver服务管理日志。
	ldapserver_uninstall.log	Oldapserver卸载日志

日志类型	日志文件名	描述
	restart_service.log	Oldapserver重启日志
	ldapserver_unlockUser.log	记录解锁Ldap用户和管理账户的日志
metric_agent	gc.log	MetricAgent JAVA虚拟机gc日志
	metric_agent.log	MetricAgent运行日志
	metric_agent_qps.log	MetricAgent内部队列长度及qps信息记录日志
	metric_agent_root.log	MetricAgent所有运行日志
	start.log	MetricAgent启停信息日志
omm	omsconfig.log	OMS配置日志
	check_oms_heartbeat.log	OMS心跳运行日志
	monitor.log	OMS监控日志
	ha_monitor.log	HA_Monitor操作日志
	ha.log	HA操作日志
	fms.log	告警日志
	fms_ha.log	告警的HA监控日志
	fms_script.log	告警控制日志
	config.log	告警配置日志
	iam.log	IAM日志
	iam_script.log	IAM控制日志
	iam_ha.log	IAM的HA监控日志
	config.log	IAM配置日志
	operatelog.log	IAM操作日志
	heartbeatcheck_ha.log	OMS心跳的HA监控日志
	install_oms.log	OMS安装日志
	pms_ha.log	监控的HA监控日志
	pms_script.log	监控控制日志
	config.log	监控配置日志
	plugin.log	监控插件运行日志

日志类型	日志文件名	描述
	pms.log	监控日志
	ha.log	HA运行日志
	cep_ha.log	CEP的HA监控日志
	cep_script.log	CEP控制日志
	cep.log	CEP日志
	config.log	CEP配置日志
	omm_gaussdba.log	gaussdb的HA监控日志
	gaussdb-<SERIAL>.log	gaussdb运行日志
	gs_ctl-<DATE>.log	gaussdb控制日志的归档日志
	gs_ctl-current.log	gaussdb控制日志
	gs_guc-current.log	gaussdb操作日志
	encrypt.log	omm加密日志
	omm_agent_ctl.log	OMA控制日志
	oma_monitor.log	OMA监控日志
	install_oma.log	OMA安装日志
	config_oma.log	OMA配置日志
	omm_agent.log	OMA运行日志
	acs.log	acs资源日志。
	aos.log	aos资源日志
	controller.log	controller资源日志
	floatip.log	floatip资源日志
	ha_ntp.log	ntp资源日志
	httpd.log	httpd资源日志
	okerberos.log	okerberos资源日志
	oldap.log	oldap资源日志
	tomcat.log	tomcat资源日志
	send_alarm.log	管理节点HA告警发送脚本运行日志
	feed_watchdog.log	feed_watchdog资源日志
timestamp	restart_stamp	NodeAgent启动时间

日志类型	日志文件名	描述
tomcat	cas.log，localhost_access_cas_log.log	cas运行日志
	catalina.log，catalina.out，host-manager.log，localhost.log，manager.log	tomcat运行日志
	localhost_access_web_log.log	记录访问FusionInsight Manager系统REST接口的日志
	web.log	web进程运行日志
	northbound_ftp_sftp.log，snmp.log	北向日志
	perfStats.log	性能数据统计日志
watchdog	watchdog.log，feed_watchdog.log	watchdog.log运行日志
update-service	omm_upd_server.log	updserver的运行日志
	omm_upd_agent.log	updagent的运行日志
	update-manager.log	updmanager的运行日志
	install.log	升级服务安装日志
	uninstall.log	升级服务卸载日志

日志类型	日志文件名	描述
	catalina.<时间>.log，catalina.out，host-manager.<时间>.log，localhost.<时间>.log，manager.<时间>.log，manager_access_log.<时间>.txt，web_service_access_log.<时间>.txt，catalina.log，gc-update-service.log.0.current，update-manager.controller，update-web-service.controller，update-web-service.log，commit_rm_distributed.log，commit_rm_upload_package.log，common_omagent_operator.log，forbid_monitor.log，initialize_package_atoms.log，initialize_unzip_pack.log，omm-upd.log，register_patch_pack.log，resume_monitor.logrollback_clear_patch.log，unregister_patch_pack.log，update-rcommupd.log，update-rcupdatemanager.log，update-service.log	升级服务运行日志
upgrade	upgrade.log_<时间>	升级OMS日志
	rollback.log_<时间>	回滚OMS日志
sudo	sudo.log	sudo脚本执行日志

日志级别

Manager中提供了如表11-66所示的日志级别。日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 11-66 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO记录系统及各事件正常运行状态信息
DEBUG	DEBUG记录系统及系统的调试信息。

日志格式

Manager的日志格式如下所示：

表 11-67 日志格式

日志类型	组件	格式	示例
Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2020-06-30 00:37:09,067 INFO [pool-1-thread-1] Completed Discovering Node. com.xxx.hadoop.om.controller.tasks.nodesetup.DiscoverNodeTask.execute(DiscoverNodeTask.java:299)

11.10.3 配置日志级别与文件大小

操作场景

如果需要在日志中调整记录的日志级别，则管理员可以修改FusionInsight Manager的日志级别。对于某个具体的服务，除了可以修改日志级别，还可以修改日志文件大小，防止磁盘空间不足日志无法保存。

对系统的影响

保存新的配置需要重启服务，此时对应的服务不可用。

修改 FusionInsight Manager 日志级别

1. 以**omm**用户登录主管理节点。
2. 执行以下命令，切换路径。

   cd ${BIGDATA_HOME}/om-server/om/sbin
3. 执行以下命令，修改日志级别。

   ./setLogLevel.sh *日志级别参数*

   日志级别参数如下，优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG，程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少：
   - “DEFAULT”：设置后恢复到默认日志级别。
   - “FATAL”：严重错误日志级别，设置后日志只会打印输出“FATAL”信息。
   - “ERROR”：错误日志级别，设置后日志打印输出“ERROR”和“FATAL”信息。
   - “WARN”：警告日志级别，设置后日志打印输出“WARN”、“ERROR”和“FATAL”信息。
   - “INFO”（默认）：提示信息日志级别，设置后日志打印输出“INFO”、“WARN”、“ERROR”和“FATAL”信息。
   - “DEBUG”：调试日志级别，设置后日志打印输出“DEBUG”、“INFO”、“WARN”、“ERROR”和“FATAL”信息。
   - “TRACE”：跟踪日志级别，设置后日志打印输出"TRACE"、“DEBUG”、“INFO”、“WARN”、“ERROR”和“FATAL”信息。

   说明

   由于开源中定义的不同，组件的日志级别定义略有差异。
4. 验证日志级别设置已生效，请下载日志并查看。请参见**日志**。

修改服务日志级别与日志文件大小

说明

KrbServer，LdapServer以及DBService不支持修改服务日志级别与日志文件大小。

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务”。

步骤3 单击服务列表中的某个服务，切换到“配置”页签。

步骤4 选择“全部配置”，左边菜单栏中展开角色实例的菜单，单击所需修改的角色所对应的“日志”。

步骤5 搜索各项参数，获取参数说明，在参数配置页面勾选所需的日志级别或修改日志文件大小。日志文件大小需填写单位“MB”。

须知

- 系统会根据配置的日志大小自动清理日志，如果需要保存更多的信息请设置一个较大的数值。为确保日志文件的完整性，建议根据实际业务量大小，在日志文件基于规则清理前，手动将日志文件备份存储至其他文件夹中。
- 个别服务不支持通过界面修改日志级别。

步骤6 单击“保存”，在“保存配置”单击“确定”。

步骤7 验证日志级别设置已生效，请下载日志并查看。

----结束

11.10.4 配置审计日志本地备份数

操作场景

集群组件的审计日志按名称分类，保存在集群各节点“/var/log/Bigdata/audit”，OMS每天凌晨3点自动备份这些审计日志目录。

各节点审计日志目录会按<节点IP>.tar.gz的文件名压缩，所有压缩文件再按<yyyy-MM-dd_HH-mm-ss>.tar.gz的文件名格式，压缩保存在主管理节点“/var/log/Bigdata/audit/bk/”，同时备管理节点会同步保存一个相同的副本。

默认情况下，OMS备份的文件最大保留个数为90，该任务指导系统管理员配置此最大保留个数。

操作步骤

步骤1 以**omm**用户登录主管理节点。

说明

用户只需在主管理节点执行此操作，不支持在备管理节点上修改审计日志备份文件数，否则可能造成集群无法正常工作。

步骤2 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/sbin

步骤3 执行以下命令，修改审计日志备份文件数。

./modifyLogConfig.sh -m *最大保留个数*

OMS备份组件审计日志默认最大保留90个，可选值为“0”到“365”，如果设置的保留个数越大，会占用更多的磁盘空间。

显示如下结果，说明修改成功：

```
Modify log config successfully
```

----结束

11.10.5 查看角色实例日志

操作场景

FusionInsight Manager支持在线直接查看各角色实例的日志内容，

操作步骤

步骤1 登录FusionInsight Manager。

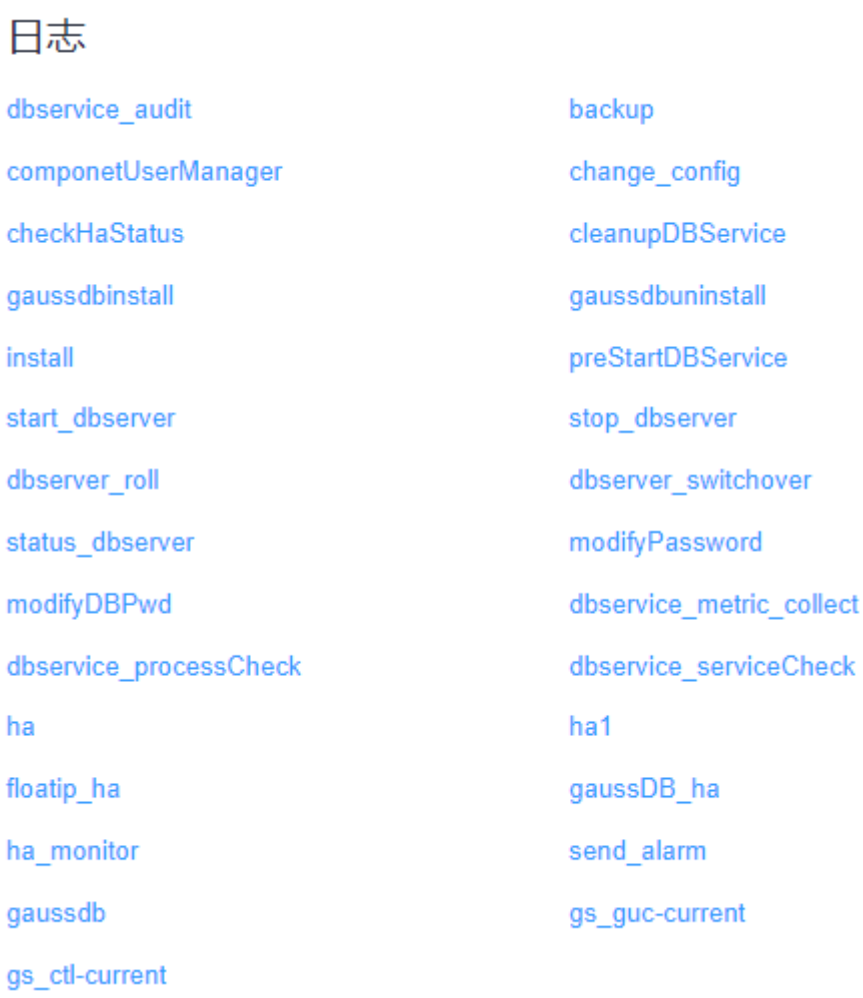
步骤2 选择“集群 > *待操作的集群名称* > 服务 > *服务名称* > 实例”，单击需要查看日志的实例名称，进入实例状态页面。

步骤3 在“日志”区域，单击要查看的日志文件名称，即可在线预览对应日志内容。

说明

- 在“主机”页面单击主机名称，在主机详情页面的“实例”区域，也可查看该主机上所有角色实例的日志文件。
- 日志内容默认最多显示100行，可单击“加载更多”按钮查看更多日志内容。单击“下载”按钮，可将该日志文件下载到本地。如需批量下载服务日志，请参考**下载日志**。

图 11-13 查看实例日志



----结束

11.11 备份恢复管理

11.11.1 备份恢复简介

概述

FusionInsight Manager提供对集群内的用户数据及系统数据的备份恢复能力，备份功能按组件提供。系统支持备份Manager的数据、组件元数据及业务数据。

备份功能支持将数据备份至本地磁盘（LocalDir）、本端HDFS（LocalHDFS）、远端HDFS（RemoteHDFS）、NAS（NFS/CIFS）、SFTP服务器（SFTP）、OBS，具体操作请参考**备份数据**。

对于支持多服务的组件，支持同服务多个实例的备份恢复功能且备份恢复操作与自身服务实例一致。

说明

MRS 3.1.0及之后版本才支持备份数据到OBS。

备份恢复任务的使用场景如下：

- 用于日常备份，确保系统及组件的数据安全。
- 当系统故障导致无法工作时，使用已备份的数据完成恢复操作。
- 当主集群完全故障，需要创建一个与主集群完全相同的镜像集群，可以使用已备份的数据完成恢复操作。

表 11-68 根据业务需要备份 Manager 配置数据

备份类型	备份内容	备份目录类型
OMS	默认备份集群管理系统中的数据库数据（不包含告警数据）以及配置数据。	• LocalDir • LocalHDFS • RemoteHDFS • NFS • CIFS • SFTP • OBS

表 11-69 根据业务需要备份组件元数据或其他数据

备份类型	备份内容	备份目录类型
DBService	备份DBService管理的组件（Loader、Hive、Spark、Oozie、Hue、CDL）的元数据。对于安装了多服务的集群，包含多个Hive和Spark服务实例的元数据。	• LocalDir • LocalHDFS • RemoteHDFS • NFS • CIFS • SFTP • OBS
Kafka	Kafka的元数据。	• LocalDir • LocalHDFS • RemoteHDFS • NFS • CIFS • OBS

备份类型	备份内容	备份目录类型
NameNode	备份HDFS元数据。添加多个NameService后，支持不同NameService的备份恢复功能且备份恢复操作与默认实例“hacluster”一致。	• LocalDir • RemoteHDFS • NFS • CIFS • SFTP • OBS
Yarn	备份Yarn服务资源池相关信息。	
HBase	HBase系统表的tableinfo文件和数据文件。	
IoTDB	IoTDB的元数据。	• LocalDir • NFS • RemoteHDFS • CIFS • SFTP
ClickHouse	ClickHouse的元数据。	• LocalDir • RemoteHDFS

表 11-70 根据业务需要备份特定组件业务数据

备份类型	备份内容	备份目录类型
HBase	备份表级别的用户数据。对于安装了多服务的集群，支持多个HBase服务实例的备份恢复功能且备份恢复操作与HBase服务实例一致。	• RemoteHDFS • NFS • CIFS • SFTP
HDFS	备份用户业务对应的目录或文件。 **说明** 加密目录不支持备份恢复。	
Hive	备份表级别的用户数据。对于安装了多服务的集群，支持多个Hive服务实例的备份恢复功能且备份恢复操作与Hive服务实例一致。	
IoTDB	备份IoTDB业务数据。	• RemoteHDFS
ClickHouse	备份表级别的用户数据。	• RemoteHDFS

需要特别说明的是，部分组件不提供单独的数据备份与恢复功能：

- Kafka支持副本特性，在创建主题时可指定多个副本来备份数据。
- CDL的数据存储在DBService与Kafka服务中，系统管理员可以通过创建备份DBService和Kafka的任务来备份数据。
- Mapreduce和Yarn的数据存放在HDFS上，故其依赖HDFS提供备份与恢复即可。

- ZooKeeper中存储的业务数据，其备份恢复能力由各上层组件按需独立实现。

原理

任务

在进行备份恢复之前，需要先创建备份恢复任务，并指定任务的参数，例如任务名称、备份数据源和备份文件保存的目录类型等等。通过执行备份恢复任务，用户可完成数据的备份恢复需求。在使用Manager执行恢复HDFS、HBase、Hive和NameNode数据时，无法访问集群。

每个备份任务可同时备份不同的数据源，每个数据源将生成独立的备份文件，每次备份的所有备份文件组成一个备份文件集，可用于恢复任务。备份任务支持将备份文件保存在Linux本地磁盘、本集群HDFS与备集群HDFS中。

备份任务提供全量备份或增量备份的策略，云数据备份任务不支持增量备份策略。如果备份的路径类型是NFS或CIFS，不建议使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

说明

任务运行规则：

- 某个任务已经处于执行状态，则当前任务无法重复执行，其他任务也无法启动。
- 周期任务自动执行时，距离该任务上次执行的时间间隔需要在120秒以上，否则任务推迟到下个周期启动。手动启动任务无时间间隔限制。
- 周期任务自动执行时，当前时间不得晚于任务开始时间120秒以上，否则任务推迟到下个周期启动。
- 周期任务锁定时无法自动执行，需要手动解锁。
- OMS、DBService、Kafka和NameNode备份任务开始执行前，若主管理节点“LocalBackup”分区可用空间小于20GB，则无法开始执行。

管理员在规划备份恢复任务时，请严格根据业务逻辑、数据存储结构、数据库或表关联关系，选择需要备份或者恢复的数据。系统默认创建间隔为1小时的周期备份任务“default-oms”、“default-*集群ID*”，支持全量备份OMS及集群的DBService、NameNode等元数据到本地磁盘。

快照

系统通过快照技术，快速备份数据。快照包含HBase快照、HDFS快照。

- HBase快照

  HBase快照是HBase表在特定时间的一个备份，该备份文件不复制业务数据，不影响RegionServer。HBase快照主要复制表的元数据，包含table descriptor，region info和HFile的引用信息。通过这些元数据信息可以恢复快照时间点之前的数据。

- HDFS快照

  HDFS快照是HDFS文件系统在特定时间点的只读备份副本，主要用于数据备份、用户误操作保护和灾难恢复的场景。

  任意HDFS目录均可以配置启用快照功能并创建对应的快照文件，为目录创建快照前系统会自动启用此目录的快照功能。创建快照不会对正常的HDFS操作有任何影响。每个HDFS目录最多可创建65536个快照。

  如果一个HDFS目录已创建快照，那么在快照完全删除以前，此目录无法删除或修改名称。该目录的上级目录或子目录也无法再创建快照。

DistCp

DistCp（distributed copy）是一个用于在本集群HDFS中或不同集群HDFS间进行大量数据复制的工具。在HBase、HDFS或Hive元数据的备份恢复任务中，如果选择将数据备份在备集群HDFS中，系统将调用DistCp完成操作。主备集群请选择安装相同版本的MRS软件版本并安装集群系统。

DistCp使用Mapreduce来影响数据的分布、异常处理及恢复和报告，此工具会把指定列表中包含的多个源文件和目录输入不同的Map任务，每个Map任务将复制列表中指定文件对应分区的数据。

使用DistCp在两个集群的HDFS间进行数据复制，集群双方需要分别配置互信（同一个FusionInsight Manager管理下的集群不需要配置互信）和启用集群间拷贝功能。集群数据备份到另一个集群的HDFS时，需要安装Yarn组件，否则备份失败。

本地快速恢复

使用DistCp将本集群HBase、HDFS和Hive数据备份在备集群HDFS中以后，本集群HDFS保留了备份数据的快照。用户可以通过创建本地快速恢复任务，直接从本集群HDFS的快照文件中恢复数据。

NAS

NAS（Network Attached Storage）是一种特殊的专用数据存储服务器，包括存储器件和内嵌系统软件，可提供跨平台文件共享功能。利用NFS（支持NFSv3、NFSv4）和CIFS（支持SMBv2、SMBv3）协议，用户可以连通MRS的业务平面与NAS服务器，将数据备份至NAS或从NAS恢复数据。

说明

- 数据备份至NAS前，系统会自动将NAS共享地址挂载为备份任务执行节点的本地分区。在备份结束后，系统会卸载备份任务执行节点的NAS共享分区。
- 为防止备份恢复失败，数据备份及恢复期间，请勿访问NAS服务器挂载至本地的共享地址，如：“/srv/BigData/LocalBackup/nas”。
- 业务数据备份至NAS时，会使用DistCp。

规格

表 11-71 备份恢复特性规格

项目	参数
备份或恢复任务最大数量（个）	100
同一集群同时运行的任务数量（个）	1
等待运行的任务最大数量（个）	199
Linux本地磁盘最大备份文件大小（GB）	600

说明

若业务数据存储在ZooKeeper中的上层组件，在备份恢复这类数据时，需确保单个备份或恢复任务的znode数量不会过大，否则会造成任务失败，并影响Zookeeper的服务性能。可通过如下方法确认单个备份或恢复任务的znode数量：

- 单个备份或恢复任务的znode数量要少于操作系统的文件句柄限制。查看句柄限制的方式如下：
  1. 使用shell命令输入：**cat /proc/sys/fs/file-max**，用于查看系统级的最大限制。
  2. 使用shell命令输入：**ulimit -n**，用于查看用户级的限制。
- 对于父目录的znode数量超过上述限制的情形，可以通过其子目录进行批量备份与恢复。使用Zookeeper提供的客户端脚本查看znode数量的方式：
  1. 在FusionInsight Manager首页，选择“集群 > *待操作集群的名称* > 服务 > Zookeeper > 实例”，查看Zookeeper各角色的管理IP。
  2. 登录客户端所在节点，执行如下命令：

     zkCli.sh -server *ip*:*port*，其中*ip*可以为任意管理IP，*port*默认值是2181。
  3. 当看到如下输出信息时，表示已经成功连接上Zookeeper服务器。
     ```
     WatchedEvent state:SyncConnected type:None path:null
     [zk: ip:port(CONNECIED) 0]
     ```
  4. 使用getusage命令查看待备份目录的znode数量，例如：

     getusage /hbase/region，输出结果中"Node count=xxxxxx"即表示region目录下存储的znode数量。

表 11-72 “default”任务规格

项目	OMS	HBase	Kafka	DBService	NameNode
备份周期	1小时				
最大备份数	168个（7天历史数据）				24个（1天历史数据）
单个备份文件最大大小	10MB	10 MB	512MB	100MB	20GB
最大占用磁盘大小	1.64GB	1.64 GB	84GB	16.41GB	480GB
备份数据保存位置	主备管理节点“*数据存放路径*/LocalBackup/”				

说明

- 默认任务保存的备份数据，请管理员根据企业运维要求，定期转移并保存到集群外部。
- 管理员可直接创建DistCp备份任务将OMS、DBService和NameNode等的数据保存到外部集群。
- 集群数据的备份任务运行时长可根据要备份的数据量除以集群与备份设备之间的网络带宽来计算得出，在实际场景中，建议将计算得出的时长乘以1.5作为任务执行时长参考值。
- 执行数据备份任务会对集群的最大IO性能产生影响，建议备份任务运行时间与集群业务高峰错开。

11.11.2 备份数据

11.11.2.1 备份 Manager 数据

操作场景

为了确保FusionInsight Manager系统日常数据安全，或者系统管理员需要对Manager进行重大操作（如扩容、减容等）前后，需要对Manager数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

管理员可以通过FusionInsight Manager创建备份Manager任务并备份数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 设置“备份对象”为“OMS”。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-73 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。

参数名称	描述
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“OMS”。

步骤7 在“OMS”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。

- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “备份时使用集群”：填写备份目录对应的集群名称。
  - “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

  - “源集群”：选择要备份数据使用的Yarn队列所在的集群。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和源集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为“*备份任务名_任务创建时间*”，用于保存数据源的备份文件。

备份文件的名称为“*版本号_数据源_任务执行时间*.tar.gz”。

----结束

11.11.2.2 备份 CDL 数据

操作场景

为了确保CDL日常数据安全，或者系统管理员需要对CDL进行重大操作（如升级或迁移等）时，需要对CDL数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

CDL的数据存储在DBService与Kafka服务中，系统管理员可以通过FusionInsight Manager创建备份DBService和Kafka任务并备份数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-74 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	• 首次全量备份，后续增量备份 • 每次都全量备份 • 每n次进行一次全量备份 **说明** • 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 • 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“DBService”和“Kafka”。

说明

若安装了多个DBService或Kafka服务，默认备份所有DBService或Kafka服务，可单击“指定服务”指定需要备份的服务。

步骤7 在“DBService”的“路径类型”，选择一个备份目录的类型。具体参数配置要求请参考步骤7。

步骤8 在“Kafka”的“路径类型”，选择一个备份目录的类型。具体参数配置要求请参考步骤7。

步骤9 单击“确定”保存。

步骤10 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。

备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

11.11.2.3 备份 ClickHouse 元数据

操作场景

为了确保ClickHouse集群中的元数据安全，或者集群用户需要对ZooKeeper进行重大操作（如升级或迁移等）时，需要对ClickHouse集群的元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

集群用户可以通过FusionInsight Manager创建备份ClickHouse任务并备份元数据。支持创建任务自动或手动备份数据。

须知

该功能仅MRS 3.1.0及之后版本支持。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果主备集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主备集群部署为普通模式，则不需要配置互信。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型，“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

创建周期备份任务，还需要填写以下参数：

- “开始时间”：表示任务第一次启动的时间。
- “周期”：表示任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
- “备份策略”：只支持“每次都全量备份”。

步骤6 在“备份配置”，勾选“元数据和其它数据”下的“ClickHouse”。

步骤7 在“ClickHouse”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，需要参考**创建ClickHouse备份任务路径选择RemoteHDFS时的环境配置**先进行环境配置。

  还需要配置以下参数：

- “目的端NameService名称”：填写备集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
- “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
- “最大备份数”：填写备份目录中可保留的备份文件集数量。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为“*备份任务名_任务创建时间*”，用于保存数据源的备份文件。备份文件的名称为“*数据源_任务执行时间.tar.gz*”。

----结束

11.11.2.4 备份 ClickHouse 业务数据

操作场景

为了确保ClickHouse日常用户的业务数据安全，或者集群用户需要对ClickHouse进行重大操作（如升级或迁移等），需要对ClickHouse数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

集群用户可以通过FusionInsight Manager创建备份ClickHouse任务。支持创建任务自动或手动备份数据。

须知

该功能仅MRS 3.1.0及之后版本支持。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主备集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主备集群部署为普通模式，则不需要配置互信。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-75 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	• 首次全量备份，后续增量备份 • 每次都全量备份 • 每n次进行一次全量备份 **说明** • 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。

步骤6 在“备份配置”，勾选“业务数据”下的“ClickHouse”。

步骤7 在“ClickHouse”的“路径类型”，选择一个备份目录的类型。

备份目录当前仅支持RemoteHDFS类型：

“RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

MRS 3.1.0和MRS 3.1.2版本选择此参数值，需要参考**创建ClickHouse备份任务路径选择RemoteHDFS时的环境配置**先进行环境配置。

MRS 3.1.3及之后版本还需要配置以下参数：

- “目的端NameService名称”：填写备集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “目的端主NameNode IP地址”：填写备集群主NameNode业务平面IP地址。
- “目的端备NameNode IP地址”：填写备集群备NameNode业务平面IP地址。
- “目的端NameNode RPC端口”：填写目的端集群HDFS基础配置“dfs.namenode.rpc.port”参数值。
- “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。

MRS 3.1.0和MRS 3.1.2版本还需要配置以下参数：

- “目的端NameService名称”：填写备集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
- “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
- “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在“备份内容”中，选择一个或多个需要备份的ClickHouse表。

支持两种方式选择备份数据：

- 添加备份数据文件

  单击导航中某个数据库的名称，将展开显示此数据库中的所有表，勾选指定的表。

  MRS 3.2.0及之后版本：

  a. 单击“添加”。
  b. 在“文件目录”框内选择要备份的表，单击“添加”添加到“备份文件”框内。
  c. 单击“确定”添加备份数据文件。

- MRS 3.2.0及之后版本：正则表达式筛选

  a. 单击“正则表达式输入”。
  b. 根据界面提示，在第一个输入框填入ClickHouse表所在的逻辑集群和数据库，需要与当前存在的逻辑集群和数据库完全匹配，例如”/default_cluster/database“。
  c. 在第二个框输入正则表达式，支持标准正则表达式。例如要筛选数据库中包含test关键字的所有的表，输入”test.*“。
  d. 单击“刷新”，在“目录名称”查看筛选的表。
  e. 单击“同步”保存筛选结果。

  说明

  - 输入正则表达式时，可以使用+和—增加或删除一条表达式。
  - 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。

- MRS 3.2.0之前版本：正则表达式筛选

  a. 单击“正则表达式输入”。
  b. 根据界面提示，在第一个输入框填写ClickHouse表所在的数据库，需要与当前存在的数据库完全匹配。例如“default”。

c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选数据库中所有的表，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的表，如**tb*1***可输入“tb\d*”。

d. 单击“刷新”，在“目录名称”查看筛选的表。

e. 单击“同步”保存筛选结果。

说明

- 输入正则表达式时，可以使用+和—增加或删除一条表达式。
- 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 待备份的目录或表不存在。
- NameService名称不正确。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为*“数据源_任务创建时间”*，数据源每次备份的最新备份文件保存在此目录中。

----结束

11.11.2.5 备份 DBService 数据

操作场景

为了确保DBService日常数据安全，或者系统管理员需要对DBService进行重大操作（如升级或迁移等）时，需要对DBService数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份DBService任务并备份数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。

- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-76 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“DBService”。

说明

若安装了多个DBService服务，默认备份所有DBService服务，可单击“指定服务”指定需要备份的DBService服务。

步骤7 在“DBService”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。

选择此参数值，还需要配置以下参数：

- “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
- “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和源集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）

  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。

备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

11.11.2.6 备份 HBase 元数据

操作场景

为了确保HBase元数据（主要包括tableinfo文件和HFile）安全，防止因HBase的系统表目录或者文件损坏导致HBase服务不可用，或者系统管理员需要对HBase系统表进行重大操作（如升级或迁移等）时，需要对HBase元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份HBase任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。

- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- HBase的“fs.defaultFS”配置参数需要与Yarn、HDFS的配置保持一致。
- 如果HBase数据存储在本地HDFS，支持将HBase元数据备份到OBS。如果HBase数据存储在OBS，则不支持数据备份。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-77 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	• 首次全量备份，后续增量备份 • 每次都全量备份 • 每n次进行一次全量备份 **说明** • 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 • 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“元数据和其它数据”下的“HBase”。

说明

若安装了多个HBase服务，默认备份所有HBase服务，可单击“指定服务”指定需要备份的HBase服务。

步骤7 在“HBase”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。

  默认保存目录为“*数据存放路径*/LocalBackup/”，例如“/srv/BigData/LocalBackup”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和源集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）

  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

11.11.2.7 备份 HBase 业务数据

操作场景

为了确保HBase日常数据安全，或者系统管理员需要对HBase进行重大操作（如升级或迁移等），需要对HBase业务数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份HBase任务并备份数据。支持创建任务自动或手动备份数据。

HBase备份业务数据时，可能存在以下场景：

- 用户创建HBase表时，“KEEP_DELETED_CELLS”属性默认值为“false”，备份该HBase表时会将已经删除的数据备份，可能导致恢复后出现垃圾数据。请根据业务需要，在创建HBase表时手动修改参数值为“true”。
- 用户在HBase表写入数据时手动指定了时间戳，且时间早于上一次该HBase表的备份时间，则在增量备份任务中可能无法备份新数据。

- HBase备份功能不支持对HBase的global或者命名空间的读取、写入、执行、创建和管理权限的访问控制列表（ACL）进行备份，恢复HBase数据后需要管理员在FusionInsight Manager上重新设置角色的权限。
- 已创建的HBase备份任务，如果本次备份任务在备集群的备份数据丢失，当下次执行增量备份时备份任务将失败，需要重新创建HBase的备份任务。若下次执行全量则备份正常。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用HDFS客户端，以**hdfs**用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建HDFS快照的目录清单，确保待备份的数据文件所在HDFS路径的父目录或子目录不存在HDFS快照，否则无法创建备份任务。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- HBase的“fs.defaultFS”配置参数需要与Yarn，HDFS的配置保持一致。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-78 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。

参数名称	描述
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“业务数据”下的“HBase > HBase”。

步骤7 在“HBase”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为20。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为100。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在“备份内容”中，选择一个或多个需要备份的HBase表。

支持两种方式选择备份数据：

- 添加备份数据文件

  单击导航中某个数据库的名称，将展开显示此数据库中的所有表，勾选指定的表。

  MRS 3.2.0及之后版本：

  a. 单击“添加”。
  b. 在“文件目录”框内选择要备份的表，单击“添加”添加到“备份文件”框内。
  c. 单击“确定”添加备份数据文件。

- 正则表达式筛选

  a. 单击“正则表达式输入”。
  b. 根据界面提示，在第一个输入框填写HBase表所在的命名空间，需要与当前存在的命名空间完全匹配。例如“default”。
  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选命名空间中所有的表，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的表，如**tb1**可输入“tb\d*”。
  d. 单击“刷新”，在“目录名称”查看筛选的表。
  e. 单击“同步”保存筛选结果。

  说明

  - 输入正则表达式时，可以使用+和—增加或删除一条表达式。
  - 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 队列名称不正确。
- 待备份的HBase表数据文件所在HDFS路径的父目录或子目录存在HDFS快照。
- 待备份的目录或表不存在。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为“*备份任务名_数据源_任务创建时间*”，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

11.11.2.8 备份 NameNode 数据

操作场景

为了确保NameNode日常数据安全，或者系统管理员需要对NameNode进行重大操作（如升级或迁移等），需要对NameNode数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份NameNode任务。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-79 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	仅支持“每次都全量备份”。 **说明** • 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 • 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“NameNode”。

步骤7 在“NameNode”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“*数据存放路径*/LocalBackup/”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：备集群NameNode的业务平面IP地址。
  - “目的端路径”：备份文件存放的位置。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）

- “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

    说明

    MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。

备份文件的名称为“*版本号_数据源_任务执行时间*.tar.gz”。

----结束

11.11.2.9 备份 HDFS 业务数据

操作场景

为了确保HDFS日常用户的业务数据安全，或者系统管理员需要对HDFS进行重大操作（如升级或迁移等），需要对HDFS数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份HDFS任务并备份数据。支持创建任务自动或手动备份数据。

说明

加密目录不支持备份恢复。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用HDFS客户端，以“hdfs”用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建HDFS快照的目录清单，确保待备份的数据文件所在HDFS路径的父目录或子目录不存在HDFS快照，否则无法创建备份任务。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-80 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。

参数名称	描述
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“HDFS”。

步骤7 在“HDFS”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在HDFS“备份内容”中，根据业务需要选择一个或多个需要备份的HDFS目录。

支持两种方式选择备份数据：

- 添加备份数据文件

  单击导航中某个数据库的名称，将展开显示此数据库中的所有表，勾选指定的表。

  MRS 3.2.0及之后版本：

  a. 单击“添加”。
  b. 在“文件目录”框内选择要备份的表，单击“添加”添加到“备份文件”框内。
  c. 单击“确定”添加备份数据文件。

- 正则表达式筛选

  a. 单击“正则表达式输入”。
  b. 根据界面提示，在第一个输入框填写目录的父目录完整路径，需要与当前存在的目录完全匹配。例如“/tmp”。
  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选父目录中所有的文件或子目录，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的文件，如**file1**可输入“file\d*”。
  d. 单击“刷新”，在“目录名称”查看筛选的目录。
  e. 单击“同步”保存筛选结果。

  说明

  - 输入正则表达式时，可以使用+和—增加或删除一条表达式。
  - 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。
  - 备份目录不可包含长期写入的文件，否则会导致备份任务失败，因此不建议对顶层目录进行操作，例如“/user”、“/tmp”、“/mr-history”。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 队列名称不正确。
- 待备份的数据文件所在HDFS路径的父目录或子目录存在HDFS快照。
- 待备份的目录或表不存在。
- NameService名称不正确。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为“*备份任务名_数据源_任务创建时间*”，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

11.11.2.10 备份 Hive 业务数据

操作场景

为了确保Hive日常用户的业务数据安全，或者系统管理员需要对Hive进行重大操作（如升级或迁移等），需要对Hive数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份Hive任务。支持创建任务自动或手动备份数据。

- Hive备份恢复功能不支持识别用户的Hive表、索引、视图等对象在业务和结构上存在的关联关系。用户在执行备份恢复任务时，需要根据业务场景管理统一的恢复点，防止影响业务正常运行。
- Hive备份恢复功能不支持Hive on RDB数据表，需要在外部数据库中单独备份恢复原始数据表。
- 已创建的Hive备份任务且包含Hive on HBase表，如果本次备份任务在备集群的备份数据丢失，当下次执行增量备份时备份任务将失败，需要重新创建Hive的备份任务。若下次执行全量则备份正常。
- 通过Manager备份功能对Hive表层级的HDFS目录做备份后，Hive表将无法被删除重建。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用HDFS客户端，以“hdfs”用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建HDFS快照的目录清单，确保待备份的数据文件所在HDFS路径的父目录或子目录不存在HDFS快照，否则无法创建备份任务。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-81 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“Hive > Hive”。

步骤7 在“Hive”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

步骤8 在“最大恢复点个数”填写备份任务在本集群中备份可保留的快照数量。

步骤9 在“备份内容”中，选择一个或多个需要备份的Hive表。

支持两种方式选择备份数据：

- 添加备份数据文件

  单击导航中某个数据库的名称，将展开显示此数据库中的所有表，勾选指定的表。

  MRS 3.2.0及之后版本：

  a. 单击“添加”。
  b. 在“文件目录”框内选择要备份的表，单击“添加”添加到“备份文件”框内。
  c. 单击“确定”添加备份数据文件。

- 正则表达式筛选

  a. 单击“正则表达式输入”。
  b. 根据界面提示，在第一个输入框填写Hive表所在的数据库，需要与当前存在的数据库完全匹配。例如“default”。
  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选数据库中所有的表，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的表，如**tb1**可输入“tb\d*”。
  d. 单击“刷新”，在“目录名称”查看筛选的表。
  e. 单击“同步”保存筛选结果。

说明

- 输入正则表达式时，可以使用＋和－增加或删除一条表达式。
- 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。

步骤10 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 队列名称不正确。
- 待备份的数据文件所在HDFS路径的父目录或子目录存在HDFS快照。
- 待备份的目录或表不存在。
- NameService名称不正确。

步骤11 单击“确定”保存。

步骤12 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为*“备份任务名_数据源_任务创建时间”*，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

11.11.2.11 备份 IoTDB 元数据

操作场景

为了确保IoTDB元数据安全，防止因IoTDB的元数据文件损坏等导致IoTDB服务不可用时，需要对IoTDB元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份IoTDB任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。
- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的Yarn队列等策略规格。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-82 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	选择数据周期备份策略。 ● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** 备份组件元数据时不支持增量备份，仅支持“每次都全量备份”。

步骤6 在“备份配置”，勾选“元数据和其它数据”区域下的“IoTDB”。

说明

若安装了多个IoTDB服务，默认备份所有IoTDB服务，可单击“指定服务”指定需要备份的服务。

步骤7 在“IoTDB”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“*数据存放路径*/LocalBackup/”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端主NameNode IP地址”：填写备集群主NameNode业务平面IP地址。
  - “目的端备NameNode IP地址”：填写备集群备NameNode业务平面IP地址。
  - “目的端NameNode RPC端口”：填写备集群HDFS基础配置“dfs.namenode.rpc.port”参数值。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “SFTP”：表示将备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “服务器共享路径”：SFTP服务器上的备份路径。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

步骤8 单击“确定”保存备份任务。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。

备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

11.11.2.12 备份 IoTDB 业务数据

操作场景

为了确保IoTDB日常数据安全，或者系统管理员需要对IoTDB进行重大操作（如升级或迁移等），需要对IoTDB业务数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份IoTDB任务并备份数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。目前IoTDB仅支持备份到HDFS中。
- 若IoTDB为普通模式，则不支持备份业务数据到安全模式HDFS。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查备集群HDFS是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-83 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。

参数名称	描述
备份策略	选择数据周期备份策略。 ● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份

步骤6 在“备份配置”，勾选“业务数据”区域下的“IoTDB > IoTDB”。

步骤7 在“IoTDB”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

“RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

选择此参数值，还需要配置以下参数：

- “目的端NameService名称”：填写备集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “目的端主NameNode IP地址”：填写备集群主NameNode业务平面IP地址。
- “目的端备NameNode IP地址”：填写备集群备NameNode业务平面IP地址。
- “目的端NameNode RPC端口”：填写备集群HDFS基础配置“dfs.namenode.rpc.port”参数值。
- “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
- “最大备份数”：填写备份目录中可保留的备份文件集数量。

步骤8 在“备份内容”中，选择一个或多个需要备份的业务数据。

支持两种方式选择备份数据：

- 添加备份数据文件

  单击导航中某个数据库的名称，将展开显示此数据库中的所有表，勾选指定的表。

  MRS 3.2.0及之后版本：

  a. 单击“添加”。

  b. 在“文件目录”框内选择要备份的表，单击“添加”添加到“备份文件”框内。

  c. 单击“确定”添加备份数据文件。

- 正则表达式筛选

  a. 单击“正则表达式输入”。

  b. 根据界面提示，在第一个输入框填写目录的父目录完整路径，需要与当前存在的目录完全匹配。例如“/root”。

  c. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选父目录中所有的文件或子目录，输入“([\s\S]*?)”。例如要筛选命名规则为字母数字组合的文件，如**file1**可输入“file\d*”。

d. 在第二个输入框输入正则表达式，支持标准正则表达式。例如要筛选包含“test”的对象，请输入“.*test.*”。例如要筛选以“test”开头的对象，请输入“test.*”。例如要输入以“test”结尾的对象，请输入“.*test”。

e. 单击“刷新”，在“目录名称”查看筛选的目录。

f. 单击“同步”保存筛选结果。

说明

- 输入正则表达式时，可以使用+和—增加或删除一条表达式。
- 如果已选择的表或目录不正确，可以单击“清除选中节点”清除勾选。
- 备份目录不可包含长期写入的文件，否则会导致备份任务失败，因此不建议对顶层目录进行操作，例如“/user”、“/tmp”、“/mr-history”。

步骤9 单击“校验”查看备份任务的配置是否正确。

校验失败可能存在以下原因：

- 目的端NameNode IP地址不正确。
- 待备份的数据不存在。

步骤10 单击“确定”保存。

步骤11 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备集群的备份路径中为每个备份任务创建子目录，目录名为*“备份任务名_数据源_任务创建时间”*，数据源每次备份的最新备份文件保存在此目录中。所有备份文件集保存在对应的快照目录中。

----结束

11.11.2.13 备份 Kafka 元数据

操作场景

为了确保Kafka元数据安全，或者系统管理员需要对ZooKeeper进行重大操作（如升级或迁移等）时，需要对Kafka元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建备份Kafka任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 如果数据要备份至远端HDFS中，需要准备一个用于备份数据的备集群，认证模式需要与主集群相同。其他备份方式不需要准备备集群。
- 如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“*数据存放路径*/LocalBackup/”是否有充足的空间。

- 如果数据要备份至NAS中，需要提前部署好NAS服务端。
- 如果数据要备份至OBS中，需要当前集群已对接OBS，并具有访问OBS的权限。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 单击“创建”。

步骤3 在“任务名称”填写备份任务的名称。

步骤4 在“备份对象”选择待操作的集群。

步骤5 在“备份类型”选择备份任务的运行类型。

“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

表 11-84 周期备份参数

参数名称	描述
开始时间	任务第一次启动的时间。
周期	任务下次启动，与上一次运行的时间间隔，支持按“小时”或按“天”。
备份策略	● 首次全量备份，后续增量备份 ● 每次都全量备份 ● 每n次进行一次全量备份 **说明** ● 备份Manager数据和组件元数据时不支持增量备份，仅支持“每次都全量备份”。 ● 如果“路径类型”要使用NFS或CIFS，不能使用增量备份功能。因为在NFS或CIFS备份时使用增量备份时，每次增量备份都会刷新最近一次全量备份的备份数据，所以不会产生新的恢复点。

步骤6 在“备份配置”，勾选“Kafka”。

步骤7 在“Kafka”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“*数据存放路径*/LocalBackup/”。

  选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “目的端NameService名称”：填写备集群的NameService名称。可以输入集群内置的远端集群的NameService名称（haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4），也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “目的端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “目的端路径”：填写备集群保存备份数据的HDFS目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录，例如“/hbase”或“/user/hbase/backup”。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “服务器共享路径”：填写用户配置的NAS服务器共享目录。（服务器共享路径不能设置为根目录，且共享路径的用户组和属组必须为nobody:nobody）
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “目的端路径”：填写保存备份数据的OBS目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。

说明

MRS 3.1.0及之后版本才支持备份数据到OBS。

步骤8 单击“确定”保存。

步骤9 在备份任务列表中已创建任务的“操作”列，选择“更多 > 即时备份”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为*“备份任务名_任务创建时间”*，用于保存数据源的备份文件。备份文件的名称为*版本号_数据源_任务执行时间.tar.gz*。

----结束

11.11.3 恢复数据

11.11.3.1 恢复 Manager 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对FusionInsight Manager系统进行重大数据调整等操作后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，需要对Manager进行恢复数据操作。

管理员可以通过FusionInsight Manager创建恢复Manager任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的Manager数据。

对系统的影响

- 恢复过程中需要重启Controller，重启时FusionInsight Manager无法登录和操作。
- 恢复过程中需要重启所有集群，集群重启时无法访问。
- Manager数据恢复后，会丢失从备份时刻到恢复时刻之间的数据，例如系统设置、用户信息、告警信息或审计信息。可能导致无法查询到数据，或者某个用户无法访问集群。
- Manager数据恢复后，系统将强制各集群的LdapServer从OLadp同步一次数据。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。

- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查OMS资源状态是否正常，检查各集群的LdapServer实例状态是否正常。如果不正常，不能执行恢复操作。
- 检查集群主机和服务的状态是否正常。如果不正常，不能执行恢复操作。
- 检查恢复数据时集群主机拓扑结构与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份。
- 检查恢复数据时集群中已添加的服务与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份。
- 停止依赖集群运行的上层业务应用。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理” 。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 选择“运维 > 备份恢复 > 恢复管理 > 创建 ”。

步骤4 在“任务名称”填写恢复任务的名称。

步骤5 在“恢复对象”选择“OMS”。

步骤6 勾选“OMS”。

步骤7 在“OMS”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “恢复时使用集群”：表示选择恢复任务执行时使用集群的名称。
  - “源NameService名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “源集群”：选择恢复数据使用的Yarn队列所在的集群。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “SFTP”：表示备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤8 单击“确定”保存。

步骤9 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤10 以**omm**用户分别登录主、备管理节点。

步骤11 执行以下命令，重新启动OMS。

sh ${BIGDATA_HOME}/om-server/om/sbin/restart-oms.sh

提示以下信息表示命令执行成功：

```
start HA successfully.
```

执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**，查看管理节点的“HAAllResOK”是否为“Normal”，并可以重新登录FusionInsight Manager表示重启OMS成功。

步骤12 在FusionInsight Manager，选择“集群 > *待操作的集群名称* > 服务 > KrbServer > 更多 > 同步配置”，单击“确定”，等待KrbServer服务配置同步过程完成。

步骤13 选择“集群 > *待操作集群的名称* > 更多 > 同步配置”，单击“确定”，等待集群配置同步成功。

步骤14 选择“集群 > *待操作集群的名称* > 更多 > 重启”，输入当前登录的用户密码确认身份，单击“确定”，等待集群重启成功。

----结束

11.11.3.2 恢复 CDL 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对CDL进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对CDL进行恢复数据操作。

CDL的元数据存储在DBService与Kafka服务中，系统管理员可以通过FusionInsight Manager创建恢复DBService与Kafka任务恢复CDL数据。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的DBService与Kafka数据。
- MRS集群中默认使用DBService保存Hive、Hue、Loader、Spark、Oozie、CDL的元数据。恢复DBService的数据将恢复全部相关组件的元数据。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。
- 元数据恢复后，Kafka的消费者在ZooKeeper上保存的offset信息将会回退，可能导致重复消费。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查DBService主备实例状态是否正常。如果不正常，不能执行恢复操作。
- 先停止Kafka服务，待恢复完成后，再启动Kafka服务。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“DBService”和“Kafka”。

说明

若安装了多个DBService或Kafka服务，请勾选需要恢复的DBservice、Kafka服务名称。

步骤8 在“DBService”的“路径类型”，选择一个备份目录的类型。具体参数配置要求请参考**步骤8**。

步骤9 在“Kafka”的“路径类型”，选择一个备份目录的类型。具体参数配置要求请参考**步骤8**。

步骤10 单击“确定”保存。

步骤11 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.3 恢复 ClickHouse 元数据

操作场景

在用户意外修改、删除或需要找回数据时，集群用户需要对ZooKeeper进行重大操作（如升级或迁移等）后，系统数据出现异常或未达到预期结果，导致ClickHouse组件全部故障无法使用，或者迁移数据到新集群的场景中，需要对ClickHouse元数据进行恢复数据操作。

集群用户可以通过FusionInsight Manager创建恢复ClickHouse任务。只支持创建任务手动恢复数据。

须知

- 该功能仅MRS 3.1.0及之后版本支持。
- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复ClickHouse元数据，建议手动备份最新ClickHouse元数据后，再执行恢复ClickHouse元数据操作。否则会丢失从备份时刻到恢复时刻之间的ClickHouse元数据。
- ClickHouse元数据恢复和业务数据恢复不能同时进行操作，否则会导致业务数据恢复失败。建议元数据恢复完成后再进行业务数据恢复。

对系统的影响

- 元数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 元数据恢复后，需要重新启动ClickHouse的上层应用。

前提条件

- 检查ClickHouse元数据备份文件保存路径。

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主备集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主备集群部署为普通模式，则不需要配置互信。

操作步骤

步骤1 在FusionInsight Manager，选择 “运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”。

在弹出的窗口中，指定一次成功执行的“备份路径”，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”中选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“元数据和其他数据”下的“ClickHouse”。

步骤8 在“ClickHouse”的“路径类型”，选择一个恢复目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示从主管理节点的本地磁盘上恢复数据。

  选择此参数值，还需要配置:
  - “源端路径”：表示要恢复的备份文件。例如，“备份任务名_数据源_任务执行时间.tar.gz”。
  - “逻辑集群”：填写备份数据的ClickHouse逻辑集群。
- “RemoteHDFS”：表示从备集群的HDFS目录恢复数据。

  MRS 3.2.0及之后版本选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端主NameNode IP地址”：填写备集群主NameNode业务平面IP地址。
  - “源端备NameNode IP地址”：填写备集群备NameNode业务平面IP地址。
  - “源端NameNode RPC端口”：填写目的端集群HDFS基础配置“dfs.namenode.rpc.port”参数值。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“备份路径/备份任务名_数据源_任务创建时间/*数据源_任务执行时间.tar.gz*”。
  - MRS 3.2.0及之后版本“逻辑集群”：填写备份数据的ClickHouse逻辑集群。

  MRS 3.1.0和MRS 3.1.2版本选择此参数值，还需要配置以下参数：

- “源端NameService名称”：填写备份数据集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
- “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“备份路径/备份任务名_数据源_任务创建时间/*数据源_任务执行时间.tar.gz*”。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，在弹出窗口上继续单击“确定”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤11 在“集群 > 服务”页面，启动ClickHouse服务。

----结束

11.11.3.4 恢复 ClickHouse 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，集群用户对ClickHouse进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对ClickHouse进行恢复数据操作。

集群用户可以通过FusionInsight Manager创建恢复ClickHouse任务并恢复数据。只支持创建任务手动恢复数据。

ClickHouse备份恢复功能不支持识别用户的ClickHouse表、索引、视图等对象在业务和结构上存在的关联关系。用户在执行备份恢复任务时，需要根据业务场景管理统一的恢复点，防止影响业务正常运行。

须知

- 该功能仅MRS 3.1.0及之后版本支持。
- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的ClickHouse数据。
- ClickHouse元数据恢复和业务数据恢复不能同时进行操作，否则会导致业务数据恢复失败。建议元数据恢复完成后再进行业务数据恢复。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动ClickHouse的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主备集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主备集群部署为普通模式，则不需要配置互信。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 规划好恢复数据保存表的数据库，数据表在HDFS的保存位置，以及访问恢复数据的用户清单。
- 检查ClickHouse备份文件保存路径。
- 停止ClickHouse的上层应用。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”的“业务数据”下，勾选“ClickHouse”。

步骤8 在“ClickHouse”的“路径类型”，选择一个备份目录的类型。

备份目录当前仅支持RemoteHDFS类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。

- “源端路径”：参考**步骤2**中获取的“备份路径”，填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.5 恢复 DBService 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对DBService进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对DBService进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复DBService任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的DBService数据。
- MRS集群中默认使用DBService保存Hive、Hue、Loader、Spark、Oozie、CDL的元数据。恢复DBService的数据将恢复全部相关组件的元数据。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。

- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查DBService主备实例状态是否正常。如果不正常，不能执行恢复操作。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“DBService”。

说明

若安装了多个DBService服务，请勾选需要恢复的DBservice服务名称。

步骤8 在“DBService”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “源NameService名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。

  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

说明

MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.6 恢复 HBase 元数据

操作场景

为了确保HBase元数据（主要包括tableinfo文件和HFile）安全，防止因HBase的系统表目录或者文件损坏导致HBase服务不可用，或者系统管理员需要对HBase系统表进行重大操作（如升级或迁移等）时，需要对HBase元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建恢复HBase任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的HBase数据。
- 建议一个恢复任务只恢复一个组件的元数据，避免因停止某个服务或实例影响其他组件的数据恢复。同时恢复多个组件数据，可能导致数据恢复失败。

  HBase元数据不能与NameNode元数据同时恢复，会导致数据恢复失败。

对系统的影响

- 元数据恢复前，需要停止HBase服务，在这期间所有上层应用都会受到影响，无法正常工作。
- 元数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 元数据恢复后，需要重新启动HBase的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。

- 检查HBase元数据备份文件保存路径。
- 恢复HBase元数据需要先停止HBase服务。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择 “运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。
  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“元数据和其他数据”下的“HBase”。

说明

若安装了多个HBase服务，请勾选需要恢复的HBase服务名称。

步骤8 在“HBase”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。
  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。
  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。

- “NFS”：表示将备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。

- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.7 恢复 HBase 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对HBase进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对HBase业务数据进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复HBase任务并恢复数据。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的HBase数据。

对系统的影响

- 恢复过程的数据还原阶段，系统会把待恢复的HBase表禁用，此时无法访问该表。还原阶段可能需要几分钟时间，此时HBase的上层应用无法正常工作。
- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动HBase的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查HBase备份文件保存路径。
- 停止HBase的上层应用。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“业务数据”下的“HBase”。

步骤8 在“HBase”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的Yarn队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。

  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

步骤9 在“数据配置”中的“备份数据”列根据业务需要勾选一个或多个需要恢复的已备份数据，并在“目标名称空间”列，指定备份数据恢复的命名空间。

“目标名称空间”建议选择一个备份命名空间不同的位置。

步骤10 在“强制覆盖”选择“true”，表示存在同名数据表时强制恢复备份的所有数据，如果数据表中存在备份后新增加的数据，那恢复后将丢失这些数据。选择“false”表示存在同名表时不执行恢复任务。

步骤11 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的命名空间，校验失败。
- 如果不满足强制覆盖的条件，校验失败。

步骤12 单击“确定”保存。

步骤13 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤14 检查是否是在全新安装，或者重新安装HBase的环境中恢复了HBase数据。

- 是，需要管理员在FusionInsight Manager上根据原有的业务规划重新设置角色的权限。
- 否，任务结束。

----结束

11.11.3.8 恢复 NameNode 数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对NameNode进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对NameNode进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复NameNode任务并恢复数据。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的NameNode数据。
- 建议一个恢复任务只恢复一个组件的元数据，避免因停止某个服务或实例影响其他组件的数据恢复。同时恢复多个组件数据，可能导致数据恢复失败。

  HBase元数据不能与NameNode元数据同时恢复，会导致数据恢复失败。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。

- 恢复数据后需要重启NameNode，重启完成前NameNode不可访问。
- 恢复数据后可能导致元数据与业务数据无法匹配，HDFS进入安全模式且HDFS服务启动失败。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 登录FusionInsight Manager，请参见**登录管理系统**。
- 在FusionInsight Manager停止所有待恢复数据的NameNode角色实例，其他的HDFS角色实例必须保持正常运行，恢复数据后重启NameNode。NameNode角色实例重启前无法访问。
- 检查NameNode备份文件保存路径是否保存在主管理节点“*数据存放路径*/LocalBackup/”。

操作步骤

步骤1 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > NameNode”，查看待恢复数据的NameNode角色实例是否已经停止，如果NameNode角色实例未停止，请停止NameNode角色实例运行。

步骤2 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤3 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。
  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤4 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤5 单击“创建”。

步骤6 在“任务名称”填写恢复任务的名称。

步骤7 在“恢复对象”选择待操作的集群。

步骤8 在“恢复配置”，勾选“NameNode”。

步骤9 在“NameNode”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。
  选择此参数值，还需要配置以下参数：

- “源端路径”：表示备份文件在本地磁盘中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
  - “目标NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。
- “NFS”：表示将备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
- “CIFS”：表示将备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “NameService名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤10 单击“确定”保存。

步骤11 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤12 在FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 更多 > 重启服务”。

在弹出窗口中输入当前登录的管理员密码确认管理员身份，单击“确定”。界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

11.11.3.9 恢复 HDFS 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对HDFS进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对HDFS进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复HDFS任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的HDFS数据。
- 对于Yarn任务运行时使用的目录（例如“/tmp/logs”、“/tmp/archived”、“/tmp/hadoop-yarn/staging”），不能进行HDFS恢复操作，否则进行恢复的Distcp任务会由于文件丢失而导致恢复失败。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动HDFS的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查HDFS备份文件保存路径。
- 停止HDFS的上层应用。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“业务数据”下的“HDFS”。

步骤8 在“HDFS”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：

  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。

- “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
- “用户名”：填写配置CIFS协议时设置的用户名。
- “密码”：填写配置CIFS协议时设置的密码。
- “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
- “队列名称”：填写备份任务执行时使用的YARN队列的名称。
- “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

● “SFTP”：表示备份文件通过SFTP协议保存到服务器中。

选择此参数值，还需要配置以下参数：

- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “服务器IP地址”：填写备份数据的服务器IP地址。
- “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
- “用户名”：填写使用SFTP协议连接服务器时的用户名。
- “密码”：填写使用SFTP协议连接服务器时的密码。
- “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “队列名称”：填写备份任务执行时使用的YARN队列的名称。
- “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。

步骤9 在“数据配置”中的“备份数据”列根据业务需要勾选一个或多个需要恢复的已备份数据，并在“目标路径”列，指定备份数据恢复后的位置。

“目标路径”建议选择一个与目的端路径不同的新路径。

步骤10 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的恢复目录，校验失败。

步骤11 单击“确定”保存。

步骤12 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.10 恢复 Hive 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对Hive进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对Hive进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复Hive任务并恢复数据。只支持创建任务手动恢复数据。

Hive备份恢复功能不支持识别用户的Hive表、索引、视图等对象在业务和结构上存在的关联关系。用户在执行备份恢复任务时，需要根据业务场景管理统一的恢复点，防止影响业务正常运行。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的Hive数据。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动Hive的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 规划好恢复数据保存表的数据库，数据表在HDFS的保存位置，以及访问恢复数据的用户清单。
- 检查Hive备份文件保存路径。
- 停止Hive的上层应用。

- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“Hive”。

步骤8 在“Hive”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

- “RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的Hive备份文件集。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “NFS”：表示备份文件通过NFS协议保存在NAS中。选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的Hive备份文件集。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。
  - “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的Hive备份文件集。
  - “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
  - “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
  - “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“100”。
- “SFTP”：表示备份文件通过SFTP协议保存到服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。

- “队列名称”：填写备份任务执行时使用的YARN队列的名称。
- “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的HDFS目录。
- “目标NameService名称”：选择备份目录对应的目标NameService名称。默认值为“hacluster”。
- “最大map数”：填写执行MapReduce任务的最大map数，默认值为“20”。
- “单个map的最大带宽(MB/s)”：填写单个map最大带宽，默认值为“1”。

步骤9 在“数据配置”的“数据选择”中，根据业务需要勾选一个或多个需要恢复的已备份数据，并分别在“目标数据库”和“目标路径”列，指定备份数据恢复后的数据库和文件保存位置。

配置约束：

- 支持恢复到原数据库，但数据表保存在一个与目的端路径不同的新路径。
- 如果恢复Hive的索引表，请同时选择恢复索引表对应的Hive数据表。
- 如果为了防止影响当前数据，选择了新的恢复目录，那么新目录需要手动授予HDFS权限，使对备份表拥有权限的用户可以访问此目录。
- 支持恢复到其他数据库。如果恢复到其他数据库，那么此数据库对应在HDFS中的目录，需要手动授予HDFS权限，使对备份表拥有权限的用户可以访问此目录。

步骤10 在“强制覆盖”选择“true”，表示存在同名数据表时强制恢复备份的所有数据，如果数据表中存在备份后新增加的数据，那恢复后将丢失这些数据。选择“false”表示存在同名表时不执行恢复任务。

步骤11 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的恢复目录，校验失败。
- 如果不满足强制覆盖的条件，校验失败。

步骤12 单击“确定”保存。

步骤13 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.11 恢复 IoTDB 元数据

操作场景

为了确保IoTDB元数据安全，防止因IoTDB的文件损坏等导致IoTDB服务不可用时，需要对IoTDB元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

系统管理员可以通过FusionInsight Manager创建恢复IoTDB任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的IoTDB数据。
- 建议一个恢复任务只恢复一个组件的元数据，避免因停止某个服务或实例影响其他组件的数据恢复。同时恢复多个组件数据，可能导致数据恢复失败。

对系统的影响

元数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。

操作步骤

步骤1 在FusionInsight Manager，选择 “运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“元数据和其它数据”区域下的“IoTDB”。

说明

若安装了多个IoTDB服务，请勾选需要恢复的IoTDB服务名称。

步骤8 在“路径类型”参数后选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。

  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “NFS”：表示备份文件通过NFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。

  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。

  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端主NameNode IP地址”：填写备集群主NameNode业务平面IP地址。
  - “源端备NameNode IP地址”：填写备集群备NameNode业务平面IP地址。
  - “源端NameNode RPC端口”：填写备集群HDFS基础配置“dfs.namenode.rpc.port”参数值。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
- “SFTP”：表示备份文件通过SFTP协议保存在服务器中。

  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写备份数据的服务器IP地址。
  - “端口号”：填写SFTP协议连接备份服务器使用的端口号，默认值为“22”。
  - “用户名”：填写使用SFTP协议连接服务器时的用户名。
  - “密码”：填写使用SFTP协议连接服务器时的密码。
  - “源端路径”：填写备份文件在备份服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。

- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

步骤11 在“集群 > 服务”页面，启动IoTDB服务。

----结束

11.11.3.12 恢复 IoTDB 业务数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对IoTDB进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，模块全部故障无法使用，或者迁移数据到新集群的场景中，需要对IoTDB进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复IoTDB任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的IoTDB数据。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动IoTDB的上层应用。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 检查IoTDB备份文件保存路径。
- 停止IoTDB的上层应用。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。

  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“业务数据”下的“IoTDB > IoTDB”。

步骤8 在“IoTDB”的“路径类型”，选择一个备份目录的类型。

备份目录支持以下类型：

“RemoteHDFS”：表示将备份文件保存在备集群的HDFS目录。

选择此参数值，还需要配置以下参数：

- “源端NameService名称”：填写备份数据集群的NameService名称。可在备集群HDFS服务的“管理NameService”页面获取，例如“hacluster”。
- “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
- “源端主NameNode IP地址”：填写备集群主NameNode业务平面IP地址。
- “源端备NameNode IP地址”：填写备集群备NameNode业务平面IP地址。
- “源端NameNode RPC端口”：填写备集群HDFS基础配置“dfs.namenode.rpc.port”参数值。
- “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/*”。
- “恢复点列表”：单击“刷新”，然后选择一个备集群上已备份的IoTDB目录。

步骤9 在“数据配置”中的“备份数据”列根据业务需要勾选一个或多个需要恢复的已备份数据，并在“目标路径”列，指定备份数据恢复后的位置。

“目标路径”建议选择一个与目的端路径不同的新路径。

步骤10 在“强制覆盖”选择“true”，表示存在同名数据表时强制恢复备份的所有数据，如果数据表中存在备份后新增加的数据，那恢复后将丢失这些数据。选择“false”表示存在同名表时不执行恢复任务。

步骤11 单击“校验”查看恢复任务的配置是否正确。

- 如果队列名称不正确，校验失败。
- 如果不存在指定的恢复目录，校验失败。

步骤12 单击“确定”保存。

步骤13 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。

- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

----结束

11.11.3.13 恢复 Kafka 元数据

操作场景

在用户意外修改、删除或需要找回数据时，系统管理员对ZooKeeper进行重大操作（如升级、重大数据调整等）后，系统数据出现异常或未达到预期结果，导致Kafka组件全部故障无法使用，或者迁移数据到新集群的场景中，需要对Kafka元数据进行恢复数据操作。

系统管理员可以通过FusionInsight Manager创建恢复Kafka任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复Kafka元数据，建议手动备份最新Kafka元数据后，再执行恢复操作。否则会丢失从备份时刻到恢复时刻之间的Kafka元数据信息。

对系统的影响

- 元数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 元数据恢复后，Kafka的消费者在ZooKeeper上保存的offset信息将会回退，可能导致重复消费。

前提条件

- 如果需要从远端HDFS恢复数据，需要准备备集群。如果主集群部署为安全模式，且主备集群不是由同一个FusionInsight Manager管理，则必须配置系统互信，请参见**配置跨Manager集群互信**。如果主集群部署为普通模式，则不需要配置互信。
- 主备集群必须已配置跨集群拷贝，请参见**启用集群间拷贝功能**。
- 主备集群上的时间必须一致，而且主备集群上的NTP服务必须使用同一个时间源。
- 先停止Kafka服务，待恢复完成后，再启动Kafka服务。
- 登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择 “运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”，打开备份任务执行历史记录。

在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：

- “备份对象”表示备份的数据源。
- “备份路径”表示备份文件保存的完整路径。
  选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤3 在FusionInsight Manager，选择“运维 > 备份恢复 > 恢复管理”。

步骤4 单击“创建”。

步骤5 在“任务名称”填写恢复任务的名称。

步骤6 在“恢复对象”选择待操作的集群。

步骤7 在“恢复配置”，勾选“Kafka”。

步骤8 在“Kafka”的“路径类型”，选择一个备份目录的类型。

选择不同的备份目录时，对应设置如下：

- “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。
  选择此参数值，还需要配置“源端路径”，表示要恢复的备份文件。例如，“*版本号_数据源_任务执行时间.tar.gz*”。
- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。
  选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“*备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “源NameService名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。
- “RemoteHDFS”：表示备份文件保存在备集群的HDFS目录。
  选择此参数值，还需要配置以下参数：
  - “源端NameService名称”：填写备份数据集群的NameService名称。可以输入集群内置的远端集群的NameService名称：haclusterX，haclusterX1，haclusterX2，haclusterX3，haclusterX4；也可输入其他已配置的远端集群NameService名称。
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “源端NameNode IP地址”：填写备集群NameNode业务平面IP地址，支持主节点或备节点。
  - “源端路径”：填写备集群保存备份数据的完整HDFS路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。
  - “队列名称”：填写备份任务执行时使用的YARN队列的名称。需和集群中已存在且状态正常的队列名称相同。
- “NFS”：表示备份文件通过NFS协议保存在NAS中。
  选择此参数值，还需要配置以下参数：
  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “CIFS”：表示备份文件通过CIFS协议保存在NAS中。

  选择此参数值，还需要配置以下参数：

  - “IP 模式”：目标IP的IP地址模式。系统会根据集群网络类型自动选择对应的IP模式，如IPv4或者IPv6。
  - “服务器IP地址”：填写NAS服务器IP地址。
  - “端口号”：填写CIFS协议连接NAS服务器使用的端口号，默认值为“445”。
  - “用户名”：填写配置CIFS协议时设置的用户名。
  - “密码”：填写配置CIFS协议时设置的密码。
  - “源端路径”：填写备份文件在NAS服务器中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

- “OBS”：表示将备份文件保存在OBS中。

  选择此参数值，还需要配置以下参数：

  - “源端路径”：填写备份文件在OBS中保存的完整路径。例如，“*备份路径/备份任务名_数据源_任务创建时间/版本号_数据源_任务执行时间.tar.gz*”。

    说明

    MRS 3.1.0及之后版本才支持将备份文件保存到OBS。

步骤9 单击“确定”保存。

步骤10 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“重试”，重试恢复任务。

须知

- 如果备份完成后删除了Kafka服务，再次重新添加服务，并恢复元数据，重启Kafka服务Broker无法启动，查看“/var/log/Bigdata/kafka/broker/server.log”日志中包含错误，错误样例如下：

  ```
  ERROR Fatal error during KafkaServer startup. Prepare to shutdown (kafka.server.KafkaServer)kafka.common.InconsistentClusterIdException: The Cluster ID kVSgfurUQFGGpHMTBqBPiw doesn't match stored clusterId Some(0Qftv9yBTAmf2iDPSlIk7g) in meta.properties. The broker is trying to join the wrong cluster. Configured zookeeper.connect may be wrong. at kafka.server.KafkaServer.startup(KafkaServer.scala:220) at kafka.server.KafkaServerStartable.startup(KafkaServerStartable.scala:44) at kafka.Kafka$.main(Kafka.scala:84) at kafka.Kafka.main(Kafka.scala)
  ```

  请查看Kafka Broker配置文件“${BIGDATA_HOME}/Fusionsight_Current/*Broker/etc/server.properties中的log.dirs”配置值，此值为kafka数据目录。进入kafka数据目录，将“meta.properties”中的“cluster.id”的值“0Qftv9yBTAmf2iDPSlIk7g”修改为上述错误日志里的最新值“kVSgfurUQFGGpHMTBqBPiw”。
- 每一个Broker所在节点都需做上述修改，修改完成后重启Kafka服务。

----结束

11.11.4 启用集群间拷贝功能

操作场景

当用户需要将保存在HDFS中的数据从当前集群备份到另外一个集群时，需要使用DistCp工具。DistCp工具依赖于集群间拷贝功能，该功能默认未启用。拷贝数据的集群双方都需要配置。

管理员可以根据以下指导，在FusionInsight Manager修改参数以启用集群间拷贝功能。启用之后即可创建将数据备份至远端HDFS（RemoteHDFS）的备份任务。

对系统的影响

启用集群间复制功能需要重启Yarn，服务重启期间无法访问。

前提条件

- 拷贝数据的集群的HDFS的参数“hadoop.rpc.protection”需使用相同的数据传输方式。默认设置为“privacy”表示加密，“authentication”表示不加密。
- 对于安全模式的集群，集群之间需要配置系统互信。

操作步骤

步骤1 登录其中一个集群的FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”，单击“全部配置”。

步骤3 左边菜单栏中选择“Yarn > 集群间拷贝”。

步骤4 修改参数“dfs.namenode.rpc-address”，在“haclusterX.remotenn1”右侧填写对端集群其中一个NameNode实例的业务IP和RPC端口，在“haclusterX.remotenn2”右侧填写对端集群另外一个NameNode实例的业务IP和RPC端口。

“haclusterX.remotenn1”和“haclusterX.remotenn2”不区分主备NameNode。NameNode RPC端口默认为“8020”，不支持通过Manager修改。

修改后参数值例如：“10.1.1.1:8020”和“10.1.1.2:8020”。

说明

- 如果本集群数据要备份至多个集群的HDFS中，可以继续配置对应的NameNode RPC地址至haclusterX1、haclusterX2、haclusterX3、haclusterX4。

步骤5 单击“保存”，并在确认对话框中单击“确定”。

步骤6 重启Yarn服务。

步骤7 登录另外一个集群的FusionInsight Manager，重复**步骤2**～**步骤6**。

----结束

11.11.5 管理本地快速恢复任务

操作场景

使用DistCp备份数据时，本集群HDFS中将保存备份数据的快照信息。FusionInsight Manager支持使用本地的快照快速恢复数据，减少从备集群恢复数据使用的时间。

管理员可以通过FusionInsight Manager与本集群HDFS保存的快照信息，创建本地快速恢复任务并执行恢复任务。

操作步骤

步骤1 登录FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在备份任务列表已创建任务的“操作”列，单击“恢复”。

步骤3 确认界面是否提示“没有可快速恢复的数据，请在恢复管理界面创建恢复任务进行恢复。”。

- 是，备份任务未在主集群产生备份数据快照，任务结束。
- 否，可以创建本地快速恢复任务，执行**步骤4**。

说明

元数据不支持快速恢复。

步骤4 在“任务名称”填写本地快速恢复任务的名称。

步骤5 在“备份配置”选择数据源。

步骤6 在“可恢复点列表”选择一个包含目标备份数据的恢复点。

步骤7 在“队列名称”填写任务执行时使用的Yarn队列的名称。需和集群中已存在且状态正常的队列名称相同。

步骤8 在“数据配置”选择需要恢复的对象。

步骤9 单击“校验”，界面显示“校验恢复任务配置成功”。

步骤10 单击“确定”。

步骤11 在恢复任务列表已创建任务的“操作”列，单击“执行”，开始执行恢复任务。

任务执行完成后，“任务状态”显示为“成功”。

----结束

11.11.6 修改备份任务

操作场景

系统管理员可以通过FusionInsight Manager修改已创建的备份任务的配置参数，以适应业务需求的变化。不支持修改任何恢复任务配置参数，只能查看恢复任务的配置参数。

对系统的影响

修改备份任务后，新的参数在下一次执行任务时生效。

前提条件

- 已创建备份任务。
- 已根据业务实际需求，规划新的备份任务策略。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，单击“配置”，打开修改配置页面。

在新页面中修改任务参数，支持修改的主要参数项如下：

- 开始时间
- 周期
- 目的端NameService名称
- 目的端NameNode IP地址
- 目的端路径
- 最大备份数
- 最大恢复点个数
- 最大map数
- 单个map的最大带宽

说明

修改某个备份任务参数“目的端路径”后，第一次执行此任务默认为全量备份。

步骤3 单击“确定”保存。

----结束

11.11.7 查看备份恢复任务

操作场景

系统管理员可以通过FusionInsight Manager查看已创建的备份恢复任务，以及任务的运行情况。

前提条件

登录FusionInsight Manager，请参见**登录管理系统**。

操作步骤

步骤1 在FusionInsight Manager，选择“运维 > 备份恢复”。

步骤2 单击“备份管理”或“恢复管理”。

步骤3 在任务列表中，查看“任务状态”与“任务进度”列获取上一次任务运行的结果。绿色表示运行成功，红色表示运行失败。

步骤4 在任务列表指定任务的“操作”列，选择“更多 > 查询历史”或单击“查询历史”，打开备份恢复任务运行记录。

在弹出的窗口中，在指定一次执行记录前单击，打开此次任务运行的日志信息。

----结束

相关任务

- 启动备份恢复任务

  在任务列表指定任务的“操作”列，选择“更多 > 即时备份”或单击“执行”，启动处于准备或失败状态的备份恢复任务。已成功执行过的恢复任务不能重新运行。

- 停止备份恢复任务

  在任务列表指定任务的“操作”列，选择“更多 > 停止”或单击“停止”，停止处于运行状态的备份恢复任务。停止成功后，该任务的“任务状态”变为“已停止”。

- 删除备份恢复任务

  在任务列表指定任务的“操作”列，选择“更多 > 删除”或单击“删除”，删除备份恢复任务。删除任务后备份的数据默认会保留。

- 挂起备份任务

  在任务列表指定任务的“操作”列，选择“更多 > 挂起”，挂起备份任务。仅支持周期备份的任务，挂起后周期备份任务不再自动执行。挂起正在执行的备份任务时，该任务会停止运行。需要解锁时，选择“更多 > 重新执行”。

11.11.8 创建 ClickHouse 备份任务路径选择 RemoteHDFS 时的环境配置

说明

本章节仅适用于MRS 3.1.0版本。

问题

通过FusionInsight Manager创建ClickHouse备份任务，路径类型选择RemoteHDFS时，如何先进行环境配置?

回答

步骤1 登录备集群的FusionInsight Manager。

步骤2 选择“集群 > 服务 > HDFS”，单击“更多 > 下载客户端”。“选择客户端类型”设置为“仅配置文件”，根据待安装客户端节点的节点类型选择正确的平台类型后（x86选择**x86_64**，ARM选择**aarch64**）单击“确定”。

步骤3 等待客户端文件包生成后根据浏览器提示下载客户端到本地并解压。

例如，客户端文件压缩包为“FusionInsight_Cluster_1_HDFS_Client.tar”，解压后得到“FusionInsight_Cluster_1_HDFS_ClientConfig_ConfigFiles.tar”继续解压该文件。解压到本地PC的“D:\FusionInsight_Cluster_1_HDFS_ClientConfig_ConfigFiles”目录下（路径中不能有空格）。

步骤4 进入客户端路径“FusionInsight_Cluster_1_HDFS_ClientConfig_ConfigFiles\”，获取“hosts”文件。

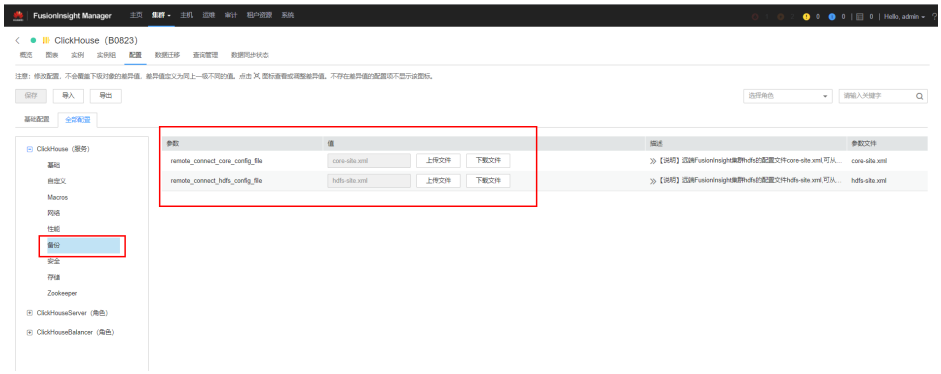
步骤5 进入客户端路径“FusionInsight_Cluster_1_HDFS_ClientConfig_ConfigFiles\HDFS\config”，获取“core-site.xml”和“hdfs-site.xml”文件。

步骤6 登录源集群的FusionInsight Manager。

步骤7 选择“集群 > 服务 > ClickHouse”，单击“配置 > 全部配置”，在“ClickHouse（服务）”下选择“备份”。

在“remote_connect_core_config_file”参数单击“上传文件”，选择**步骤5**中准备的“core-site.xml”文件上传。

在“remote_connect_hdfs_config_file”参数单击“上传文件”，选择**步骤5**中准备的“hdfs-site.xml”文件上传。



步骤8 单击“保存”，确认提示信息，单击“确定”保存配置。保存完成后，单击“完成”。

步骤9 选择“集群 > 服务 > ClickHouse”，单击“实例”，查看“ClickHouseServer”的实例IP。

步骤10 以**root**用户分别登录ClickHouseServer实例对应的主机节点，检查“/etc/hosts”文件内是否包含**步骤4**中主机信息，如未包含，添加**步骤4**中主机信息到“/etc/hosts”文件中。

----结束

11.12 安全管理

11.12.1 安全概述

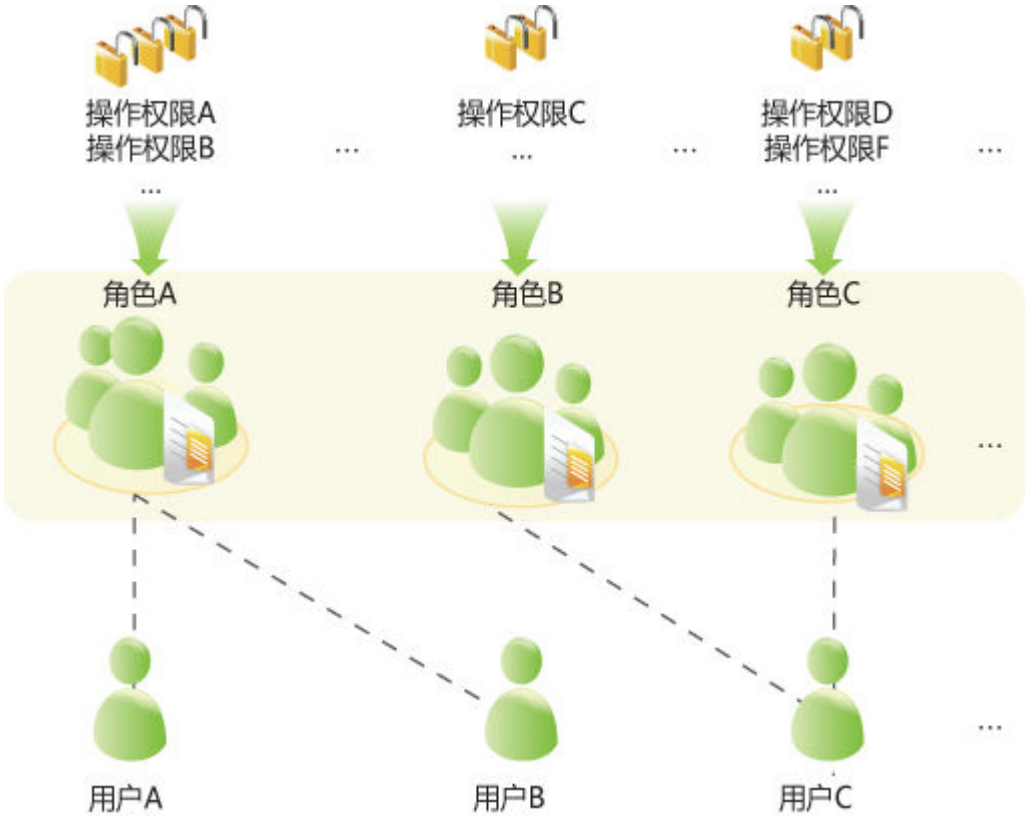
11.12.1.1 权限模型

基于角色的权限控制

FusionInsight通过采用RBAC（role-based access control，基于角色的权限控制）方式对大数据系统进行权限管理，将系统中各组件零散的权限管理功能集中呈现和管理，对普通用户屏蔽掉了内部的权限管理细节，对管理员简化了权限管理的操作方法，提升权限管理的易用性和用户体验。

FusionInsight权限模型由“用户－用户组－角色－权限”四类对象构成。

图 11-14 权限模型



- **权限**

  由组件侧定义，允许访问组件某个资源的能力。不同组件针对自己的资源，有不同的权限。

  例如：

  - HDFS针对文件资源权限，有读、写、执行等权限。
  - HBase针对表资源权限，有创建、读、写等权限。

- **角色**

  组件权限的一个集合，一个角色可以包含多个组件的多个权限，不同的角色也可以拥有同一个组件的同一个资源的权限。

- **用户组**

  用户的集合，当用户组关联某个或者多个角色后，该用户组内的用户就将拥有这些角色所定义的组件权限。

  不同用户组可以关联同一个角色，一个用户组也可以不关联任何角色，该用户组原则上将不具有任何组件资源的权限。

  说明

  部分组件针对特定的默认用户组，系统默认赋予了部分权限。

- **用户**

  系统的访问者，每个用户的权限由该用户关联的用户组和角色所对应的权限构成，用户需要加入用户组或者关联角色来获得对应的权限。

基于策略的权限控制

Ranger组件通过PBAC（policy-based access control，基于策略的权限控制）方式进行权限管理，可对HDFS、Hive、HBase等组件进行更加细粒度的数据访问控制。

说明

组件同时只支持一种权限控制机制，当组件启用Ranger权限控制策略后，通过FusionInsight Manager创建的角色中关于该组件的权限将失效（HDFS与Yarn的组件ACL规则仍将生效），用户需通过Ranger管理界面添加策略进行资源的赋权。

Ranger的权限模型由多条权限策略组成，权限策略主要由以下几方面组成：

- 资源

  组件所提供的可由用户访问的对象，例如HDFS的文件或文件夹、Yarn中的队列、Hive中的数据库/表/列等。

- 用户

  系统的访问者，每个用户的权限由该用户关联的策略来获得。LDAP中的用户、用户组、角色信息会周期性的同步至Ranger。

- 权限

  策略中针对资源可配置各种访问条件，例如文件的读写，具体可以配置允许条件、拒绝条件以及例外条件等。

11.12.1.2 权限机制

FusionInsight采用LDAP存储用户和用户组的数据；角色的定义信息保存在关系数据库中，角色和权限的对应关系则保存在组件侧。

FusionInsight使用Kerberos进行统一认证。

用户权限校验流程大致如下：

1. 客户端（用户终端或FusionInsight组件服务）调用FusionInsight认证接口。
2. FusionInsight使用登录用户名和密码，到Kerberos进行认证。
3. 如果认证成功，客户端会发起访问服务端（FusionInsight组件服务）的请求。
4. 服务端会根据登录的用户，找到其属于的用户组和角色。
5. 服务端获得用户组拥有的所有权限和角色拥有的所有权限的并集。
6. 服务端判断客户端是否有权限访问其请求的资源。

示例场景（RBAC）：

HDFS中有三个文件fileA、fileB、fileC。

- 定义角色roleA对fileA有读和写权限，角色roleB对fileB有读权限。
- 定义groupA属于roleA；groupB属于roleB。
- 定义userA属于groupA和roleB，userB属于GroupB。

当userA登录成功并访问HDFS时：

1. HDFS获得useA属于的所有角色（roleB）。
2. HDFS同时还会获得userA属于的所有用户组所属于的角色（roleA）。
3. 此时，userA拥有roleA和roleB对应权限的并集。

4. 因此对于fileA，则userA有读写权限；对fileB，有读权限；对于fileC，无任何权限。

同理userB登录后：

1. userB只拥有roleB对应的权限。
2. 对于fileA，则userB无权限；对fileB，有读权限；对于fileC，无任何权限。

11.12.1.3 认证策略

大数据平台用户需要对用户进行身份认证，防止不合法用户访问集群。安全模式或者普通模式的集群均提供认证能力。

安全模式

安全模式的集群统一使用Kerberos认证协议进行安全认证。Kerberos协议支持客户端与服务端进行相互认证，提高了安全性，可有效消除使用网络发送用户凭据进行模拟认证的安全风险。集群中由KrbServer服务提供Kerberos认证支持。

Kerberos用户对象

Kerberos协议中，每个用户对象即一个principal。一个完整的用户对象包含两个部分信息：用户名和域名。在运维管理或应用开发的场景中，需要在客户端认证用户身份后才能连接到集群服务端。系统操作运维与业务场景中主要使用的用户分为“人机”用户和“机机”用户。二者主要区别在于“机机”用户密码由系统随机生成。

Kerberos认证

Kerberos认证支持两种方式：密码认证及keytab认证。认证有效时间默认为24小时。

- 密码认证：通过输入用户正确的密码完成身份认证。主要在运维管理场景中使用“人机”用户进行认证，命令为**kinit** *用户名*。
- keytab认证：keytab文件包含了用户principal和用户凭据的加密信息。使用keytab文件认证时，系统自动使用加密的凭据信息进行认证无需输入用户密码。主要在组件应用开发场景中使用“机机”用户进行认证。keytab文件也支持在**kinit**命令中使用。

普通模式

普通模式的集群不同组件使用原生开源的认证机制，不支持**kinit**认证命令。FusionInsight Manager（含DBService、KrbServer和LdapServer）使用的认证方式为用户名密码方式。组件使用的认证机制如**表11-85**所示。

表 11-85 组件认证方式一览表

服务	认证方式
IoTDB	simple认证
CDL	无认证
ClickHouse	simple认证
Flume	无认证

服务	认证方式
HBase	● WebUI：无认证 ● 客户端：simple认证
HDFS	● WebUI：无认证 ● 客户端：simple认证
Hive	simple认证
Hue	用户名密码认证
Kafka	无认证
Loader	● WebUI：用户名密码认证 ● 客户端：无认证
Mapreduce	● WebUI：无认证 ● 客户端：无认证
Oozie	● WebUI：用户名密码认证 ● 客户端：simple认证
Spark2x	● WebUI：无认证 ● 客户端：simple认证
Storm	无认证
Yarn	● WebUI：无认证 ● 客户端：simple认证
ZooKeeper	simple认证

认证方式解释如下：

- “simple认证”：在客户端连接服务端的过程中，默认以客户端执行用户（例如操作系统用户“root”或“omm”）自动进行认证，管理员或业务用户不显式感知认证，不需要**kinit**完成认证过程。
- “用户名密码认证”：使用集群中“人机”用户的用户名与密码进行认证。
- “无认证”：默认任意的用户都可以访问服务端。

11.12.1.4 鉴权策略

安全模式

大数据平台用户完成身份认证后，系统还需要根据实际权限管理配置，选择是否对用户进行鉴权，确保系统用户拥有资源的有限或全部权限。如果系统用户权限不足，需要由系统管理员为用户授予各个组件对应的权限后，才能访问资源。安全模式或者普通模式集群均提供鉴权能力，组件的具体权限项在两种模式中相同。

新安装的安全模式集群默认即安装了Ranger服务并启用了Ranger鉴权，用户可以通过组件的权限插件对组件资源的访问设置细粒度的安全访问策略。若不需使用Ranger进

行鉴权，管理员可在服务页面手动停用Ranger鉴权，停用Ranger鉴权后，访问组件资源的时系统将继续基于FusionInsight Manager的角色模型进行权限控制。

安全模式集群中，支持使用Ranger鉴权的组件包括：HDFS、Yarn、Kafka、Hive、HBase、Storm、Spark2x、Impala、CDL。

从历史版本升级的集群，用户访问组件资源时默认不使用Ranger鉴权，管理员可在安装了Ranger服务后手动启用Ranger鉴权。

安全版本的集群所有组件默认统一对及访问进行鉴权，不支持关闭鉴权功能。

普通模式

普通模式的集群不同组件使用各自原生开源的鉴权行为，详细鉴权机制如**表11-86**所示。

在安装了Ranger服务的普通模式集群中，Ranger可以支持基于OS用户进行组件资源的权限控制，支持启用Ranger鉴权的组件包括：HBase、HDFS、Hive、Spark2x、Yarn。

表 11-86 普通模式组件鉴权一览表

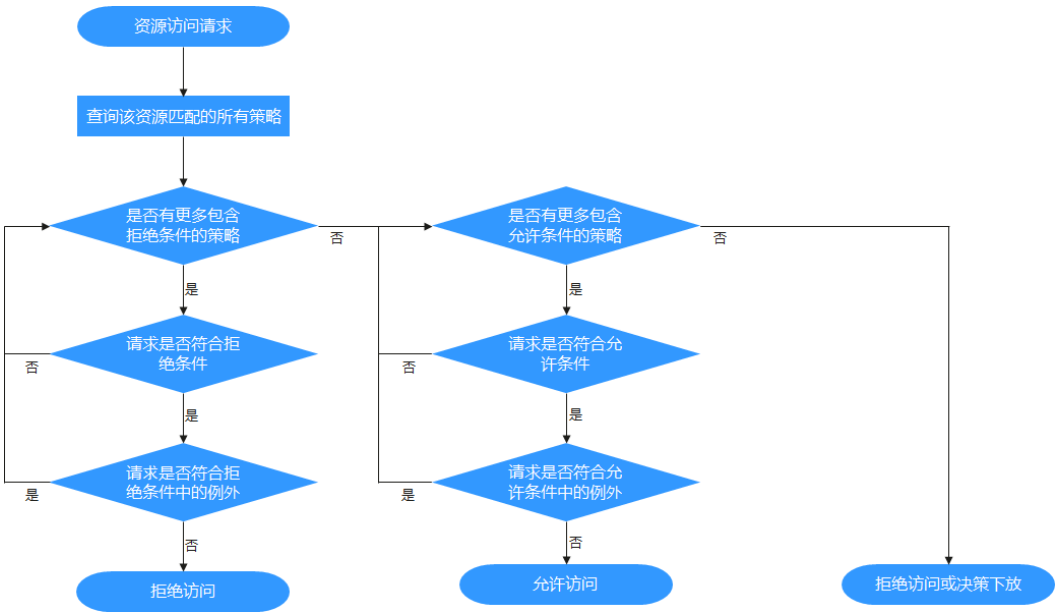
服务	是否鉴权	是否支持开关鉴权
IoTDB	鉴权	不支持修改
ClickHouse	鉴权	不支持修改
Flume	无鉴权	不支持修改
HBase	无鉴权	支持修改
HDFS	鉴权	支持修改
Hive	无鉴权	不支持修改
Hue	无鉴权	不支持修改
Kafka	无鉴权	不支持修改
Loader	无鉴权	不支持修改
Mapreduce	无鉴权	不支持修改
Oozie	鉴权	不支持修改
Spark2x	无鉴权	不支持修改
Storm	无鉴权	不支持修改
Yarn	无鉴权	支持修改
ZooKeeper	鉴权	支持修改
CDL	无鉴权	不支持修改

Ranger 权限策略条件判断优先级

配置资源的权限策略时，可配置针对该资源的允许条件（Allow Conditions）、允许例外条件（Exclude from Allow Conditions）、拒绝条件（Deny Conditions）以及拒绝例外条件（Exclude from Deny Conditions），以满足不同场景下的例外需求。

不同条件的优先级由高到低为：拒绝例外条件 > 拒绝条件 > 允许例外条件 > 允许条件。

系统判断流程可参考下图所示，如果组件资源请求未匹配到Ranger中的权限策略，系统默认将拒绝访问。但是对于HDFS和Yarn，系统会将决策下放给组件自身的访问控制层继续进行判断。



例如要将一个文件夹FileA的读写权限授权给用户组groupA，但是该用户组内某个用户UserA除外，这时可以增加一个允许条件及一个例外条件即可实现。

11.12.1.5 用户账号一览表

用户分类

MRS集群提供以下3类用户，请系统管理员定期修改密码，不建议使用默认密码。

说明

本章节介绍MRS集群内的相关默认用户信息。

用户类型	使用说明
系统用户	• 通过FusionInsight Manager创建，是系统操作运维与业务场景中主要使用的用户，包含两种类型： – “人机”用户：用于在FusionInsight Manager的操作运维场景，以及在组件客户端操作的场景。创建此类型用户时需要参考**创建用户**设置“密码”和“确认密码”。 – “机机”用户：用于系统应用开发的场景。 • 用于OMS系统进程运行的用户。
系统内部用户	集群提供的用于Kerberos认证、进程通信、保存用户组信息和关联用户权限的内部用户。系统内部用户不建议在操作与维护的场景下使用。请通过**admin**用户操作，或联系系统管理员根据业务需要创建新用户。
数据库用户	• 用于OMS数据库管理和数据访问的用户。 • 用于业务组件（Hue、Hive、Loader、Oozie、Ranger和DBService）数据库的用户。

系统用户

说明

- 需要使用操作系统中**root**用户，所有节点**root**用户需设置为统一的密码。
- 需要使用操作系统中**ldap**用户，此账号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作系统管理员维护。

类别	用户名称	初始密码	描述	密码修改方法
系统管理员	admin	用户创建集群时自定义	FusionInsight Manager的管理员。 **说明** **admin**用户默认不具备其他组件的管理权限，例如访问组件原生界面时，需要使用具备对应组件管理权限的用户才可以访问到完整内容。	请参见**修改admin密码**。
节点操作系统用户	ommdba	随机密码	创建系统数据库的用户。在管理节点生成，属于操作系统用户，无需设置为统一的密码。该用户不能用于远程登录。	请参见**修改操作系统用户密码**。
	omm	Bigdata123@	系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。	

系统内部用户

类别	默认用户	初始密码	描述	密码修改方法
Kerberos管理员	kadmin/admin	Admin@123	用于增加、删除、修改及查询Kerberos上的用户账号。	请参见**修改Kerberos管理员密码**。
OMS Kerberos管理员	kadmin/admin	Admin@123	用于增加、删除、修改及查询OMS Kerberos上的用户账号。	请参见**修改OMS Kerberos管理员密码**。
LDAP管理员	cn=root,dc=hadoop,dc=com	• MRS 3.1.2之前版本：LdapChangeMe@123 • MRS 3.1.2及之后版本：系统随机生成	用于增加、删除、修改及查询LDAP用户账号信息。	• MRS 3.1.2之前版本：请参见**修改LDAP管理员和LDAP用户密码（含OMS LDAP）**。 • MRS 3.1.2及之后版本：请参见**修改OMS服务配置参数**
OMS LDAP管理员	cn=root,dc=hadoop,dc=com	• MRS 3.1.2之前版本：LdapChangeMe@123 • MRS 3.1.2及之后版本：系统随机生成	用于增加、删除、修改及查询OMS LDAP用户账号信息。	
LDAP用户	cn=pg_search_dn,ou=Users,dc=hadoop,dc=com	系统随机生成	用于查询LDAP中存储的用户和用户组信息。	

类别	默认用户	初始密码	描述	密码修改方法
OMS LDAP用户	cn=pg_search_dn,ou=Users,dc=hadoop,dc=com	系统随机生成	用于查询OMS LDAP中存储的用户和用户组信息。	
LDAP管理账户	cn=krbkdc,ou=Users,dc=hadoop,dc=com	• MRS 3.1.2之前版本：LdapChangeMe@123 • MRS 3.1.2及之后版本：系统随机生成	用于查询Kerberos组件认证账户信息。	• MRS 3.1.2之前版本：请参见**修改LDAP管理账户密码**。 • MRS 3.1.2及之后版本：请参见**修改OMS服务配置参数**
	cn=krbadmin,ou=Users,dc=hadoop,dc=com	• MRS 3.1.2之前版本：LdapChangeMe@123 • MRS 3.1.2及之后版本：系统随机生成	用于增加、删除、修改及查询Kerberos组件认证账户信息。	

类别	默认用户	初始密码	描述	密码修改方法
组件运行用户	hdfs	Hdfs@123	HDFS系统管理员，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看webUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统 3. 登录FTP服务	请参见**修改组件运行用户密码**。
	hbase	Hbase@123	HBase，HBase1~4系统管理员，用户权限： ● 集群管理权限：表的Enable、Disable操作，触发MajorCompact，ACL操作 ● 授权或回收权限，集群关闭等操作相关的权限 ● 表管理权限：建表、修改表、删除表等操作权限 ● 数据管理权限：表级别、列族级别以及列级别的数据读写权限 ● 登录HMaster WebUI界面 ● 登录FTP服务	

类别	默认用户	初始密码	描述	密码修改方法
	mapred	Mapred@123	MapReduce系统管理员，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 登录FTP服务 ● 登录Yarn WebUI界面	
	zookeeper	ZooKeeper@123	ZooKeeper系统管理员，用户权限： ● 对Zookeeper上所有节点的增删改查权限 ● 对Zookeeper上所有节点的配额修改查询权限	
	rangeradmin	Rangeradmin@123	Ranger的系统管理权限，用户权限。 ● Ranger Web UI的管理权限 ● 使用Ranger鉴权的各组件管理权限	
	rangerauditor	Rangerauditor@123	Ranger系统的默认审计用户。	

类别	默认用户	初始密码	描述	密码修改方法
	hive	Hive@123	Hive系统管理员，用户权限： 1. Hive管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	
	hive1	Hive1@123	Hive1系统管理员，用户权限： 1. Hive1管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hive2	Hive2@123	Hive2系统管理员，用户权限： 1. Hive2管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	
	hive3	Hive3@123	Hive3系统管理员，用户权限： 1. Hive3管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	

类别	默认用户	初始密码	描述	密码修改方法
	hive4	Hive4@123	Hive4系统管理员，用户权限： 1. Hive4管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限。 4. Ranger策略的管理权限。	
	kafka	Kafka@123	Kafka的系统管理员，用户权限: ● Topic的创建、删除、生产、消费、配置修改。 ● Cluster的元数据控制、配置修改、副本迁移、leader选举、acl管理。 ● ConsumerGroup Offset的提交、查询、删除。 ● DelegationToken的查询。 ● Transaction的查询、提交。	
	storm	Admin@123	storm的系统管理员。 用户权限：storm任务提交。	

类别	默认用户	初始密码	描述	密码修改方法
	ranger usersync	系统随机生成	用于同步用户及用户组的内部用户。	
	rangertagsync	系统随机生成	用于同步标签的内部用户。	
	oms/manager	系统随机生成	用于Controller和NodeAgent认证的用户，拥有“supergroup”组权限。	
	backup/manager	系统随机生成	用于运行备份恢复任务的用户，拥有“supergroup”、“wheel”和“ficommon”组权限。配置跨系统互信后拥有访问互信系统HDFS、HBase、Hive、ZooKeeper数据的权限。	
	hdfs/hadoop.*<系统域名>*	系统随机生成	HDFS系统启动用户，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看WebUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统 3. 登录FTP服务	

类别	默认用户	初始密码	描述	密码修改方法
	mapred/hadoop.*<系统域名>*	系统随机生成	MapReduce系统启动用户，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 登录FTP服务 ● 登录Yarn WebUI界面	
	mr_zk/hadoop.*<系统域名>*	系统随机生成	用于MapReduce访问ZooKeeper。	
	hbase/hadoop.*<系统域名>*	系统随机生成	HBase系统启动过程用于内部组件之间认证的用户。	
	hbase/zkclient.*<系统域名>*	系统随机生成	安全集群下，HBase做ZooKeeper认证时使用的用户。	
	thrift/hadoop.*<系统域名>*	系统随机生成	ThriftServer系统启动用户。	
	thrift/*<hostname>*	系统随机生成	ThriftServer系统访问HBase的用户，拥有HBase所有NameSpace和表的读、写、执行、创建和管理的权限。*<hostname>*表示集群中安装ThriftServer节点的主机名。	

类别	默认用户	初始密码	描述	密码修改方法
	hive/hadoop.*<系统域名>*	系统随机生成	Hive系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	
	hive1/hadoop.*<系统域名>*	系统随机生成	Hive1系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive1管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	

类别	默认用户	初始密码	描述	密码修改方法
	hive2/hadoop.*<系统域名>*	系统随机生成	Hive2系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive2管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	
	hive3/hadoop.*<系统域名>*	系统随机生成	Hive3系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive3管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	

类别	默认用户	初始密码	描述	密码修改方法
	hive4/hadoop.*<系统域名>*	系统随机生成	Hive4系统启动过程用于内部组件之间认证的用户，用户权限： 1. Hive4管理员权限： ● 数据库的创建、删除、修改 ● 表的创建、查询、修改、删除 ● 数据的查询、插入、加载 2. HDFS文件操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 3. 提交、停止MapReduce任务的权限	
	loader/hadoop.*<系统域名>*	系统随机生成	Loader系统启动与Kerberos认证用户。	
	HTTP/*<hostname>*	系统随机生成	用于连接各组件的HTTP接口，*<hostname>*表示集群中节点主机名。	
	hue	系统随机生成	Hue系统启动与Kerberos认证用户，并用于访问HDFS和Hive。	
	flume	系统随机生成	Flume系统启动用户，用于访问HDFS和Kafka，对HDFS目录“/flume”有读写权限。	
	flume_server	系统随机生成	Flume系统启动用户，用于访问HDFS和Kafka，对HDFS目录“/flume”有读写权限。	

类别	默认用户	初始密码	描述	密码修改方法
	spark2x/hadoop.*<系统域名>*	系统随机生成	Spark2x系统管理员用户，用户权限： 1、Spark2x服务启动用户 2、提交Spark2x任务的权限	
	spark_zk/hadoop.*<系统域名>*	系统随机生成	用于Spark2x访问ZooKeeper。	
	spark2x1/hadoop.*<系统域名>*	系统随机生成	Spark2x1系统管理员用户，用户权限： 1. Spark2x1服务启动用户 2. 提交Spark2x任务的权限	
	spark2x2/hadoop.*<系统域名>*	系统随机生成	Spark2x2系统管理员用户，用户权限： 1. Spark2x2服务启动用户 2. 提交Spark2x任务的权限	
	spark2x3/hadoop.*<系统域名>*	系统随机生成	Spark2x3系统管理员用户，用户权限： 1. Spark2x3服务启动用户 2. 提交Spark2x任务的权限	
	spark2x4/hadoop.*<系统域名>*	系统随机生成	Spark2x4系统管理员用户，用户权限： 1. Spark2x4服务启动用户 2. 提交Spark2x任务的权限	
	zookeeper/hadoop.*<系统域名>*	系统随机生成	ZooKeeper系统启动用户。	

类别	默认用户	初始密码	描述	密码修改方法
	zkcli/hadoop.*<系统域名>*	系统随机生成	登录Zookeeper服务器用户。	
	oozie	系统随机生成	Oozie系统启动与Kerberos认证用户。	
	kafka/hadoop.*<系统域名>*	系统随机生成	用于Kafka安全认证。	
	storm/hadoop.*<系统域名>*	系统随机生成	Storm系统启动用户。	
	storm_zk/hadoop.*<系统域名>*	系统随机生成	用于Worker进程访问ZooKeeper。	
	flink/hadoop.*<系统域名>*	系统随机生成	Flink服务的内部用户。	
	check_ker_M	系统随机生成	系统内部测试Kerberos服务功能是否正常的用户。	
	clickhouse/hadoop.<系统域名>	系统随机生成	用于ClickHouse安全认证。该用户为内部用户，只允许在集群内部使用。	
	default	无	ClickHouse内部用户，仅非安全模式下可使用的管理员用户	
	rangeradmin/hadoop.<系统域名>	系统随机生成	Ranger系统启动用户，用于内部组件之间认证。	
	tez	系统随机生成	TezUI系统启动与Kerberos认证用户，并用于访问Yarn。	

类别	默认用户	初始密码	描述	密码修改方法
	K/M	系统随机生成	Kerberos内部功能用户，不能删除，不支持密码修改，未安装Kerberos服务的节点无法使用内部账户。	无
	kadmin/changepw	系统随机生成		
	kadmin/history	系统随机生成		
	krbtgt/*<系统域名>*	系统随机生成		
LDAP用户	admin	无	FusionInsight Manager的管理员。 主组为compcommon，不具备组权限，具备Manager_administrator角色的权限。	LDAP用户不支持登录与认证，无密码修改方法。
	backup		主组为compcommon	
	backup/manager		主组为compcommon	
	oms		主组为compcommon	
	oms/manager		主组为compcommon	
	clientregister		主组为compcommon	
	zookeeper		主组为hadoop	
	zookeeper/hadoop.*<系统域名>*		主组为hadoop	
	zkcli		主组为hadoop	

类别	默认用户	初始密码	描述	密码修改方法
	zkcli/hadoop.*<系统域名>*		主组为hadoop	
	flume		主组为hadoop	
	flume_server		主组为hadoop	
	hdfs		主组为hadoop	
	hdfs/hadoop.*<系统域名>*		主组为hadoop	
	mapred		主组为hadoop	
	mapred/hadoop.*<系统域名>*		主组为hadoop	
	mr_zk		主组为hadoop	
	mr_zk/hadoop.*<系统域名>*		主组为hadoop	
	hue		主组为supergroup	
	hive		主组为hive	
	hive/hadoop.*<系统域名>*		主组为hive	
	hive1		主组为hive1	
	hive1/hadoop.*<系统域名>*		主组为hive1	
	hive2		主组为hive2	
	hive2/hadoop.*<系统域名>*		主组为hive2	

类别	默认用户	初始密码	描述	密码修改方法
	hive3		主组为hive3	
	hive3/hadoop.*<系统域名>*		主组为hive3	
	hive4		主组为hive4	
	hive4/hadoop.*<系统域名>*		主组为hive4	
	hbase		主组为hadoop	
	hbase/hadoop.*<系统域名>*		主组为hadoop	
	thrift		主组为hadoop	
	thrift/hadoop.*<系统域名>*		主组为hadoop	
	oozie		主组为hadoop	
	hbase/zkclient.*<系统域名>*		主组为hadoop	
	loader		主组为hadoop	
	loader/hadoop.*<系统域名>*		主组为hadoop	
	spark2x		主组为hadoop	
	spark2x/hadoop.*<系统域名>*		主组为hadoop	
	spark_zk		主组为hadoop	

类别	默认用户	初始密码	描述	密码修改方法
	spark2x1		主组为hadoop	
	spark2x1/hadoop.*<系统域名>*		主组为hadoop	
	spark2x2		主组为hadoop	
	spark2x2/hadoop.*<系统域名>*		主组为hadoop	
	spark2x3		主组为hadoop	
	spark2x3/hadoop.*<系统域名>*		主组为hadoop	
	spark2x4		主组为hadoop	
	spark2x4/hadoop.*<系统域名>*		主组为hadoop	
	kafka		主组为kafkaadmin	
	kafka/hadoop.*<系统域名>*		主组为kafkaadmin	
	storm		主组为stormadmin	
	storm/hadoop.*<系统域名>*		主组为stormadmin	
	storm_zk		主组为storm	

类别	默认用户	初始密码	描述	密码修改方法
	storm_zk/hadoop.*<系统域名>*		主组为storm	
	kms/hadoop		主组为kmsadmin	
	knox		主组是compcommon	
	executor		主组是compcommon	
	rangeradmin		主组为supergroup	
	rangeradmin/hadoop.<系统域名>		主组为supergroup	
	rangerusersync		主组为supergroup	
	rangertagsync		主组为supergroup	
	rangerauditor		主组为compcommon	

说明

用户可登录FusionInsight Manager后，选择“系统 > 权限 > 域和互信”，查看“本端域”参数，即为当前系统域名。上表中系统内部用户的用户名所包含的系统域名所有字母为小写。

例如“本端域”参数为“9427068F-6EFA-4833-B43E-60CB641E5B6C.COM”，则HDFS默认启动用户为“hdfs/hadoop.9427068f-6efa-4833-b43e-60cb641e5b6c.com”。

数据库用户

系统数据库用户包含OMS数据库用户、DBService数据库用户。

类别	默认用户	初始密码	描述	密码修改方法
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作。	请参见**修改OMS数据库管理员密码**。

类别	默认用户	初始密码	描述	密码修改方法
	omm	ChangeMe@123456	OMS数据库数据访问用户。	请参见**修改OMS数据库访问用户密码**。
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户。	• MRS 3.1.2之前版本：请参见**修改组件数据库用户密码**。 • MRS 3.1.2及之后版本：暂不支持
	compdbuser	随机密码	MRS 3.1.2及之后版本：DBService组件中GaussDB数据库的管理员用户，用于业务运维场景。该账户密码已失效，首次使用需重置。	请参见**修改DBService数据库compdbuser用户密码**。
	hive	• MRS 3.1.2之前版本：HiveUser@ • MRS 3.1.2及之后版本：随机密码	Hive连接DBService数据库hivemeta的用户。	• MRS 3.1.2之前版本：请参见**修改组件数据库用户密码**。 • MRS 3.1.2及之后版本：请参见**重置组件数据库用户密码**。
	hive1	• MRS 3.1.2之前版本：HiveUser@ • MRS 3.1.2及之后版本：随机密码	Hive1连接DBService数据库hivemeta1的用户。	

类别	默认用户	初始密码	描述	密码修改方法
	hive2	• MRS 3.1.2之前版本：HiveUser@ • MRS 3.1.2及之后版本：随机密码	Hive2连接DBService数据库hivemeta2的用户。	
	hive3	• MRS 3.1.2之前版本：HiveUser@ • MRS 3.1.2及之后版本：随机密码	Hive3连接DBService数据库hivemeta3的用户。	
	hive4	• MRS 3.1.2之前版本：HiveUser@ • MRS 3.1.2及之后版本：随机密码	Hive4连接DBService数据库hivemeta4的用户。	
	hive*N*/*N*	• MRS 3.1.2之前版本：HiveUser@ • MRS 3.1.2及之后版本：随机密码	安装多服务时，Hive-*N*连接DBService数据库hive*N*meta的用户。 例如Hive-1服务连接DBService数据库hive1meta的用户为hive11。	

类别	默认用户	初始密码	描述	密码修改方法
	hue	• MRS 3.1.2之前版本：HueUser@123 • MRS 3.1.2及之后版本：随机密码	Hue连接DBService数据库hue的用户。	
	sqoop	• MRS 3.1.2之前版本：SqoopUser@ • MRS 3.1.2及之后版本：随机密码	Loader连接DBService数据库sqoop的用户。	
	sqoop*N*	• MRS 3.1.2之前版本：SqoopUser@ • MRS 3.1.2及之后版本：随机密码	安装多服务时，Loader-*N*连接DBService数据库sqoop*N*的用户。 例如Loader-1服务连接DBService数据库sqoop1的用户为sqoop1。	
	oozie	• MRS 3.1.2之前版本：OozieUser@ • MRS 3.1.2及之后版本：随机密码	Oozie连接DBService数据库oozie的用户。	

类别	默认用户	初始密码	描述	密码修改方法
	oozie/V	• MRS 3.1.2之前版本：OozieUser@ • MRS 3.1.2及之后版本：随机密码	安装多服务时，Oozie-/V连接DBService数据库oozie/V的用户。 例如Oozie-1服务连接DBService数据库oozie1的用户为oozie1。	
	rangeradmin	• MRS 3.1.2之前版本：OozieUser@ • MRS 3.1.2及之后版本：随机密码	Ranger连接DBservice数据库ranger的用户。	
	kafkaui	随机密码	KafkaUI连接DBService数据库kafka的用户。 该用户仅在MRS 3.1.2及之后版本中存在。	
	flink	随机密码	Flink连接DBService数据库flink的用户。 该用户仅在MRS 3.1.2及之后版本中存在。	

11.12.1.6 默认权限信息一览

角色

默认角色	描述
Manager_administrator	Manager管理员，具有Manager所有权限。 可创建一级租户，可创建、修改新的用户组，指定用户权限，以满足不同用户对系统的管理需求。
Manager_operator	Manager操作员，具有**主页**、**集群**、**主机**、**运维**页签所有权限。

默认角色	描述
Manager_auditor	Manager审计员，具有**审计**页签的所有权限。 可查看和管理Manager系统审计日志的权限。
Manager_viewer	Manager查看员，具有**主页、集群、主机、告警与事件、系统>权限**相关信息的查看权限和下载客户端权限。（下载客户端的权限仅MRS 3.2.0及之后版本支持）
Manager_tenant	Manager租户管理员。 可为当前用户所属于的非叶子租户创建子租户并管理。具有“**运维 > 告警**”页面下“**告警**”、“**事件**”的查看权限。
System_administrator	系统管理员，具有Manager的管理员权限及所有组件服务管理员的权限。
default	为集群default租户创建的默认角色。拥有Yarn组件default队列的管理权限。非首个安装集群的default租户默认角色为“c*<集群ID>*_default”。
Manager_administrator_180	FusionInsight Manager系统管理员组。系统内部角色，仅限组件间内部使用。
Manager_auditor_181	FusionInsight Manager系统审计员组。系统内部角色，仅限组件间内部使用。
Manager_operator_182	FusionInsight Manager系统操作员组。系统内部角色，仅限组件间内部使用。
Manager_viewer_183	FusionInsight Manager系统查看员组。系统内部角色，仅限组件间内部使用。
System_administrator_186	系统管理员组。系统内部角色，仅限组件间内部使用。
Manager_tenant_187	租户系统用户组。系统内部角色，仅限组件间内部使用。
default_1000	为租户创建的用户组。系统内部角色，仅限组件间内部使用。

用户组

类型	默认用户组	描述
集群默认用户组	hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。

类型	默认用户组	描述
	hadoopmanager	将用户加入此用户组，可获得HDFS和Yarn的组件运维管理员权限。对HDFS来说，运维管理员可以访问NameNode WebUI，还能进行手动主备倒换等操作。对Yarn来说，运维管理员可以执行Yarn集群的管理操作，例如访问ResourceManager WebUI，管理NodeManager节点，刷新队列，设置NodeLabel等，但不能提交任务。
	hetuadmin	HetuEngine管理员用户组，属于该组的用户拥有在HSConsole页面操作的权限。
	hive	普通用户组。Hive用户必须属于该用户组。
	hive1	普通用户组。Hive1用户必须属于该用户组。
	hive2	普通用户组。Hive2用户必须属于该用户组。
	hive3	普通用户组。Hive3用户必须属于该用户组。
	hive4	普通用户组。Hive4用户必须属于该用户组。
	kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限，才能访问对应Topic。
	kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
	kafkasuperuser	Kafka的Topic读写用户组。添加入本组的用户，拥有所有Topic的读写权限。
	storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。
	stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。
	supergroup	这个用户组内的用户具有HBase，HDFS和Yarn的管理员权限，并且可以使用Hive。
	yarnviewgroup	Yarn任务只读用户组。将用户加入此用户组，可获得Yarn和Mapreduce界面上任务的只读权限。
	check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
	compcommon	系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
操作系统默认用户组	wheel	系统内部运行用户“omm”的主组。
	ficommon	系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

说明

如果当前集群不是在FusionInsight Manager内第一次安装的集群，集群内除Manager以外其他组件对应的默认用户组名称为“c*<集群ID>_默认用户组名*”，例如“c2_hadoop”。

用户

请参见**用户账号一览表**。

服务相关用户安全参数

- **HDFS**

  参数“dfs.permissions.superusergroup”表示HDFS最高权限管理员组，默认值为“supergroup”。

- **Spark2x以及对应多实例**

  参数“spark.admin.acls”表示Spark2x的管理员列表，列表中成员有权限管理所有Spark任务，若用户未加入此列表则无法管理所有Spark任务。默认值为“admin”。

11.12.1.7 FusionInsight Manager 安全功能

通过FusionInsight Manager的以下模块，可以方便的完成用户权限数据的查看和设置。

- 用户管理：提供用户的增、删、改、查基本功能，提供用户绑定用户组和角色的功能。

  具体请参见**用户管理**。

- 用户组管理：提供用户组的增、删、改、查基本功能，提供用户组绑定角色的功能。

  具体请参见**用户组管理**。

- 角色管理：提供角色的增、删、改、查基本功能，提供角色绑定某个或者多个组件的资源访问权限的功能。

  具体请参见**角色管理**。

- 租户管理：提供租户的增、删、改、查基本功能以及租户与组件资源的绑定关系。FusionInsight为了便于管理，为每个租户都会默认产生一个角色。如果定义租户拥有某些资源的权限，则租户对应的角色就拥有这些资源的权限。

  具体请参见**租户资源**。

11.12.2 账户管理

11.12.2.1 账户安全设置

11.12.2.1.1 解锁 LDAP 用户和管理账户

操作场景

管理员在LDAP用户和管理账户被锁定时，需要在管理节点解锁集群LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”以及LDAP管理账户

“cn=krbkdc,ou=Users,dc=hadoop,dc=com”和
“cn=krbadmin,ou=Users,dc=hadoop,dc=com”。

说明

Ldap用户或管理账户连续使用错误密码操作Ldap次数大于5次时，会造成LDAP用户或管理账户被锁定。用户被锁定之后，5分钟后会自动解锁。

操作步骤

步骤1 以**omm**用户登录主管理节点。

步骤2 执行以下命令，切换到目录：

cd ${BIGDATA_HOME}/om-server/om/ldapserver/ldapserver/local/script

步骤3 执行以下命令，解锁LDAP用户或管理账户：

./ldapserver_unlockUsers.sh *USER_NAME*

其中，*USER_NAME*表示将要解锁的用户名称。

例如，解锁LDAP管理账户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”的方法如下：

./ldapserver_unlockUsers.sh krbkdc

运行脚本之后，在ROOT_DN_PASSWORD之后输入krbkdc用户密码，显示如下结果，说明解锁成功：

```
Unlock user krbkdc successfully.
```

----结束

11.12.2.1.2 解锁系统内部用户

操作场景

若服务出现异常状态，有可能是系统内部用户被锁定，请及时解锁，否则会影响集群正常运行。系统内部用户列表请参见**用户账号一览表**。系统内部用户无法使用FusionInsight Manager解锁。

前提条件

根据**用户账号一览表**获取LDAP管理员“cn=root,dc=hadoop,dc=com”的默认密码。

操作步骤

步骤1 使用以下方法确认系统内部用户是否被锁定：

1. 查询oldap端口：
   a. 登录FusionInsight Manager，选择“系统 > OMS > oldap > 修改配置”。
   b. “Ldap服务监听端口”参数值即为oldap端口。
2. 查询域名方法：
   a. 登录FusionInsight Manager，选择“系统 > 权限 > 域和互信”。
   b. “本端域”参数即为域名。

例如当前系统域名为“9427068F-6EFA-4833-B43E-60CB641E5B6C.COM”。

3. 在集群内节点上以**omm**用户执行以下命令查询密码认证失败次数：

   ldapsearch -H ldaps://*OMS浮动IP地址*:*OLdap端口* **-LLL -x -D cn=root,dc=hadoop,dc=com -b krbPrincipalName=***系统内部用户名*@*当前域名*,cn=*当前域名*,cn=krbcontainer,dc=hadoop,dc=com -w** *LDAP管理员密码* **-e ppolicy | grep krbLoginFailedCount**

   例如，查看oms/manager用户认证失败次数：

   ldapsearch -H ldaps://10.5.146.118:21750 -LLL -x -D cn=root,dc=hadoop,dc=com -b krbPrincipalName=oms/manager@9427068F-6EFA-4833-B43E-60CB641E5B6C.COM,cn=9427068F-6EFA-4833-B43E-60CB641E5B6C.COM,cn=krbcontainer,dc=hadoop,dc=com -w *cn=root,dc=hadoop,dc=com用户密码* **-e ppolicy | grep krbLoginFailedCount**

   ```
   krbLoginFailedCount: 5
   ```

4. 登录FusionInsight Manager，选择“系统 > 权限 > 安全策略 > 密码策略”。
5. 查看“密码连续错误次数”参数值，若小于等于“krbLoginFailedCount”参数值，则用户已被锁定。

   说明

   查看运行日志，也可以确认系统内部用户是否被锁定。

步骤2 以**omm**用户登录主管理节点，执行以下命令解锁。

sh ${BIGDATA_HOME}/om-server/om/share/om/acs/config/unlockuser.sh --userName *系统内部用户名*

例如，**sh ${BIGDATA_HOME}/om-server/om/share/om/acs/config/unlockuser.sh --userName oms/manager**

----结束

11.12.2.1.3 修改集群组件鉴权配置开关

操作场景

集群部署为安全模式或者普通模式时，HDFS和ZooKeeper默认会对访问服务的用户进行鉴权，没有权限的用户无法访问HDFS和ZooKeeper中的资源。集群部署为普通模式时，HBase和Yarn默认不会对访问用户进行鉴权，所有用户可以访问HBase和Yarn中的资源。

管理员可以根据业务实际需要，在普通模式集群中配置开启HBase和Yarn鉴权，或关闭HDFS和ZooKeeper鉴权。

对系统的影响

修改开关后服务的配置将过期，需要重启对应的服务使配置生效。

开启 HBase 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“hbase.coprocessor.region.classes”、“hbase.coprocessor.master.classes”和“hbase.coprocessor.regionserver.classes”。

将协处理器参数“org.apache.hadoop.hbase.security.access.AccessController”添加到以上参数原有参数值末尾，使用英文逗号与原有协处理器分隔。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

关闭 HBase 鉴权

说明

关闭HBase鉴权后，原有的权限数据会继续保留。如果需要删除权限信息，请在关闭鉴权后，进入hbase shell删除表hbase:acl。

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HBase > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“hbase.coprocessor.region.classes”、“hbase.coprocessor.master.classes”和“hbase.coprocessor.regionserver.classes”。

将协处理器参数“org.apache.hadoop.hbase.security.access.AccessController”去除。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

关闭 HDFS 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“dfs.namenode.acls.enabled”和“dfs.permissions.enabled”。

- “dfs.namenode.acls.enabled”表示是否启用HDFS ACL，默认为“true”启用ACL，请修改为“false”。
- “dfs.permissions.enabled”表示是否为HDFS启用权限检查，默认为“true”启用权限检查，请修改为“false”。修改后HDFS中的目录和文件的属主、属组以及权限信息保持不变。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

开启 Yarn 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > Yarn > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“yarn.acl.enable”。

“yarn.acl.enable”表示是否为Yarn启用权限检查。

- 普通模式下默认为“false”不启用权限检查，如果要启用，请修改为“true”。
- 安全模式下默认为“true”，表示开启鉴权。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

关闭 ZooKeeper 鉴权

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > ZooKeeper > 配置”。

步骤3 单击“全部配置”。

步骤4 搜索参数“skipACL”。

“skipACL”表示是否跳过ZooKeeper权限检查，默认为“no”启用权限检查，请修改为“yes”。

步骤5 单击“保存”，单击“确定”。

等待界面提示操作完成。

----结束

11.12.2.1.4 使用普通模式集群用户在非集群节点登录

操作场景

集群安装为普通模式时，各组件客户端不支持安全认证且无法使用**kinit**命令，所以集群外的节点默认无法使用集群中的用户，可能导致在这些节点访问某个组件服务端时用户鉴权失败。

如果需要在集群外节点以组件用户身份访问集群资源，管理员需为集群外节点设置同名用户可通过SSH协议登录节点的功能，并以登录操作系统用户身份连接集群各组件服务端。

前提条件

- 集群外的节点需要与集群的业务平面是连通的。
- 集群的KrbServer服务运行状态正常。
- 获取集群外的节点**root**用户密码。
- 集群已规划并添加“人机”用户，并获取认证凭据文件。请参见**创建用户**和**导出认证凭据文件**。

操作步骤

步骤1 以**root**用户登录到需要添加用户的节点。

步骤2 执行以下命令：

rpm -qa | grep pam和**rpm -qa| grep krb5-client**

界面一共显示以下rpm包：

```
pam_krb5-32bit-2.3.1-47.12.1
pam-modules-32bit-11-1.22.1
yast2-pam-2.17.3-0.5.211
pam-32bit-1.1.5-0.10.17
pam_mount-32bit-0.47-13.16.1
pam-config-0.79-2.5.58
pam_krb5-2.3.1-47.12.1
pam-doc-1.1.5-0.10.17
pam-modules-11-1.22.1
pam_mount-0.47-13.16.1
pam_ldap-184-147.20
pam-1.1.5-0.10.17
krb5-client-1.6.3
```

步骤3 检查操作系统实际是否已安装清单中的rpm包？

- 是，执行**步骤5**。
- 否，执行**步骤4**。

步骤4 从操作系统镜像中获取缺少的rpm包，并上传文件到当前目录，然后执行以下命令安装rpm包：

rpm -ivh *.rpm

说明

安装的RPM包可能带来安全风险，请用户对操作系统进行加固时考虑安装这些RPM包所带来的风险。

安装完成后执行**步骤5**。

步骤5 执行以下命令，配置pam使用Kerberos认证。

pam-config --add --krb5

说明

如果需要在非集群节点取消Kerberos认证与系统用户登录，以“root”用户执行**pam-config --delete --krb5**命令。

步骤6 解压认证凭据文件得到“krb5.conf”，并使用WinSCP将此配置文件上传到集群外节点的“/etc”目录，执行以下命令设置权限使其他用户可以访问，例如“604”：

chmod 604 /etc/krb5.conf

步骤7 以**root**用户继续在连接会话中执行以下命令为“人机”用户添加对应的操作系统用户，并指定用户主组为“root”。

此操作系统用户密码与在Manager创建“人机”用户时设置的初始密码相同。

useradd *用户名* **-m -d /home/admin_test -g root -s /bin/bash**

例如，“人机”用户名为“admin_test”，执行以下命令：

useradd admin_test -m -d /home/admin_test -g root -s /bin/bash

说明

第一次使用新添加的操作系统用户通过SSH协议登录节点时，首次输入用户密码系统提示密码过期，第二次输入用户密码后系统提示修改密码。请输入一个同时满足节点操作系统及集群密码复杂度的新密码。

----结束

11.12.2.2 修改系统用户密码

11.12.2.2.1 修改 admin 密码

操作场景

“admin”是FusionInsight Manager的系统管理员账号，建议用户通过FusionInsight Manager定期修改密码，提高系统安全性。

操作步骤

步骤1 登录FusionInsight Manager。

需使用“admin”登录。

步骤2 移动鼠标到界面右上角的“Hello, admin”。

在弹出菜单中单击“修改密码”。

步骤3 分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

默认密码复杂度要求：

- 密码字符长度为8～64位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为~`!?,.;-_'(){}[]/<>@#$%^&*+|\=。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码。
- 不可与最近N次使用过的密码相同，N为**密码策略配置**中“重复使用规则”的值。

----结束

11.12.2.2.2 修改操作系统用户密码

操作场景

安装FusionInsight Manager时系统自动在集群每个节点上创建用户“omm”和“ommdba”，建议管理员定期修改集群节点操作系统用户“omm”、“ommdba”的登录密码，以提升系统运维安全性。

各节点“omm”、“ommdba”无需设置为统一的密码。

前提条件

- 获取待修改密码“omm”、“ommdba”用户对应节点的IP地址。
- 修改**omm**和**ommdba**用户需要获取**root**用户密码。

修改操作系统用户密码

步骤1 以**root**登录待修改密码节点。

步骤2 执行如下命令，修改用户密码。

passwd *ommdba*

Red Hat系统显示：

```
Changing password for user ommdba.
New password:
```

步骤3 输入用户的新密码。操作系统的密码修改策略由用户实际使用的操作系统类型决定。

```
Retype New Password:
Password changed.
```

----结束

11.12.2.3 修改系统内部用户密码

11.12.2.3.1 修改 Kerberos 管理员密码

操作场景

管理员应定期修改Kerberos管理员“kadmin”的密码，以提升系统运维安全性。

修改此用户密码将同步修改OMS Kerberos管理员密码。

前提条件

已在集群内的任一节点安装了客户端，并获取此节点IP地址。

操作步骤

步骤1 以**root**用户通过节点IP地址登录安装了客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/hadoopclient”。

cd /opt/hadoopclient

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，修改kadmin/admin密码。此操作对所有服务器生效。如果密码丢失将不可找回，请妥善保管。

kpasswd kadmin/admin

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为~`!?,.;-_'(){}[]/<>@#$%^&*+|\=。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码，例如Admin@12345。
- 不可与最近N次使用过的密码相同，N为**密码策略配置**中“重复使用规则”的值。

----结束

11.12.2.3.2 修改 OMS Kerberos 管理员密码

操作场景

管理员应定期修改OMS Kerberos管理员“kadmin”的密码，以提升系统运维安全性。

修改此用户密码将同步修改Kerberos管理员密码。

操作步骤

步骤1 以**omm**用户登录任意管理节点。

步骤2 执行以下命令，切换到目录。

cd ${BIGDATA_HOME}/om-server/om/meta-0.0.1-SNAPSHOT/kerberos/scripts

步骤3 执行以下命令，配置环境变量。

source component_env

步骤4 执行以下命令，修改kadmin/admin密码。此操作对所有服务器生效。如果密码丢失将不可找回，请妥善保管。

kpasswd kadmin/admin

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!?,.;-_'(){}[]/<>@#$%^&*+|\=中的4种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码，例如Admin@12345。
- 不可与最近N次使用过的密码相同，N为**密码策略配置**中“重复使用规则”的值。

----结束

11.12.2.3.3 修改 LDAP 管理员和 LDAP 用户密码（含 OMS LDAP）

说明

该章节仅适用于MRS 3.1.0版本。MRS 3.1.0之后版本请参考**修改OMS服务配置参数**章节。

操作场景

建议管理员定期修改集群的LDAP管理员用户“cn=root,dc=hadoop,dc=com”和LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码，以提升系统运维安全性。

修改上述用户密码将同步修改OMS LDAP管理员或用户密码。

说明

旧版本集群升级到新版本后，LDAP管理员密码将继承旧集群的密码策略，为保证系统安全，建议集群升级后及时修改密码。

对系统的影响

- 修改LdapServer服务的用户密码为高危操作，需要重启KrbServer和LdapServer服务。重启KrbServer可能会导致集群中的节点短时间内出现执行**id**命令查询不到用户的现象，请谨慎执行。
- 修改LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码后，可能会导致该用户在组件LDAP上被锁定。因此，建议修改密码后对该用户进行解锁，解锁方法请参见**解锁LDAP用户和管理账户**章节。

前提条件

修改LDAP用户“cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码前需先确认该用户没有被锁定，在集群主管理节点上执行如下命令：

说明

oldap端口查询方法：

1. 登录FusionInsight Manager，选择“系统 > OMS > oldap > 修改配置”；
2. “Ldap服务监听端口”参数值即为oldap端口。

ldapsearch -H ldaps://*OMS浮动地址:OLdap端口* -LLL -x -D cn=pg_search_dn,ou=Users,dc=hadoop,dc=com -W -b cn=pg_search_dn,ou=Users,dc=hadoop,dc=com -e ppolicy

输入LDAP用户pg_search_dn的密码，出现如下提示表示该用户被锁定，则需要解锁用户，具体请参见**解锁LDAP用户和管理账户**。

说明

LDAP用户pg_search_dn的密码为系统随机生成，具体可在主节点的“/etc/sssd/sssd.conf”或“/etc/ldap.conf”文件中获取。

```
ldap_bind: Invalid credentials (49); Account locked
```

操作步骤

步骤1 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > LdapServer”。

步骤2 选择 “更多 > 修改数据库密码”，在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤3 在“修改密码”对话框的“用户信息”中选择需要修改密码的用户名。

步骤4 在“旧密码”输入旧密码，“新密码”和“确认密码”输入新密码。

默认密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符`~!@#$%^&*()-_=+|[{}];,<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与当前密码相同。

步骤5 勾选“我已阅读此信息并了解其影响”，单击“确定”确认修改并重启服务。

----结束

11.12.2.3.4 修改 LDAP 管理账户密码

说明

该章节仅适用于MRS 3.1.0版本。MRS 3.1.0之后版本请参考**修改OMS服务配置参数**章节。

操作场景

建议管理员定期修改集群LDAP管理账户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”和“cn=krbadmin,ou=Users,dc=hadoop,dc=com”的密码，以提升系统运维安全性。

对系统的影响

- 修改密码后需要重启KrbServer服务。
- 修改密码后需要确认LDAP管理账户“cn=krbkdc,ou=Users,dc=hadoop,dc=com”和“cn=krbadmin,ou=Users,dc=hadoop,dc=com”是否被锁定，在集群主管理节点上执行如果下命令查看krbkdc是否被锁定（krbadmin用户方法类似）：

  说明

  oldap端口查询方法：

  1. 登录FusionInsight Manager，选择“系统 > OMS > oldap > 修改配置”；
  2. “Ldap服务监听端口”参数值即为oldap端口。

  ldapsearch -H ldaps://*OMS_FLOAT_ IP地址:OLdap端口* -LLL -x -D cn=krbkdc,ou=Users,dc=hadoop,dc=com -W -b cn=krbkdc,ou=Users,dc=hadoop,dc=com -e ppolicy

  输入LDAP管理账户krbkdc的密码，出现如下提示表示该用户被锁定，则需要解锁用户，具体请参见**解锁LDAP用户和管理账户**。

  ```
  ldap_bind: Invalid credentials (49); Account locked
  ```

前提条件

已确认主管理节点IP地址。

操作步骤

步骤1 以**omm**用户通过管理节点IP登录主管理节点。

步骤2 执行以下命令，切换到目录。

cd ${BIGDATA_HOME}/om-server/om/meta-0.0.1-SNAPSHOT/kerberos/scripts

步骤3 执行以下命令，修改LDAP管理账户密码。

./okerberos_modpwd.sh

输入旧密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符`~!@#$%^&*()-_=+|[{}];,<.>/?中的3种类型字符。
- 不可与当前密码相同。

显示如下结果，说明修改成功：

```
Modify kerberos server password successfully.
```

步骤4 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > KrbServer > 更多 > 重启服务”。

验证用户身份后不勾选“同时重启上层服务”，单击“确定”重启KrbServer服务。

----结束

11.12.2.3.5 修改组件运行用户密码

操作场景

建议管理员定期修改集群内组件运行用户的密码，以提升系统运维安全性。

组件运行用户，根据初始密码是否是系统随机生成，可分为两类：

- 密码随机生成的，用户类型为“机机”用户。
- 密码不是随机生成的，用户类型为“人机”用户。

对系统的影响

初始密码为系统随机生成的组件运行用户，在修改密码后需要重启集群，重启期间会造成业务暂时中断。

前提条件

已在集群内的任一节点安装了客户端，并获取此节点IP地址。

操作步骤

步骤1 以客户端安装用户，登录安装了客户端的节点。

步骤2 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，输入kadmin/admin用户密码后进入kadmin控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“Admin@123”，首次登录后会提示该密码过期，请按照提示修改密码。如果密码丢失将不可找回，请妥善保管。

步骤5 执行以下命令，修改系统内部组件运行用户密码。此操作对所有服务器生效。

cpw *系统内部用户名*

例如：**cpw oms/manager**

oms为示例用户，请根据实际用户替换，比如admin。

默认密码复杂度要求：

- 密码字符长度最小为8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符5种类型字符中的4种。支持的特殊字符为~`!?,.;-_'(){}[]/<>@#$%^&*+|\=。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可以为常见的易破解密码，例如Admin@12345。
- 不可与最近N次使用过的密码相同，N为**配置密码策略**中“重复使用规则”的值。此策略只影响“人机”用户。

说明

执行如下命令，可以查看用户的信息。

getprinc *系统内部用户名*

例如：**getprinc oms/manager**

步骤6 确认修改密码的用户，用户类型是哪种?

- 用户类型为“机机”用户，执行**步骤7**。
- 用户类型为“人机”用户，密码修改完成，任务结束。

步骤7 登录FusionInsight Manager。

步骤8 选择“集群 > *待操作的集群名称* > 更多 > 重启”。

步骤9 在弹出窗口中，输入当前登录的用户密码确认身份，单击“确定”。

步骤10 在确认重启的对话框中，单击“确定”。

步骤11 等待界面提示重启成功。

----结束

11.12.2.4 修改默认数据库用户密码

11.12.2.4.1 修改 OMS 数据库管理员密码

操作场景

建议管理员定期修改OMS数据库管理员的密码，以提升系统运维安全性。

操作步骤

步骤1 以**root**用户登录主管理节点。

说明

ommdba用户密码不支持在备管理节点修改，否则集群无法正常工作。只需在主管理节点执行修改操作，无需在备管理节点操作。

步骤2 执行以下命令，切换用户。

su - omm

步骤3 执行以下命令，切换目录。

cd $OMS_RUN_PATH/tools

步骤4 执行以下命令，修改**ommdba**用户密码。

mod_db_passwd ommdba

步骤5 输入**ommdba**的原密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

显示如下结果，说明修改成功：

```
Congratulations, update [ommdba] password successfully.
```

----结束

11.12.2.4.2 修改 OMS 数据库访问用户密码

操作场景

建议管理员定期修改OMS数据库访问用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启OMS服务，服务在重启时系统无法访问。

操作步骤

步骤1 在FusionInsight Manager选择“系统 > OMS > gaussDB > 修改密码”。

步骤2 在omm用户所在行，单击“操作”列下的“修改密码”。

步骤3 在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤4 根据界面信息，输入新旧密码。

密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。
- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

步骤5 单击“确定”，等待界面提示操作成功。

步骤6 在**omm**用户所在行，单击“操作”列下的“重启OMS服务”。

步骤7 在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤8 在确定重启的对话框中，单击“确定”，重新启动OMS服务。

----结束

11.12.2.4.3 修改组件数据库用户密码

操作场景

建议管理员定期修改组件数据库用户的密码，以提升系统运维安全性。

说明

该章节仅适用于MRS 3.1.0版本。MRS 3.1.0之后版本请参考**重置组件数据库用户密码**章节。

对系统的影响

修改密码需要重启服务，服务在重启时无法访问。

操作步骤

步骤1 在FusionInsight Manager选择“集群 > *待操作的集群名称* > 服务”。

步骤2 单击待重置数据库用户密码的服务名称，然后在服务概览页面中单击“停止服务”。

在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

确认停止服务的影响后，等待服务成功停止。

步骤3 单击待修改数据库用户密码的服务，选择“更多 > 修改数据库密码”，在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

步骤4 根据界面信息，输入新旧密码。

密码复杂度要求：

- 组件数据库用户密码字符长度为8～32。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];",<.>/?中的3种类型字符。

- 不可和用户名相同或用户名的倒序字符相同。
- 不可与前20个历史密码相同。

步骤5 勾选“我已阅读此信息并了解其影响”，单击“确定”。

步骤6 密码修改完成后，选择“更多 > 重启服务”，在弹出窗口中输入当前登录的用户密码，单击“确定”，勾选“同时重启上层服务。”，单击“确定”开始重启服务。

----结束

11.12.2.4.4 重置组件数据库用户密码

操作场景

MRS集群中组件连接DBService数据库的用户默认密码为随机密码，建议管理员定期重置组件数据库用户的密码，以提升系统运维安全性。

说明

该章节仅适用于MRS 3.1.2及之后版本。MRS 3.1.2之前版本请参考**修改组件数据库用户密码**章节。

对系统的影响

重置密码需要停止并重启服务，服务在此期间无法访问。

操作步骤

步骤1 在FusionInsight Manager选择“集群 > *待操作的集群名称* > 服务”。

步骤2 单击待重置数据库用户密码的服务名称，例如“Kafka”，然后在服务概览页面中单击“停止服务”。

在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

确认停止服务的影响后，等待服务成功停止。

步骤3 在服务概览页面选择“更多 > 重置数据库密码”。

在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”。

继续勾选“我已阅读此信息并了解其影响。”，单击“确定”。

步骤4 密码重置完成后，在服务概览页面中单击“启动服务”。

步骤5 在弹出窗口单击“确定”，等待服务启动完成。

----结束

11.12.2.4.5 修改 DBService 数据库 compdbuser 用户密码

操作场景

建议管理员定期修改OMS数据库管理员的密码，以提升系统运维安全性。

操作步骤

步骤1 登录FusionInsight Manager界面，选择“集群 > 服务 > DBService > 实例”，查看DBService主节点IP地址信息。

步骤2 以**root**用户登录DBService主节点。

说明

compuserdb用户密码不支持在备DBService节点修改。只需在主管理节点执行修改操作，无需在备管理节点操作。

步骤3 切换到“$DBSERVER_HOME”目录，执行以下命令配置环境变量。

su - omm

cd $DBSERVER_HOME

source .dbservice_profile

步骤4 执行如下命令，使用DBService数据库的**omm**用户修改compdbuser用户密码。

gsql -U omm -W *DBService数据库的omm用户密码* **-d postgres -p 20051 -c "alter user compdbuser identified by '***新密码***' valid until '***超期时间***';"**

说明

- 新密码复杂度要求：
  - 密码字符长度为16～32位。
  - 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
  - 不可和用户名相同或用户名的倒序字符相同。
  - 不可与前20个历史密码相同。
- 超期时间格式为xxxx-xx-xx，例如：2020-10-31。

显示如下结果，则修改成功：

```
ALTER ROLE
```

----结束

11.12.3 安全加固

11.12.3.1 加固策略

加固 Tomcat

在FusionInsight Manager软件安装及使用过程中，针对Tomcat基于开源做了如下功能增强：

- 升级Tomcat版本为官方稳定版本。
- 设置应用程序之下的目录权限为500，对部分目录支持写权限。
- 系统软件安装完成后自动清除Tomcat安装包。
- 应用程序目录下针对工程禁用自动部署功能，只部署了web、cas和client三个工程。

- 禁用部分未使用的http方法，防止被他人利用攻击。
- 更改Tomcat服务器默认shutdown端口号和命令，避免被黑客捕获利用关闭服务器，降低对服务器和应用的威胁。
- 出于安全考虑，更改“maxHttpHeaderSize”的取值，给服务器管理员更大的可控性，以控制客户端不正常的请求行为。
- 安装Tomcat后，修改Tomcat版本描述文件。
- 为了避免暴露Tomcat自身的信息，更改Connector的Server属性值，使攻击者不易获知服务器的相关信息。
- 控制Tomcat自身配置文件、可执行文件、日志目录、临时目录等文件和目录的权限。
- 关闭会话facade回收重用功能，避免请求泄漏风险。
- CookieProcessor使用LegacyCookieProcessor，避免cookie中的敏感数据泄漏。

加固 LDAP

在安装完集群后，针对LDAP做了如下功能增强：

- LDAP配置文件中管理员密码使用SHA加密，当升级openldap版本为2.4.39或更高时，主备LDAP节点服务自动采用SASL External机制进行数据同步，避免密码信息被非法获取。
- 集群中的LDAP服务默认支持SSLv3协议，可安全使用。当升级openldap版本为2.4.39或更高时，LDAP将自动使用TLS1.0以上的协议通讯，避免未知的安全风险。

加固 JDK

- 如果客户端程序使用了AES256加密算法，则需要对JDK进行安全加固，具体操作如下：

  获取与JDK版本对应的JCE（Java Cryptography Extension）文件。JCE文件解压后包含“local_policy.jar”和“US_export_policy.jar”。拷贝此jar包到如下路径并替换文件：

  - Linux：“JDK安装目录/jre/lib/security”
  - Windows：“JDK安装目录\jre\lib\security”

  说明

  请访问Open JDK开源社区获取JCE文件。

- 如果客户端程序需要支持SM4加密算法，则需要更新jar包：

  在“客户端安装目录/JDK/jdk/jre/lib/ext/”目录下获取“SMS4JA.jar”，并拷贝到如下目录：

  - Linux：“JDK安装目录/jre/lib/ext/”
  - Windows：“JDK安装目录\jre\lib\ext\”

11.12.3.2 配置受信任 IP 访问 LDAP

操作场景

默认情况下，部署在OMS和集群中的LDAP服务允许任意IP访问。如果需要只允许受信任的IP地址访问LDAP服务，可以配置iptables过滤列表的INPUT策略。

对系统的影响

配置受信任IP访问LDAP以后，未配置的IP无法访问LDAP。扩容前，新增加的IP需要配置为受信任的IP。

前提条件

- 根据安装规划，收集集群内全部节点的管理平面IP、业务平面IP和所有浮动IP。
- 获取集群内节点的root用户和密码。

操作步骤

配置OMS LDAP信任的IP地址

步骤1 确定管理节点IP地址，请参见**登录管理节点**。

步骤2 登录FusionInsight Manager，请参见**登录管理系统**。

步骤3 选择“系统 > OMS”，在“服务”选择“oldap > 修改配置”，查看OMS LDAP端口号，即“Ldap服务监听端口”参数值。默认为“21750”。

步骤4 以**root**用户通过主管理节点的IP地址登录主管理节点。

步骤5 执行以下命令，查看iptables过滤列表中INPUT策略。

iptables -L

例如未配置任何规则时，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
```

步骤6 执行以下命令，将集群使用的所有IP地址配置为受信任的IP。每个IP需要添加一次。

iptables -A INPUT -s *受信任IP地址* **-p tcp --dport** *端口号* **-j ACCEPT**

例如，将10.0.0.1配置为受信任的IP，可以访问端口21750，执行：

iptables -A INPUT -s 10.0.0.1 -p tcp --dport 21750 -j ACCEPT

步骤7 执行以下命令，将全部IP地址配置为不受信任的IP。已配置为信任IP不受此规则影响。

iptables -A INPUT -p tcp --dport *端口号* **-j DROP**

例如，配置全部IP不能访问端口21750，执行：

iptables -A INPUT -p tcp --dport 21750 -j DROP

步骤8 执行以下命令，查看iptables过滤列表中修改后INPUT策略。

iptables -L

例如配置一个受信任IP后，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
ACCEPT     tcp  --  10.0.0.1             anywhere            tcp dpt:21750
DROP       tcp  --  anywhere             anywhere            tcp dpt:21750
```

步骤9 执行以下命令，查看iptables过滤列表中存在的规则及相对应的编号。

iptables -L -n --line-number

```
Chain INPUT (policy ACCEPT)
num target     prot opt source               destination
1   DROP       tcp  --  0.0.0.0/0            0.0.0.0/0           tcp dpt:21750
```

步骤10 根据实际需求，可执行以下命令，删除iptables过滤列表中的规则。

iptables -D INPUT *待删除的编号*

例如，删除编号为1的规则，执行：

iptables -D INPUT 1

步骤11 以**root**用户通过备管理节点的IP地址登录备管理节点，并重复**步骤5**到**步骤10**。

配置集群LDAP信任的IP地址

步骤12 登录FusionInsight Manager。

步骤13 选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 实例”，查看LDAP服务对应的节点。

步骤14 切换到“配置”，查看集群LDAP端口号，即“LDAP_SERVER_PORT”参数值。默认为“21780”。

步骤15 以**root**用户通过LDAP服务的IP地址登录LDAP节点。

步骤16 执行以下命令，查看iptables过滤列表中INPUT策略。

iptables -L

例如未配置任何规则时，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
```

步骤17 执行以下命令，将集群使用的所有IP地址配置为受信任的IP。每个IP需要添加一次。

iptables -A INPUT -s *受信任IP地址* **-p tcp --dport** *端口号* **-j ACCEPT**

例如，将10.0.0.1配置为受信任的IP，可以访问端口21780，执行：

iptables -A INPUT -s 10.0.0.1 -p tcp --dport 21780 -j ACCEPT

步骤18 执行以下命令，将全部IP地址配置为不受信任的IP。已配置为信任IP不受此规则影响。

iptables -A INPUT -p tcp --dport *端口号* **-j DROP**

例如，配置全部IP不能访问端口21780，执行：

iptables -A INPUT -p tcp --dport 21780 -j DROP

步骤19 执行以下命令，查看iptables过滤列表中修改后INPUT策略。

iptables -L

例如配置一个受信任IP后，INPUT策略显示如下：

```
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
ACCEPT     tcp  --  10.0.0.1            anywhere            tcp dpt:21780
DROP       tcp  --  anywhere            anywhere            tcp dpt:21780
```

步骤20 执行以下命令，查看iptables过滤列表中存在的规则及相对应的编号。

iptables -L -n --line-number

```
Chain INPUT (policy ACCEPT)
num target     prot opt source               destination
1   DROP       tcp  --  0.0.0.0/0            0.0.0.0/0           tcp dpt:21780
```

步骤21 根据实际需求，可执行以下命令，删除iptables过滤列表中的规则。

iptables -D INPUT *待删除的编号*

例如，删除编号为1的规则，执行：

iptables -D INPUT 1

步骤22 以**root**用户通过另一个LDAP服务的IP地址登录LDAP节点，并重复**步骤16**到**步骤21**。

----结束

11.12.3.3 加密 HFile 和 WAL 内容

加密 HFile 和 WAL 内容

须知

- 设置HFile和WAL为SMS4加密或AES加密方式对系统的影响较大，一旦操作失误会导致数据丢失。不推荐使用此功能。
- 使用Bulkload批量导入的数据不支持加密。

缺省情况下，HBase中的HFile和WAL（Write ahead log）内容是不加密的。如果用户需要对其进行加密，可通过如下操作进行配置。

步骤1 在任一安装HBase服务节点，使用**omm**用户执行如下命令创建密钥。

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh *<path>/hbase.jks <type> <length> <alias>*

- */<path>/hbase.jks*表示生成的jks文件存储路径。
- <type>表示加密的类型，支持SMS4或AES。
- <length>表示密钥的长度，SMS4支持16位长度，AES支持128位长度。
- *<alias>*为密钥文件的别名，第一次生成时请使用缺省值“omm”。

例如，生成SMS4加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks SMS4 16 omm

生成AES加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks AES 128 omm

说明

- 集群的操作用户需要有*<path>/hbase.jks*目录的“rw”权限，且要求目录已存在。
- 运行命令后需要再输入4遍相同的*<password>*。其中**步骤3**中进行加密的密码与此步骤的密码相同。

步骤2 将生成的密钥文件分发到集群中所有节点的相同目录下，并为**omm**用户配置该文件的读写权限。

说明

- 请管理员根据企业安全要求，选择安全的操作步骤分发密钥。
- 如果在使用过程中，有节点出现密钥文件丢失的情况，请按照此步骤从其他节点拷贝到该节点。

步骤3 在FusionInsight Manager界面中，设置“hbase.crypto.keyprovider.parameters.encryptedtext”参数的值为密文密码，设置“hbase.crypto.keyprovider.parameters.uri”参数的值为密钥路径和名称。

- “hbase.crypto.keyprovider.parameters.uri”格式为：**jceks://***<key_Path_Name>*。

  *<key_Path_Name>*填写密钥的存储路径，例如“/home/hbase/conf/hbase.jks”则对应参数值为“jceks:///home/hbase/conf/hbase.jks”。
- “hbase.crypto.keyprovider.parameters.encryptedtext”格式为：*<encrypted_password>*。

  *<encrypted_password>*填写创建密钥时的密文密码，参数值显示为密文。使用**omm**用户在安装HBase服务的节点，执行如下命令获取对应加密后的密码：

  sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh

  说明

  运行命令后需要输入<password>。该密码与**步骤1**中手动输入的密码相同。

步骤4 在FusionInsight Manager界面中，设置“hbase.crypto.key.algorithm”参数值为“SMS4”或“AES”，使HFile的内容采用SMS4或AES的方式加密。

步骤5 在FusionInsight Manager界面中，设置“hbase.crypto.wal.algorithm”参数值为“SMS4”或“AES”，使WAL的内容采用SMS4或AES的方式加密。

步骤6 在FusionInsight Manager界面中，将“hbase.regionserver.wal.encryption”参数值修改为“true”。

步骤7 保存设置，并重启HBase服务使其生效。

步骤8 在创建HBase表时，需要通过设置加密方式开启加密功能，<type>表示加密的类型，“d”为加密列族。

- 通过命令行创建表时，直接为列族设置加密方式为SMS4或AES。

  create *'<table name>'*, {*NAME => 'd'*, ***ENCRYPTION => '***<type>*'***}
- 使用代码创建表时，在代码中添加如下信息设置加密方式为SMS4或AES。

```
public void testCreateTable()
{
   String tableName = "user";
   Configuration conf = getConfiguration();
   HTableDescriptor htd = new HTableDescriptor(TableName.valueOf(tableName));

   HColumnDescriptor hcd = new HColumnDescriptor("d");
```

```
    //设置加密方式为SMS4或AES。
    hcd.setEncryptionType("<type>");
    htd.addFamily(hcd);

    HBaseAdmin admin = null;
    try
    {
        admin = new HBaseAdmin(conf);

        if(!admin.tableExists(tableName))
        {
            admin.createTable(htd);
        }
    }
    catch (IOException e)
    {
        e.printStackTrace();
    }
    finally
    {
        if(admin != null)
        {
            try
            {
                admin.close();
            }
            catch (IOException e)
            {
                e.printStackTrace();
            }
        }
    }
}
```

步骤9 如果用户已按照**步骤1**到**步骤7**配置了SMS4或AES加密，但是在执行**步骤8**创建表时，未设置对应的加密参数，使得插入的数据未经加密。

此时，您可以执行如下步骤对之前插入的数据进行加密。

1. 针对表执行flush动作，将内存中的数据导入到HFile中。

   flush *'<table_name>'*

2. 执行以下步骤修改表属性。

   disable *'<table_name>'*

   alter *'<table_name>'*,**NAME=>** *'<column_name>'*,**ENCRYPTION =>** *'<type>'*

   enable *'<table_name>'*

3. 插入一条新的数据，然后flush表。

   说明

   必须要插入一条新的数据，HFile才会生成新的HFile，使得之前插入的未加密数据进行重写加密。

   put *'<table_name>'*, **'id2','f1:c1','value22222222222222222222222222222222'**

   flush *'<table_name>'*

4. 执行如下步骤重写HFile。

   major_compact'*<table_name>'*

须知

执行此步骤时，HBase表是被禁用的，不能提供表服务。请谨慎使用。

----结束

修改密钥文件

须知

修改密钥文件的操作对系统影响较大，一旦操作失误会导致数据丢失。不推荐使用此功能。

在**加密HFile和WAL内容**操作中需要生成对应的密钥文件并设置密码，为确保系统安全，在运行一段时间后，用户可修改密钥，使用新的密钥文件对HFile和WAL内容进行加密。

步骤1 使用**omm**用户执行如下命令生成新的密钥文件。

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh *<path>/hbase.jks <type> <length> <alias-new>*

- *<path>/hbase.jks*:表示生成的hbase.jks文件的存储路径。该路径和文件名称需与**加密HFile和WAL内容**章节生成的密钥文件相同。
- *<alias-new>*：表示密钥文件的别名，请使用与旧密钥文件不同的名字。
- *<type>*表示加密的类型，支持SMS4或AES。
- *<length>*表示密钥的长度，SMS4支持16位长度，AES支持128位长度。

例如，生成SMS4加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks SMS4 16 omm_new

生成AES加密的密钥执行：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks AES 128 omm_new

说明

- 集群的操作用户需要有*<path>/hbase.jks*目录的“rw”权限，且要求目录已存在。
- 运行命令后需要再输入3遍相同的*<password>*，该密码表示密钥文件的密码，请直接使用旧文件的密码，不会产生安全风险。

步骤2 将生成的密钥文件分发到集群中所有节点的相同目录下，并为**omm**用户配置该文件的读写权限。

说明

请管理员根据企业安全要求，选择安全的操作步骤分发密钥。

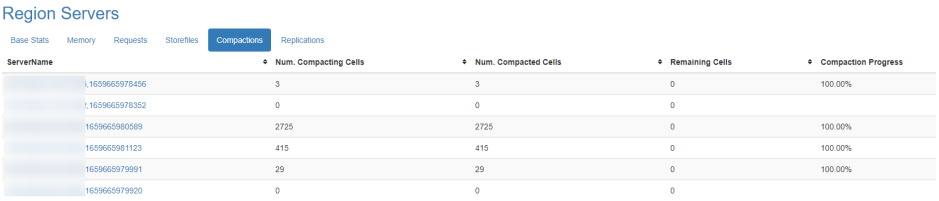
步骤3 在FusionInsight Manager的HBase服务配置界面中增加自定义配置项，设置“hbase.crypto.master.key.name”为“omm_new”，设置“hbase.crypto.master.alternate.key.name”为“omm”，然后保存配置。

步骤4 重启HBase服务，使配置生效。

步骤5 在HBase shell中执行**major compact**命令，生成基于新的加密算法的HFile文件。

major_compact *'<table_name>'*

步骤6 从HMaster的网页中可以查看到major compact进度。



步骤7 所有的“Compaction Progress”都为100%且“Remaining KVs”都为0时，使用**omm**用户执行如下命令销毁旧的密钥文件：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh *<path>/hbase.jks <alias-old>*

- *<path>/hbase.jks*:表示生成的“hbase.jks”文件的存储路径。该路径和文件名称需与**加密HFile和WAL内容**章节生成的密钥文件相同。
- *<alias-old>*：表示要删除的旧密钥文件的别名。

例如：

sh ${BIGDATA_HOME}/FusionInsight_HD_8.1.0.1/install/FusionInsight-HBase-2.2.3/hbase/bin/hbase-encrypt.sh /home/hbase/conf/hbase.jks omm

说明

集群的操作用户需要有*<path>/hbase.jks*目录的“rw”权限，且要求目录已存在。

步骤8 再执行**步骤2**，重新分发更新后的密钥文件。

步骤9 从FusionInsight Manager中删除**步骤3**中新增HBase自定义配置项“hbase.crypto.master.alternate.key.name”。

步骤10 再执行**步骤4**使配置生效。

----结束

11.12.3.4 配置 Hadoop 安全参数

设置安全通道加密

默认情况下，组件间的通道是不加密的。您可以配置如下参数，设置安全通道是加密的。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > *服务名* > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

说明

修改配置参数后需要重启对应服务以使配置生效。

表 11-87 参数说明

服务名称	配置参数	说明	缺省值
HBase	hbase.rpc.protection	设置HBase通道是否加密，包含HBase客户端访问HBase服务端的RPC（remote procedure call）通道，HMaster和RegionServer间的RPC通道。设置为“privacy”表示通道加密，认证、完整性和隐私性功能都全部开启，设置为“integrity”表示不加密，只开启认证和完整性功能，设置为“authentication”表示不加密，仅要求认证报文，不要求完整性和隐私性。 **说明** privacy会对传输内容进行加密，包括用户token等敏感信息，以确保传输信息的安全，但是该方式对性能影响很大，对比另外两种方式，会带来约60%的读写性能下降。请根据企业安全要求修改配置，且客户端与服务端中该配置项需使用相同设置。	-
HDFS	dfs.encrypt.data.transfer	设置客户端访问HDFS的通道和HDFS数据传输通道是否加密。HDFS数据传输通道包括DataNode间的数据传输通道，客户端访问DataNode的DT（Data Transfer）通道。设置为“true”表示加密，默认不加密。	false
HDFS	dfs.encrypt.data.transfer.algorithm	设置客户端访问HDFS的通道和HDFS数据传输通道是否加密。只有在dfs.encrypt.data.transfer配置项设置为true，此参数才会生效。 缺省值为“3des”，表示采用3DES算法进行加密。此处的值还可以设置为“rc4”，避免出现安全隐患，不推荐设置为该值。	3des

服务名称	配置参数	说明	缺省值
HDFS	hadoop.rpc.protection	设置Hadoop中各模块的RPC通道是否加密。包括： ● 客户端访问HDFS的RPC通道。 ● HDFS中各模块间的RPC通道，如DataNode与NameNode间。 ● 客户端访问Yarn的RPC通道。 ● NodeManager和ResourceManager间的RPC通道。 ● Spark访问Yarn，Spark访问HDFS的RPC通道。 ● MapReduce访问Yarn，Mapreduce访问HDFS的RPC通道。 ● HBase访问HDFS的RPC通道。 默认设置为“privacy”表示加密，“authentication”表示不加密。 **说明** 您可以在HDFS组件的配置界面中设置该参数的值，设置后全局生效，即Hadoop中各模块的RPC通道是否加密全部生效。	● 安全模式：privacy ● 普通模式：authentication

Web 最大并发连接数限制

为了保护Web服务器的可靠性，当访问的用户连接数达到一定数量之后，对新增用户的连接进行限制。防止大量同时登录和访问，导致服务不可用，同时避免DDOS攻击。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > *服务名* > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

表 11-88 参数说明

服务名称	配置参数	说明	缺省值
HDFS/Yarn	hadoop.http.server.MaxRequests	设置各组件Web的最大并发连接数限制。	2000
Spark2x	spark.connection.maxRequest	JobHistory允许的最大请求连接数。	5000

11.12.3.5 配置 HBase 允许修改操作的 IP 地址白名单

当HBase集群开启Replication功能时，为了保护主备集群的HBase数据一致性，对备集群HBase增加了数据修改操作的保护。当备集群HBase接收到数据修改操作的RPC请求时，首先检查发出该请求的用户的权限，只有HBase管理用户才有修改权限；其次检查发出该请求的IP的有效性，备集群只接收来自IP白名单中的机器发起的修改请求。IP白名单通过配置项“hbase.replication.allowedIPs”配置。

在FusionInsight Manager系统中，选择“集群 > 服务 > HBase > 配置”，在搜索框中输入参数名称。

表 11-89 参数说明

配置参数	说明	默认值
hbase.replication.allowedIPs	仅允许指定IP地址的复制请求。支持逗号分隔型regex模式。以下模式均支持： ● Regex模式 例如: 10.18.40.* , 10.18.* , 10.18.40.11 ● Range模式（只能指定八位字节的最后一个的范围） 例如: 10.18.40.[10-20] 参数值默认为空，为空时IP白名单为备集群RegionServer的IP，表示只接受来自备集群RegionServer的修改请求。	N/A

11.12.3.6 更新集群密钥

操作场景

在安装集群时，系统将自动生成加密密钥key值以对集群的部分安全信息（例如所有数据库用户密码、密钥文件访问密码等）进行加密存储。在集群安装成功后，如果原始密钥不慎意外泄露或者需要使用新的密钥，系统管理员可以通过以下操作手动更改密钥值。

对系统的影响

- 更新集群密钥后，集群中新增加一个随机生成的新密钥，用于加密解密新保存的数据。旧的密钥不会删除，用于解密旧的加密数据。在修改安全信息后，例如修改数据库用户密码，新密码将使用新的密钥加密。
- 更新集群密钥需要停止集群，集群停止时无法访问。

前提条件

- 已确认主备管理节点IP。请参见**登录管理节点**。
- 停止依赖集群运行的上层业务应用。

操作步骤

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 停止”，输入当前登录的用户密码确认身份。

在确认停止的对话框单击“确定”，等待界面提示停止成功。

步骤3 以**omm**用户登录主管理节点。

步骤4 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤5 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-server/om/tools

步骤6 执行以下命令，更新集群密钥。

sh updateRootKey.sh

根据界面提示，输入**y**：

```
The root key update is a critical operation.
Do you want to continue?(y/n):
```

界面提示以下信息表示更新密钥成功：

```
Step 4-1: The key save path is obtained successfully.
...
Step 4-4: The root key is sent successfully.
```

步骤7 在FusionInsight Manager界面，选择“集群 > *待操作集群的名称* > 启动”。

在弹出窗口中单击“确定”，等待界面提示启动成功。

----结束

11.12.3.7 加固 LDAP

配置 LDAP 防火墙策略

在双平面组网的集群中，由于LDAP部署在业务平面中，为保证LDAP数据安全，建议通过配置整个集群对外的防火墙策略，关闭LDAP相关端口。

步骤1 登录FusionInsight Manager。

步骤2 选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 配置”。

步骤3 查看“LDAP_SERVER_PORT”参数值，即为LdapServer的服务端口。

步骤4 根据客户的实际防火墙环境，配置整个集群对外的防火墙策略，将该端口关闭，以保证数据安全。

----结束

开启 LDAP 审计日志输出

用户可以通过设置LDAP服务的审计日志输出级别，将审计内容输出至系统日志信息中（如“/var/log/messages”），用于查看用户的活动信息及操作指令信息。

说明

LDAP的审计日志开启后，会产生大量日志信息，严重影响集群性能，请谨慎开启。

步骤1 登录任一LdapServer节点。

步骤2 执行以下命令，编辑“slapd.conf.consumer”文件，将“loglevel”的值设置为“256”（loglevel定义可以在OS上使用**man slapd.conf**命令查看）。

cd ${BIGDATA_HOME}/FusionInsight_BASE_8.1.0.1/install/FusionInsight-ldapserver-2.7.0/ldapserver/local/template

vi slapd.conf.consumer

```
...
pidfile         [PID_FILE_SLAPD_PID]
argsfile        [PID_FILE_SLAPD_ARGS]
loglevel    256
...
```

步骤3 登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > LdapServer > 更多 > 重启服务”，验证当前用户身份后重启服务。

----结束

11.12.3.8 配置 Kafka 数据传输加密

操作场景

Kafka客户端和Broker之间的数据传输默认采用明文传输，客户端可能部署在不受信任的网络中，传输的数据可能遭到泄漏和篡改。

操作步骤

默认情况下，组件间的通道是不加密的。用户可以配置如下参数，设置安全通道为加密的。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

说明

配置后应重启对应服务使参数生效。

Kafka服务端的传输加密相关配置参数如**表11-90**所示。

表 11-90 Kafka 服务端传输加密参数

配置项	描述	默认值
ssl.mode.enable	是否开启SSL对应服务。如果设置为“true”，那么Broker启动过程中会启动SSL的相关服务。	false

配置项	描述	默认值
security.inter.broker.protocol	Broker间通信协议。支持PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL这四种协议类型。	SASL_PLAINTEXT

“ssl.mode.enable”配置为“true”后，Broker会开启SSL、SASL_SSL两种协议的服务，然后服务端或者客户端才能配置相关的SSL协议，进行传输加密通信。

11.12.3.9 配置 HDFS 数据传输加密

设置 HDFS 安全通道加密

默认情况下，组件间的通道是不加密的。您可以配置如下参数，设置安全通道为加密的。

参数修改入口：在FusionInsight Manager系统中，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 配置”，展开“全部配置”页签。在搜索框中输入参数名称。

说明

配置后应重启对应服务使参数生效。

表 11-91 参数说明

配置项	描述	默认值
hadoop.rpc.protection	**须知** ● 设置后需要重启服务生效，且不支持滚动重启。 ● 设置后需要重新下载客户端配置，否则HDFS无法提供读写服务。 ● 设置后需要重启executor，否则将导致管控面作业管理及文件管理功能不可用。 设置Hadoop中各模块的RPC通道是否加密。通道包括： ● 客户端访问HDFS的RPC通道。 ● HDFS中各模块间的RPC通道，如DataNode与NameNode间。 ● 客户端访问Yarn的RPC通道 ● NodeManager和ResourceManager间的RPC通道。 ● Spark访问Yarn，Spark访问HDFS的RPC通道。 ● Mapreduce访问Yarn，MapReduce访问HDFS的RPC通道。 ● HBase访问HDFS的RPC通道。 **说明** 设置后全局生效，即Hadoop中各模块的RPC通道的加密属性全部生效。	● 安全模式：privacy ● 普通模式：authentication **说明** ● “authentication”：只进行认证，不加密。 ● “integrity”：进行认证和一致性校验。 ● “privacy”：进行认证、一致性校验、加密。
dfs.encrypt.data.transfer	设置客户端访问HDFS的通道和HDFS数据传输通道是否加密。HDFS数据传输通道包括DataNode间的数据传输通道，客户端访问DataNode的DT（Data Transfer）通道。设置为“true”表示加密，默认不加密。 **说明** ● 仅当hadoop.rpc.protection设置为privacy时使用。 ● 业务数据传输量较大时，默认启用加密对性能影响严重，使用时请注意。 ● 如果互信集群的一端集群配置了数据传输加密，则对端集群也需配置同样的数据传输加密。	false

配置项	描述	默认值
dfs.encrypt.data.transfer.algorithm	设置客户端访问HDFS的通道和HDFS数据传输通道的加密算法。只有在dfs.encrypt.data.transfer配置项设置为“true”，此参数才会生效。 **说明** 缺省值为“3des”，表示采用3DES算法进行加密。此处的值还可以设置为“rc4”，避免出现安全隐患，不推荐设置为该值。	3des
dfs.encrypt.data.transfer.cipher.suites	可以设置为空或“AES/CTR/NoPadding”，用于指定数据加密的密码套件。如果不指定此参数，则使用“dfs.encrypt.data.transfer.algorithm”参数指定的加密算法进行数据加密。默认值为“AES/CTR/NoPadding”。	AES/CTR/NoPadding

11.12.3.10 配置 Controller 与 Agent 间通信加密

操作场景

安装集群后Controller和Agent之间需要进行数据通信，在通信的过程中采用了Kerberos认证，出于对集群性能的考虑，通信过程默认不加密，对于一些安全要求较高用户可以采用以下方式进行加密。

对系统的影响

- 执行加密操作时，会自动重启Controller和所有Agent，重启期间会造成FusionInsight Manager暂时中断。
- 大集群下会导致管理节点性能有所下降，建议集群不超过200节点时开启该功能。

前提条件

已确认主备管理节点IP。

操作步骤

步骤1 以**omm**用户登录到主管理节点。

步骤2 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤3 执行以下命令，切换目录。

cd ${CONTROLLER_HOME}/sbin

步骤4 执行以下命令启用通信加密：

./enableRPCEncrypt.sh -t

执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**，查看主管理节点Controller的“ResHAStatus”是否为“Normal”，并可以重新登录FusionInsight Manager表示更改成功。

步骤5 如果需要关闭加密模式，执行以下命令：

./enableRPCEncrypt.sh -f

执行**sh ${BIGDATA_HOME}/om-server/om/sbin/status-oms.sh**，查看主管理节点Controller的“ResHAStatus”是否为“Normal”，并可以重新登录FusionInsight Manager表示更改成功。

----结束

11.12.3.11 更新 omm 用户 ssh 密钥

操作场景

在安装集群时，系统将自动为**omm**用户生成ssh认证私钥和公钥，用来建立节点间的互信。在集群安装成功后，如果原始私钥不慎意外泄露或者需要使用新的密钥时，系统管理员可以通过以下操作手动更改密钥值。

前提条件

- 已停止集群。
- 修改时禁止同时进行其他管理类操作。

操作步骤

步骤1 以**omm**用户登录到需要替换ssh密钥的节点。

如果该节点是Manager管理节点，务必在主管理节点上执行相关操作。

步骤2 执行以下命令，防止超时退出。

TMOUT=0

说明

执行完本章节操作后，请及时恢复超时退出时间，执行命令**TMOUT=***超时退出时间*。例如：**TMOUT=600**，表示用户无操作600秒后超时退出。

步骤3 执行以下命令，为节点生成新的密钥：

- 如果当前节点是Manager管理节点，执行以下命令：

  sh ${CONTROLLER_HOME}/sbin/update-ssh-key.sh

- 如果当前节点是非管理节点，执行以下命令：

  sh ${NODE_AGENT_HOME}/bin/update-ssh-key.sh

执行上述命令时界面提示“Succeed to update ssh private key.”信息，表示ssh密钥生成成功。

步骤4 执行以下命令将该节点的公钥拷贝到主管理节点：

scp ${HOME}/.ssh/id_rsa.pub** *oms_ip*:**${HOME}/.ssh/id_rsa.pub_bak

oms_ip：表示主管理节点IP。

根据提示输入**omm**用户密码完成文件拷贝。

步骤5 以**omm**用户登录到主管理节点。

步骤6 执行以下命令，防止超时退出：

TMOUT=0

步骤7 执行以下命令，切换目录：

cd ${HOME}/.ssh

步骤8 执行以下命令添加新的公钥信息：

cat id_rsa.pub_bak >> authorized_keys

步骤9 执行以下命令移动临时公钥文件到其他目录，例如，移动到“/tmp”目录。

mv -f id_rsa.pub_bak /tmp

步骤10 拷贝主管理节点的authorized_keys文件到集群内其他节点：

scp authorized_keys *node_ip*:**/${HOME}/.ssh/authorized_keys**

node_ip：集群内其他节点IP，不支持多个IP。

步骤11 执行以下命令无需输入密码确认私钥替换完成：

ssh *node_ip*

node_ip：集群内其他节点IP，不支持多个IP。

步骤12 登录FusionInsight Manager，在“主页”中单击待操作集群名称后的“ ··· > 启动”，启动集群。

----结束

11.12.4 安全维护

11.12.4.1 账户维护建议

建议系统管理员对账户例行检查，检查的内容包括：

- 操作系统、FusionInsight Manager以及各组件的账户是否有必要，临时账户是否已删除。
- 各类账户的权限是否合理。不同的管理员拥有不同的权限。
- 对各类账户的登录、操作记录进行检查和审计。

11.12.4.2 密码维护建议

用户身份验证是应用系统的门户。用户的账户和密码的复杂性、有效期等需根据客户的安全要求进行配置。

对密码的维护建议如下：

1. 专人保管操作系统密码。
2. 密码需要满足一定的强度要求，例如密码最少字符数、混合大小写等。
3. 密码传递时注意加密，尽量避免通过邮件传递密码。
4. 密码需要加密存储。
5. 系统移交时提醒企业用户更改密码。
6. 定期修改密码。

11.12.4.3 日志维护建议

利用日志记录来帮助发现非法操作、非法登录用户等异常情况。系统对于重要业务的操作需要记录日志。通过日志文件来定位异常。

定期检查日志

定期查看系统日志，若发现有非法操作、非法登录用户等异常情况，应根据异常情况进行相应的处理。

定期备份日志

FusionInsight Manager和集群提供的审计日志记录了用户活动信息和操作信息，可通过FusionInsight Manager导出审计日志。当系统中的审计日志过多时，可通过配置转储参数，将审计日志转储到指定服务器，避免引起集群节点磁盘空间不足。

维护责任人

网络监控工程师、系统维护工程师

11.12.5 安全声明

JDK 使用声明

MRS是一个大数据集群，为用户提供分布式的数据分析计算能力。本产品自带的JDK为OpenJDK，主要使用场景如下：

- 平台服务运行及维护使用。
- Linux客户端运行时使用（主要为业务提交、应用运维等）。

JDK 风险说明

系统对自带的JDK进行了权限控制，只有属于FusionInsight平台相关群组的用户才有权限访问，且平台部署在客户内网，安全风险较低。

JDK 加固

JDK加固相关操作请参考**加固策略**的“加固JDK”部分。

Hue 组件包含公网 IP 的说明

Hue组件使用的ipaddress，requests，Django等第三方包的测试用例及其注释包含的公网IP，组件在提供服务时不涉及这些IP，Hue组件的配置文件中不涉及公网IP。

12 MRS Manager 操作指导（适用于 2.x 及之前）

12.1 MRS Manager 简介

概述

MRS为用户提供海量数据的管理及分析功能，快速从结构化和非结构化的海量数据中挖掘您所需要的价值数据。开源组件结构复杂，安装、配置、管理过程费时费力，MRS Manager提供了企业级的大数据集群的统一管理平台：

- 提供集群状态的监控功能，您能快速掌握服务及主机的健康状态。
- 提供图形化的指标监控及定制，您能及时的获取系统的关键信息。
- 提供服务属性的配置功能，满足您实际业务的性能需求。
- 提供集群、服务、角色实例的操作功能，满足您一键启停等操作需求。

系统界面简介

MRS Manager提供统一的集群管理平台，帮助用户快捷、直观的完成集群的运行维护。MRS Manager请参考**访问MRS Manager（MRS 2.x及之前版本）**页面访问。

各操作入口的详细功能如**表12-1**所示。

表 12-1 界面操作入口功能描述

界面	功能描述
系统概览	提供柱状图、折线图、表格等多种图表方式展示所有服务的状态、各服务的主要监控指标、主机的状态统计。用户可以定制关键监控信息面板，并拖动到任意位置。系统概览支持数据自动刷新。
服务管理	提供服务监控、服务操作向导以及服务配置，帮助用户对服务进行统一管理。

界面	功能描述
主机管理	提供主机监控、主机操作向导，帮助用户对主机进行统一管理。
告警管理	提供告警查询、告警处理指导功能。帮助用户及时发现产品故障及潜在隐患，并进行定位排除，以保证系统正常运行。
审计管理	提供审计日志查询及导出功能。帮助用户查阅所有用户活动及操作。
租户管理	提供统一租户管理平台。
系统设置	用户可以进行监控和告警配置管理、备份管理。

当用户进入到“系统设置”的各子功能页面后，提供快捷方式跳转到其他System子功能页面，如表12-2所示。

快捷跳转操作示例如下所示。

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“系统设置”界面，任意单击一个功能链接，进入具体功能界面。

例如在“备份恢复管理”区域中单击“备份管理”，进入到“备份管理”界面。

步骤3 将鼠标移动到浏览器窗口的左边界，弹出“系统设置”黑色快捷菜单。鼠标移出该菜单后，该菜单收起。

步骤4 在弹出的快捷菜单上，可以单击某个功能链接直接跳转到对应的功能界面。

例如选择“维护 > 日志导出”，进入“日志导出”界面。

----结束

表 12-2 集群的 System 快捷菜单

菜单子标题	功能链接
备份恢复管理	备份管理
	恢复管理
维护	日志导出
	审计日志导出
	健康检查
监控和告警配置	Syslog配置
	阈值管理
	SNMP配置
	监控指标转储配置

菜单子标题	功能链接
	资源贡献排名配置
权限配置	用户管理
	用户组管理
	角色管理
	密码策略配置
	OMS数据库密码修改
资源管理	静态服务池
补丁管理	补丁管理

参考信息

MapReduce服务是云的一项数据分析服务，用于海量数据的管理和分析，简称MRS。

MRS通过MRS Manager管理大数据组件，例如Hadoop生态体系中的组件。因此，云MRS和MRS Manager管理界面上的部分概念需要区别，具体解释如表12-3：

表 12-3 差异对比参考

名词概念	云MRS	MRS Manager
MapReduce服务	表示云上的数据分析云服务，简称为MRS，包括Hive、Spark、Yarn、HDFS和ZooKeeper等组件。	为租户集群中的大数据组件提供的统一管理平台。

12.2 查看集群运行任务

操作场景

用户在MRS Manager进行操作触发运行任务时，会显示任务运行的过程与进度。关闭任务窗口后，需要通过任务管理功能，打开任务窗口。

MRS Manager默认保留10个最近运行的任务。例如重启服务、同步服务配置和执行健康检查。

操作步骤

步骤1 在MRS Manager，单击 ，打开“任务列表”。

“任务列表”可查看的信息包含：“任务名”、“状态”、“进度”、“开始时间”和“结束时间”。

步骤2 单击指定的任务名称，可查看任务执行过程中的详细信息。

----结束

12.3 监控管理

12.3.1 系统概览

MRS Manager支持将集群中所有部署角色的节点，按管理节点、控制节点和数据节点进行分类，分别计算关键主机监控指标在每类节点上的变化趋势，并在报表中按用户自定义的周期显示分布曲线图。如果一个主机属于多类节点，那么对应的指标将被统计多次。

该任务指导用户了解MRS集群的概览、及在MRS Manager查看、自定义与导出节点监控指标报表。

操作步骤

步骤1 登录MRS Manager，具体请参考**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager选择“系统概览”。

步骤3 在“时间区间”选择需要查看监控数据的时间段。可供选择的选项如下：

- 实时
- 最近3小时
- 最近6小时
- 最近24小时
- 最近一周
- 最近一个月
- 最近三个月
- 最近六个月
- 自定义：选择自定义时，在时间范围内自行选择需要查看的时间。

步骤4 单击“查看”可以查看相应时间区间的监控数据。

- MRS Manager在“服务概览”显示各个服务的“健康状态”和“角色数”。
- 单击曲线图表上侧的 图标，可显示具体的指标说明信息。

步骤5 自定义监控指标报表。

1. 单击“定制”，勾选需要在MRS Manager显示的监控指标。

   MRS Manager支持统计的指标共14个，界面最多显示12个定制的监控指标。

   - 集群主机健康状态统计
   - 集群网络读速率统计
   - 主机网络读速率分布
   - 主机网络写速率分布
   - 集群磁盘写速率统计

- 集群磁盘占用率统计
- 集群磁盘信息
- 主机磁盘占用率分布
- 集群磁盘读速率统计
- 集群内存占用率统计
- 主机内存占用率分布
- 集群网络写速率统计
- 主机CPU占用率分布
- 集群CPU占用率统计

2. 单击“确定”保存并显示所选指标。

   说明

   单击“清除”可批量取消全部选中的指标项。

步骤6 用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每60秒刷新一次”：刷新间隔60秒。
- “每120秒刷新一次”：刷新间隔120秒。
- “停止刷新”：停止刷新。

  说明

  勾选“全屏”会将“系统概览”窗口最大化。

步骤7 导出监控指标报表。

1. 选择报表的时间范围。可供选择的选项如下：
   - 实时
   - 最近3小时
   - 最近6小时
   - 最近24小时
   - 最近一周
   - 最近一个月
   - 最近三个月
   - 最近六个月
   - 自定义：选择自定义时，自行选择需要导出报表的时间。
2. 单击“导出”，Manager将生成指定时间范围内、已勾选的集群监控指标报表文件，请选择一个位置保存，并妥善保管该文件。

   说明

   如果需要查看指定时间范围的监控指标对应的分布曲线图，请单击“查看”，界面将显示用户自定义时间范围内选定指标的分布曲线图。

----结束

12.3.2 管理服务和主机监控

用户可以在日常使用中，可以在MRS Manager管理所有服务（含角色实例）和主机的状态及指标信息：

- 状态信息，包括运行、健康、配置及角色实例状态统计。
- 指标信息，各服务的主要监控指标项。
- 导出监控指标。

说明

用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

管理服务监控

步骤1 在MRS Manager，单击“服务管理”。

服务列表中标题包含“服务”、“操作状态”、“健康状态”、“配置状态”、“角色数”和“操作”。

- 服务操作状态描述如表12-4所示。

表 12-4 服务操作状态

状态	描述
已启动	服务已启动。
已停止	服务已停止。
启动失败	用户启动操作失败。
停止失败	用户停止操作失败。
未知	后台系统重启后，服务的初始状态。

- 服务健康状态如表12-5所示。

表 12-5 服务健康状态

状态	描述
良好	该服务中所有角色实例正常运行。
故障	至少一个角色实例运行状态为“故障”或被依赖的服务状态不正常。
未知	该服务中所有角色实例状态为“未知”。

状态	描述
正在恢复	后台系统正在尝试自动启动服务。
亚健康	该服务所依赖的服务状态不正常，异常服务的相关接口无法被外部调用。

- 服务配置状态如表12-6所示。

  表 12-6 服务配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
过期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
同步中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“服务”列按升序排列，单击**服务**、**操作状态**、**健康状态**或**配置状态**可修改排列方式。

步骤2 单击列表中指定服务名称，查看服务状态及指标信息。

步骤3 定制、导出监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。
3. 单击“导出”，导出当前查看的指标数据。

----结束

管理角色实例监控

步骤1 在MRS Manager，单击“服务管理”，在服务列表中单击服务指定名称。

步骤2 单击“实例”，查看角色状态。

角色实例列表中包含实例信息的**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**操作状态**、**健康状态**及**配置状态**。

- 角色实例的状态如表12-7所示。

  表 12-7 角色实例状态

状态	描述
已启动	角色实例已启动。

状态	描述
已停止	角色实例已停止。
启动失败	用户启动操作失败。
停止失败	用户停止操作失败。
退服中	角色实例正在退服。
已退服	角色实例已退服。
入服中	角色实例正在入服。
未知	后台系统重启后，角色实例的初始状态。

- 角色实例的健康状态如表12-8所示。

  表 12-8 角色实例健康状态

状态	描述
良好	该角色实例正常运行。
恢复中	后台系统正在尝试自动启动角色实例。
故障	该角色实例运行异常，如PID不存在，无法访问端口。
未知	角色实例所在主机与后台系统未连接。
亚健康	该角色实例部分正常运行

- 角色实例的配置状态如表12-9所示。

  表 12-9 角色实例配置状态

状态	描述
已同步	系统中最新的配置信息已生效。
过期	参数修改后，最新的配置未生效。需重启相应服务生效最新配置信息。
失败	参数配置过程中出现通信或读写异常。尝试使用“同步配置”恢复。
同步中	参数配置进行中。
未知	无法获取当前配置状态。

默认以“角色”列按升序排列，单击**角色**、**主机名**、**管理IP**、**业务IP**、**机架**、**操作状态**、**健康状态**或**配置状态**可修改排列方式。

支持在“角色”筛选相同角色的全部实例。

单击“高级搜索”，在角色搜索区域中设置搜索条件，单击“搜索”，查看指定的角色信息。单击“重置”清除输入的搜索条件。支持模糊搜索条件的部分字符。

步骤3 单击列表中指定角色实例名称，查看角色实例状态及指标信息。

步骤4 定制、导出监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。
3. 单击“导出”，导出当前查看的指标数据。

----结束

管理主机监控

步骤1 在MRS Manager，单击“主机管理”，看所有主机状态。

主机列表中包括主机名称、管理IP、业务IP、机架、网络速度、操作状态、健康状态、磁盘使用率、内存使用率、CPU使用率。

- 主机操作状态如**表12-10**所示。

表 12-10 主机操作状态

状态	描述
正常	主机及主机上的服务角色正常运行。
已隔离	主机被用户隔离，主机上的服务角色停止运行。

- 主机健康状态描述如**表12-11**所示。

表 12-11 主机健康状态

状态	描述
良好	主机心跳检测正常。
故障	主机心跳超时未上报。
未知	执行添加操作时，主机的初始状态。

默认以“主机名称”列按升序排列，单击**主机名称**、**管理IP**、**业务IP**、**机架**、**网络速度**、**操作状态**、**健康状态**、**磁盘使用率**、**内存使用率**或**CPU使用率**可修改排列方式。

单击“高级搜索”，在搜索区域中，设置查询条件，单击“搜索”，查看指定的主机。单击“重置”清除输入的搜索条件。支持模糊搜索条件的部分字符。

步骤2 单击列表中指定的主机名称，查看单个主机状态及指标。

步骤3 定制、导出监控图表。

1. 在“图表”区域框中，单击“定制”自定义服务监控指标。
2. 在“时间区间”选择查询时间，单击“查看”显示该时间段内的监控数据。

3. 单击“导出”，导出当前查看的指标数据。

----结束

12.3.3 管理资源分布

用户需要了解服务和主机关键监控指标中最高、最低或平均监控数据形成的曲线，即资源分布情况时，可以在MRS Manager上查看，支持查询1小时以内的监控数据。

用户也可以在MRS Manager上修改资源分布，使服务和主机的资源分布图表中，可以按自定义的数值显示一条或多条最高、最低监控数据形成的曲线。

部分监控指标的资源分布不记录。

操作步骤

- 查看服务监控指标的资源分布
  a. 在MRS Manager，单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“资源贡献排名”。
     “指标”中选择服务的关键指标，MRS Manager将显示过去1小时内指标的资源分布情况。
- 查看主机监控指标的资源分布
  a. 单击“主机管理”。
  b. 单击主机列表中指定的主机名称。
  c. 单击“资源贡献排名”。
     “指标”中选择主机的关键指标，MRS Manager将显示过去1小时内指标的资源分布情况。
- 配置资源分布
  a. 在MRS Manager，单击“系统设置”。
  b. 在“配置”区域“监控和告警配置”下，单击“资源贡献排名配置”。
  c. 修改资源分布的显示数量。
     - “TOP数量”填写最大值的显示数量。
     - “BOTTOM数量”填写最小值的显示数量。

     说明

     最大值与最小值的资源分布显示数量总和不能大于5。
  d. 单击“确定”保存设置。
     界面右上角提示“保存TOP数量和BOTTOM数量成功。”。

12.3.4 配置监控指标转储

用户可以在MRS Manager界面上配置监控指标数据对接参数，使集群内各监控指标数据通过FTP或SFTP协议保存到指定的FTP服务器，与第三方系统进行对接。FTP协议未加密数据可能存在安全风险，建议使用SFTP。

MRS Manager支持采集当前管理的集群内所有监控指标数据，采集的周期有30秒、60秒和300秒三种。监控指标数据在FTP服务器保存时，会根据采集周期分别保存在不同

的监控文件中，监控文件命名规则为“集群名称_metric_监控指标数据采集的周期_文件保存时间.log”。

前提条件

转储服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问转储服务器的IP地址和指定端口。转储服务器的FTP服务正常。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“监控指标转储配置”。

步骤3 表12-12介绍转储参数。

表 12-12 转储参数

参数名称	参数说明	是否必选
FTP IP地址	指定监控指标数据对接后存放监控文件的FTP服务器。	是
FTP端口	指定连接FTP服务器的端口。	是
FTP用户名	指定登录FTP服务器的用户名。	是
FTP密码	指定登录FTP服务器的密码。	是
保存路径	指定监控文件在FTP服务器保存的路径。	是
转储时间间隔（秒）	指定监控文件在FTP服务器保存的周期，单位为秒。	是
转储模式	指定监控文件发送时使用的协议。可选协议为“FTP”和“SFTP”。	是
SFTP服务公钥	指定FTP服务器的公共密钥，“模式”选择“SFTP”时此参数生效。建议配置公共密钥，否则可能存在安全风险。	否

步骤4 单击“确定”，设置完成。

----结束

12.4 告警管理

12.4.1 查看与手动清除告警

操作场景

用户可以在MRS Manager查看、清除告警。

一般情况下，告警处理后，系统自动清除该条告警记录。当告警不具备自动清除功能且用户已确认该告警对系统无影响时，可手动清除告警。

在MRS Manager界面可查看最近十万条告警（包括未清除的、手动清除的和自动清除的告警）。如果已清除告警超过十万条达到十一万条，系统自动将最早的一万条已清除告警转存，转存路径为主管理节点“${BIGDATA_HOME}/OMSV100R001C00x8664/workspace/data”。第一次转存告警时自动生成目录。

说明

用户可以选择页面自动刷新间隔的设置，也可以单击 马上刷新。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

操作步骤

步骤1 在MRS Manager，单击“告警管理”，在告警列表查看告警信息。

- 告警列表每页默认显示最近的十条告警。
- 默认以“产生时间”列按降序排列，单击“告警ID”、“告警名称”、“告警级别”、“产生时间”、“定位信息”或“操作”可修改排列方式。
- 支持在“告警级别”筛选相同级别的全部告警。结果包含已清除和未清除的告警。
- 分别单击 、 、 或 可以快速筛选级别为“致命”、“严重”、“一般”或“警告”的告警。

步骤2 单击“高级搜索”显示告警搜索区域，设置查询条件后，单击“搜索”，查看指定的告警信息。单击“重置”清除输入的搜索条件。

说明

“开始时间”和“结束时间”表示时间范围的开始时间和结束时间，可以搜索此时间段内产生的告警。

查看“告警参考”章节告警帮助，按照帮助指导处理告警。如果某些场景中告警由于MRS依赖的其他云服务产生，可能需要联系对应云服务运维人员处理。

步骤3 处理完告警后，若需手动清除，单击“清除告警”，手动清除告警。

说明

如果有多个告警已完成处理，可选中一个或多个待清除的告警，单击“清除告警”，批量清除告警。每次最多批量清除300条告警。

----结束

12.4.2 配置监控与告警阈值

操作场景

配置监控与告警阈值用于关注各指标的健康情况。勾选“发送告警”后，当监控数据达到告警阈值，系统将会触发一条告警信息，将在“告警管理”中出现此告警信息。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“阈值配置”，依据规划选择监控指标并设置其基线。

步骤3 单击某一指标例如“CPU使用率”，单击“添加规则”。

步骤4 在“配置”对话框中填写监控指标规则参数。

表 12-13 监控指标规则参数

参数名	参数解释	参数值
规则名称	规则名称	CPU_MAX（举例）
参考日期	查看某指标的历史参考数据	2014/11/06（举例）
阈值类型	选择某指标的最大值或最小值，类型为“最大值”表示指标的实际值大于设置的阈值时系统将产生告警，类型为“最小值”表示指标的实际值小于设置的阈值时系统将产生告警。	• 最大值 • 最小值
告警级别	告警级别	• 致命 • 严重 • 一般 • 提示
时间范围	设置规则生效时监控指标的具体时间段	从00:00到23:59（举例）
阈值	设置规则监控指标的阈值	设置数值 80（举例）
日期	设置规则生效的日期类型	• 工作日 • 周末 • 其它
添加日期	日期选择“其他”时该参数生效。可选择多个日期。	11/30（举例）

步骤5 单击“确定”。界面右上角弹出提示“模板保存成功。”。

“发送告警”默认已勾选。Manager会检查监控指标数值是否满足阈值条件，若连续检查且不满足的次数等于“平滑次数”设置的值则发送告警，支持自定义。“检查周期(秒)”表示Manager检查监控指标的时间间隔。

步骤6 在新添加规则所在的行，单击“操作”下的“应用”，界面右上角弹出提示规则xx应用成功，完成添加。单击“操作”下的“取消应用”，界面右上角弹出提示规则xx取消成功。

----结束

12.4.3 配置 Syslog 北向参数

操作场景

该任务指导用户以Syslog方式将MRS Manager的告警事件上报到指定的监控运维系统中。

须知

Syslog协议未做加密，传输数据容易被窃取，存在安全风险。

前提条件

对接服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问对接服务器的IP地址和指定端口。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“Syslog配置”。

“Syslog服务”的开关默认为关闭，单击启用Syslog服务。

步骤3 设置表12-14所示的对接参数。

表 12-14 对接参数

参数区域	参数名称	参数说明
Syslog协议	服务IP	设置对接服务器IP地址。
	服务端口	设置对接端口。
	协议	设置协议类型，取值范围： ● “TCP” ● “UDP”

参数区域	参数名称	参数说明
	安全级别	设置上报消息的严重程度，取值范围： ● “Informational” ● “Emergency” ● “Alert” ● “Critical” ● “Error” ● “Warning” ● “Notice” ● “Debug”
	Facility	设置产生日志的模块。
	标识符	设置产品标识，默认为“MRS Manager”。
报告信息	报文格式	设置告警报告的消息格式，具体要求请参考界面帮助。
	报告告警类型	设置需要上报的告警类型。 ● “故障”表示Manager产生告警时会上报Syslog告警消息。 ● “清除”表示清除Manager告警时会上报Syslog告警消息。 ● “事件”表示Manager产生事件时会上报Syslog告警消息。
	报告告警级别	设置需要上报的告警级别。支持“提示”、“一般”、“严重”和“致命”。
未恢复告警上报设置	周期上报未恢复告警	设置是否按指定周期上报未清除的告警。“周期上报未恢复告警”的开关默认为关闭，单击启用此功能。

参数区域	参数名称	参数说明
	间隔时间（分钟）	设置周期上报未恢复告警到远程Syslog服务的时间间隔，当“周期上报未恢复告警”开关打开时启用。单位为分钟，默认值为“15”，取值范围为5分钟到一天（1440分钟）。
心跳设置	上报心跳	设置是否开启周期上报Syslog心跳消息。“周期上报未恢复告警”的开关默认为关闭，单击启用此功能。
	心跳周期（分钟）	设置周期上报心跳的时间间隔，当“上报心跳”开关打开时启用。单位为分钟，默认值为“15”，取值范围为1-60。
	心跳报文	设置心跳上报的内容，当“上报心跳”开关打开时启用，不能为空。支持数字、字母、下划线、竖线、冒号、空格、英文逗号和句号等字符，长度小于等于256。

说明

设置周期上报心跳报文后，在某些集群容错自动恢复的场景下（例如主备管理节点倒换）可能会出现报文上报中断的现象，此时等待自动恢复即可。

步骤4 单击“确定”，设置完成。

----结束

12.4.4 配置 SNMP 北向参数

操作场景

该任务指导用户采用SNMP协议把MRS Manager的告警、监控数据集成到网管平台。

前提条件

对接服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问对接服务器的IP地址和指定端口。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“配置”区域“监控和告警配置”下，单击“SNMP配置”。

“SNMP服务”的开关默认为关闭，单击启用SNMP服务。

步骤3 设置表12-15所示的对接参数。

表 12-15 对接参数

参数名称	参数说明
版本	SNMP协议版本号，取值范围： ● v2c：低版本，安全性较低 ● v3：高版本，安全性比v2c高 推荐使用v3版本。
本地端口	本地端口，默认值“20000”，取值范围“1025”到“65535”。
读团体名	该参数仅在设置“版本”为v2c时存在，用于设置只读团体名。
写团体名	该参数仅在设置“版本”为v2c时存在，用于设置可写团体名。
安全用户名	该参数仅在设置“版本”为v3时存在，用于设置协议安全用户名。
认证协议	该参数仅在设置“版本”为v3时存在，用于设置认证协议，推荐选择SHA。
认证密码	该参数仅在设置“版本”为v3时存在，用于设置认证密钥。
确认认证密码	该参数仅在设置“版本”为v3时存在，用于确认认证密钥。
加密协议	该参数仅在设置“版本”为v3时存在，用于设置加密协议，推荐选择AES256。
加密密码	该参数仅在设置“版本”为v3时存在，用于设置加密密钥。
确认加密密码	该参数仅在设置“版本”为v3时存在，用于确认加密密钥。

说明

- “认证密码”和“加密密码”密码长度为8到16位，至少需要包含大写字母、小写字母、数字、特殊字符中的3种类型字符。两个密码不能相同。两个密码不可和安全用户名或安全用户名的逆序字符相同。
- 使用SNMP协议从安全方面考虑，需要定期修改“认证密码”和“加密密码”密码。
- 使用SNMP v3版本时，安全用户在5分钟之内连续鉴权失败5次将被锁定，5分钟后自动解锁。

步骤4 单击“Trap目标”下的“添加Trap目标”，在弹出的“添加Trap目标”对话框中填写以下参数：

- 目标标识：Trap目标标识，一般指接收Trap的网管或主机标识。长度限制1～255字节，一般由字母或数字组成。
- 目标IP：目标IP。可使用A、B、C类IP地址，要求可与管理节点的管理平面IP地址互通。
- 目标端口：接收Trap的端口，要求与对端保持一致，取值范围“0”～“65535”。
- Trap团体名：该参数仅在设置Version为v2c时存在，用于设置主动上报团体名。

单击“确定”，设置完成，退出“添加Trap目标”对话框。

步骤5 单击“确定”，设置完成。

----结束

12.5 对象管理

12.5.1 对象管理简介

MRS集群包含了各类不同的基本对象，不同对象的描述介绍如表12-16所示：

表 12-16 MRS 基本对象概览

对象	描述	举例
服务	可以完成具体业务的一类功能集合。	例如KrbServer服务和LdapServer服务。
服务实例	服务的具体实例，一般情况下可使用服务表示。	例如KrbServer服务。
服务角色	组成一个完整服务的一类功能实体，一般情况下可使用角色表示。	例如KrbServer由KerberosAdmin角色和KerberosServer角色组成。
角色实例	服务角色在主机节点上运行的具体实例。	例如运行在Host2上的KerberosAdmin，运行在Host3上的KerberosServer。
主机	一个弹性云服务器，可以运行Linux系统。	例如Host1～Host5。
机架	一组包含使用相同交换机的多个主机集合的物理实体。	例如Rack1，包含Host1～Host5。
集群	由多台主机组成的可以提供多种服务的逻辑实体。	例如名为Cluster1的集群由（Host1～Host5）5个主机组成，提供了KrbServer和LdapServer等服务。

12.5.2 查看配置

用户可以在MRS Manager上查看服务（含角色）和角色实例的配置。

操作步骤

- 查看服务的配置。
  a. 在MRS Manager，单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“服务配置”。
  d. 在“参数类别”选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
  e. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
     在服务节点下的参数属于服务配置参数，在角色节点下的参数是角色配置参数。
  f. 在“非默认”选项中选择“非默认”，界面上显示参数值为非默认值的参数。
- 查看角色实例的配置。
  a. 在MRS Manager，单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“实例”页签。
  d. 单击角色实例列表中指定的角色实例名称。
  e. 单击“实例配置”。
  f. 在“参数类别”选择“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
  g. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。
  h. 在“非默认”选项中选择“非默认”，界面上显示参数值为非默认值的参数。

12.5.3 管理服务操作

用户可以在MRS Manager：

- 启动操作状态为“停止”、“停止失败”或“启动失败”服务，以使用该服务。
- 停止不再使用或异常服务。
- 重启异常或配置过期的服务，以恢复或生效服务功能。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在指定服务所在行，单击“启动”、“停止”和“重启”执行启动、停止和重启操作。

服务之间存在依赖关系。对某服务执行启动、停止和重启操作时，与该服务存在依赖关系的服务将受到影响。

具体影响如下：

- 启动某服务，该服务依赖的下层服务需先启动，服务功能才可生效。
- 停止某服务，依赖该服务的上层服务将无法提供功能。
- 重启某服务，依赖该服务且启动的上层服务需重启后才可生效。

----结束

12.5.4 配置服务参数

用户可以根据实际业务场景，在MRS Manager中快速查看和修改服务默认的配置，及导出或导入配置。

对系统的影响

- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。
- 集群中只剩下一个DBService角色实例时，不支持修改DBService服务的参数。

操作步骤

- 修改服务参数。
  a. 单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“服务配置”。
  d. 在“参数类别”选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的根节点分别为服务名称和角色名称。
  e. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

     修改某个参数的值后需要取消修改，可以单击恢复。

     说明

     如果需要批量修改服务某个角色多个实例的配置，可以使用主机组实现实例参数的批量配置。在“角色”选择角色名称，然后在“主机”打开“<选择主机>”。“主机组名”填写一个名称，“主机”列表中勾选要修改的主机并加入“已选择的主机”，单击“确定”添加主机组。添加的主机组可以在“主机”中选择，且仅在当前页面有效，刷新页面后将无法保存。
  f. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

     界面提示“操作成功。”，单击“完成”，服务成功启动。

     说明

     更新YARN服务队列的配置且不重启服务时，选择“更多 > 刷新队列”更新队列使配置生效。
- 导出服务配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 单击“服务配置”。

d. 单击“导出服务配置”，选择一个位置保存配置文件。

- 导入服务配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 单击“服务配置”。
  d. 单击“导入服务配置”。
  e. 选择一个指定的配置文件。
  f. 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”。
     界面提示“操作成功。”，单击“完成”，服务成功启动。

12.5.5 配置服务自定义参数

MRS各个组件支持开源的所有参数，在MRS Manager支持修改部分关键使用场景的参数，且部分组件的客户端可能不包含开源特性的所有参数。如果需要修改其他Manager未直接支持的组件参数，用户可以在Manager通过自定义配置项功能为组件添加新参数。添加的新参数最终将保存在组件的配置文件中并在重启后生效。

对系统的影响

- 配置服务属性后，需要重启此服务，重启期间无法访问服务。
- 配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

前提条件

用户已充分了解需要新添加的参数意义、生效的配置文件以及对组件的影响。

操作步骤

步骤1 在MRS Manager界面，单击“服务管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“服务配置”。

步骤4 在“参数类别”选择“全部配置”。

步骤5 在左侧导航栏选择“自定义”，Manager将显示当前组件的自定义参数。

“参数文件”显示保存用户新添加的自定义参数的配置文件。每个配置文件中可能支持相同名称的开源参数，设置不同参数值后生效结果由组件加载配置文件的顺序决定。自定义参数支持服务级别与角色级别，请根据业务实际需要选择。不支持单个角色实例添加自定义参数。

步骤6 根据配置文件与参数作用，在对应参数项所在行“名称”列输入组件支持的参数名，在“值”列输入此参数的参数值。

- 支持单击 和 增加或删除一条自定义参数。第一次单击 添加自定义参数后才支持删除操作。
- 修改某个参数的值后需要取消修改，可以单击 恢复。

步骤7 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

任务示例

配置Hive自定义参数

Hive依赖于HDFS，默认情况下Hive访问HDFS时是HDFS的客户端，生效的配置参数统一由HDFS控制。例如HDFS参数“ipc.client.rpc.timeout”影响所有客户端连接HDFS服务端的RPC超时时间，如果用户需要单独修改Hive连接HDFS的超时时间，可以使用自定义配置项功能进行设置。在Hive的“core-site.xml”文件增加此参数可被Hive服务识别并代替HDFS的设置。

步骤1 在MRS Manager界面，选择 “服务管理 > Hive > 服务配置”。

步骤2 在“参数类别”选择“全部配置”。

步骤3 在左侧导航栏选择Hive服务级别“自定义”，Manager将显示Hive支持的服务级别自定义参数。

步骤4 在“core-site.xml”对应参数“core.site.customized.configs”的“名称：”输入“ipc.client.rpc.timeout”，“值：”输入新的参数值，例如“150000”。单位为毫秒。

步骤5 单击“保存配置”，勾选“重新启动受影响的服务或实例。”并单击“是”重启服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

12.5.6 同步服务配置

操作场景

当用户发现部分服务的“配置状态”为“过期”或“失败”时，您可以尝试使用同步配置功能，以恢复配置状态。或者集群中所有服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

对系统的影响

同步服务配置后，需要重启配置过期的服务。重启时对应的服务不可用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在服务列表中，单击指定服务名称。

步骤3 在服务状态及指标信息上方，选择“更多 > 同步配置”。

步骤4 在弹出窗口根据提示输入密码，单击“确认”完成身份验证。身份验证后勾选“重启配置过期的服务或实例。”，并单击“确定”重启配置过期的服务。

界面提示“操作成功”，单击“完成”，服务成功启动。

----结束

12.5.7 管理角色实例操作

操作场景

用户可以在MRS Manager启动操作状态为“停止”、“停止失败”或“启动失败”角色实例，以使用该角色实例，也可以停止不再使用或异常的角色实例，或者重启异常的角色实例，以恢复角色实例功能。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 单击服务列表中指定的服务名称。

步骤3 单击“实例”页签。

步骤4 勾选待操作角色实例前的复选框。

步骤5 选择“更多 > 启动实例”、“停止实例”或“重启实例”，执行相应操作。

----结束

12.5.8 配置角色实例参数

操作场景

用户可以根据实际业务场景，在MRS Manager中快速查看及修改角色实例默认的配置。支持导出或导入配置。

对系统的影响

配置HBase、HDFS、Hive、Spark、Yarn、Mapreduce服务属性后，需要重新下载并更新客户端配置文件。

操作步骤

- 修改角色实例参数。
  a. 单击“服务管理”。
  b. 单击服务列表中指定的服务名称。
  c. 单击“实例”页签。
  d. 单击角色实例列表中指定的角色实例名称。
  e. 单击“实例配置”页签。
  f. 在“参数类别”选择“全部配置”，界面上将显示该角色实例的全部配置参数导航树。
  g. 在导航树选择指定的参数，修改参数值。支持在“搜索”输入参数名直接搜索并显示结果。

     修改某个参数的值后需要取消修改，可以单击 恢复。

h. 单击“保存配置”，勾选“重启角色实例”并单击“确定”，重启角色实例。

界面提示“操作成功。”，单击“完成”，角色实例成功启动。

- 导出角色实例配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 选中某角色或单击“实例”。
  d. 选择指定主机上某角色实例。
  e. 单击“实例配置”。
  f. 单击“导出实例配置”，导出指定角色实例配置数据并选择一个位置保存。
- 导入角色实例配置参数。
  a. 单击“服务管理”。
  b. 选中某项服务。
  c. 选中某角色或单击“实例”。
  d. 选择指定主机上某角色实例。
  e. 单击“实例配置”。
  f. 单击“导入实例配置”，导入指定角色实例配置数据。
  g. 单击“保存配置”，勾选“重启角色实例。”并单击“确定”。
  界面提示“操作成功。”，单击“完成”，角色实例成功启动。

12.5.9 同步角色实例配置

操作场景

当用户发现角色实例的“配置状态”为“过期”或“失败”时，可以在MRS Manager尝试使用同步配置功能，同步角色实例的配置数据与后台配置数据，以恢复配置状态。

对系统的影响

同步配置角色实例后需要重启配置过期的角色实例。重启时对应的角色实例不可用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”，选择服务名称。

步骤2 单击“实例”页签。

步骤3 在角色实例列表中，单击指定角色实例名称。

步骤4 在角色实例状态及指标信息上方，选择“更多 > 同步配置”。

步骤5 在弹出窗口勾选“重启配置过期的服务或实例。”，并单击“确定”重启角色实例。

界面提示“操作成功。”，单击“完成”，角色实例成功启动。

----结束

12.5.10 退服和入服务角色实例

操作场景

某个Core或Task节点出现问题时，可能导致整个集群状态显示为“异常”。MRS集群支持将数据存储在不同Core节点，用户可以在MRS Manager指定角色实例退服，使退服的角色实例不再提供服务。在排除故障后，可以将已退服的角色实例入服。

MRS 1.6.0之前版本支持退服、入服的角色实例包括：

- HDFS的DataNode角色实例
- Yarn的NodeManager角色实例

MRS 1.6.0及以后版本支持退服、入服的角色实例包括：

- HDFS的DataNode角色实例
- Yarn的NodeManager角色实例
- HBase的RegionServer角色实例
- Kafka的Broker角色实例

限制：

- 当DataNode数量少于或等于HDFS的副本数时，不能执行退服操作。例如HDFS副本数为3时，则系统中少于4个DataNode，将无法执行退服，Manager在执行退服操作时会等待30分钟后报错并退出执行。
- Kafka Broker数量少于或等于副本数时，不能执行退服。例如Kafka副本数为2时，则系统中少于3个节点，将无法执行退服，Manager执行退服操作时会失败并退出执行。
- 已经退服的角色实例，必须执行入服操作启动该实例，才能重新使用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 单击服务列表中相应服务。

步骤3 单击“实例”页签。

步骤4 勾选指定角色实例名称前的复选框。

步骤5 选择“更多 > 退服”或“入服”执行相应的操作。

说明

实例退服操作未完成时在其他浏览器窗口重启集群中相应服务，可能导致MRS Manager提示停止退服，实例的“操作状态”显示为“已启动”。实际上后台已将该实例退服，请重新执行退服操作同步状态。

----结束

12.5.11 管理主机操作

操作场景

当主机故障异常时，用户可能需要在MRS Manager停止主机上的所有角色，对主机进行维护检查。故障清除后，启动主机上的所有角色恢复主机业务。

操作步骤

步骤1 单击“主机管理”。

步骤2 勾选待操作主机前的复选框。

步骤3 选择“更多 > 启动所有角色”或“停止所有角色”执行相应操作。

----结束

12.5.12 隔离主机

操作场景

用户发现某个主机出现异常或故障，无法提供服务或影响集群整体性能时，可以临时将主机从集群可用节点排除，使客户端访问其他可用的正常节点。在为集群安装补丁的场景中，也支持排除指定节点不安装补丁。

该任务指导用户在MRS Manager上根据实际业务或运维规划手工将主机隔离。隔离主机仅支持隔离非管理节点。

对系统的影响

- 主机隔离后该主机上的所有角色实例将被停止，且不能对主机及主机上的所有实例进行启动、停止和配置等操作。
- 主机隔离后无法统计并显示该主机硬件和主机上实例的监控状态及指标数据。

操作步骤

步骤1 在MRS Manager单击“主机管理”。

步骤2 勾选待隔离主机前的复选框。

步骤3 选择“更多 > 隔离主机”。

步骤4 在“隔离主机”，单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功隔离，“操作状态”显示为“已隔离”

说明

已隔离的主机，可以取消隔离重新加入集群，请参见**取消隔离主机**。

----结束

12.5.13 取消隔离主机

操作场景

用户已排除主机的异常或故障后，需要将主机隔离状态取消才能正常使用。

该任务指导用户在MRS Manager上取消隔离主机。

前提条件

- 主机状态为“已隔离”。
- 主机的异常或故障已确认修复。

操作步骤

步骤1 在MRS Manager单击“主机管理”。

步骤2 勾选待取消隔离主机前的复选框。

步骤3 选择“更多 > 取消隔离主机”。

步骤4 在“取消隔离主机”，单击“确定”。

界面提示“操作成功。”，单击“完成”，主机成功取消隔离，“操作状态”显示为“正常”。

步骤5 单击已取消隔离主机的名称，显示主机“状态”，单击“启动所有角色”。

----结束

12.5.14 启动及停止集群

操作场景

集群是包含着服务组件的集合。用户可以启动或者停止集群中所有服务。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在服务列表上方，选择“更多 > 启动集群”或“停止集群”执行相应的操作。

----结束

12.5.15 同步集群配置

操作场景

当MRS Manager显示全部服务或部分服务的“配置状态”为“过期”或“失败”时，用户可以尝试使用同步配置功能，以恢复配置状态。

- 若集群中所有服务的配置状态为“失败”时，同步集群的配置数据与后台配置数据。
- 若集群中某些服务的配置状态为“失败”时，同步指定服务的配置数据与后台配置数据。

对系统的影响

同步集群配置后，需要重启配置过期的服务。重启时对应的服务不可用。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 在服务列表上方，选择“更多 > 同步配置”。

步骤3 在弹出窗口输入当前登录用户密码进行身份验证，身份验证后勾选“重启配置过期的服务或实例。”，并单击“确定”，重启配置过期的服务。

界面提示“操作成功”，单击“完成”，集群成功启动。

----结束

12.5.16 导出集群的配置数据

操作场景

为了满足实际业务的需求，用户可以在MRS Manager中将集群所有配置数据导出，导出文件用于快速更新服务配置。

操作步骤

步骤1 在MRS Manager，单击“服务管理”。

步骤2 选择“更多 > 导出集群配置”。

导出文件用于更新服务配置，请参见**配置服务参数**中**导入服务配置参数**。

----结束

12.6 日志管理

12.6.1 关于日志

日志描述

MRS集群的日志保存路径为“/var/log/Bigdata”。日志分类见下表：

表 12-17 日志分类一览表

日志类型	日志描述
安装日志	安装日志记录了Manager、集群和服务安装的程序信息，可用于定位安装出错的问题。
运行日志	运行日志记录了集群各服务运行产生的运行轨迹信息及调试信息、状态变迁、未产生影响的潜在问题和直接的错误信息。
审计日志	审计日志中记录了用户活动信息和用户操作指令信息，可用于安全事件中定位问题原因及划分事故责任。

MRS日志目录清单见下表：

表 12-18 日志目录一览表

文件目录	日志内容
/var/log/Bigdata/audit	组件审计日志。
/var/log/Bigdata/controller	日志采集脚本日志。 controller进程日志。 controller监控日志。
/var/log/Bigdata/dbservice	DBService日志。
/var/log/Bigdata/flume	Flume日志。
/var/log/Bigdata/hbase	HBase日志。
/var/log/Bigdata/hdfs	HDFS日志。
/var/log/Bigdata/hive	Hive日志。
/var/log/Bigdata/httpd	httpd日志。
/var/log/Bigdata/hue	Hue日志。
/var/log/Bigdata/kerberos	Kerberos日志。
/var/log/Bigdata/ldapclient	LDAP客户端日志。
/var/log/Bigdata/ldapserver	LDAP服务端日志。
/var/log/Bigdata/loader	Loader日志。
/var/log/Bigdata/logman	logman脚本日志管理日志。
/var/log/Bigdata/mapreduce	MapReduce日志。
/var/log/Bigdata/nodeagent	NodeAgent日志。
/var/log/Bigdata/okerberos	OMS Kerberos日志。
/var/log/Bigdata/oldapserver	OMS LDAP日志。
/var/log/Bigdata/omm	oms：“omm”服务端的复杂事件处理日志、告警服务日志、HA日志、认证与授权管理日志和监控服务运行日志。 oma：“omm”代理端的安装运行日志。 core：“omm”代理端与“HA”进程失去响应的dump日志。
/var/log/Bigdata/spark	Spark日志。
/var/log/Bigdata/sudo	omm执行sudo命令产生的日志。
/var/log/Bigdata/timestamp	时间同步管理日志。
/var/log/Bigdata/tomcat	Tomcat日志。
/var/log/Bigdata/yarn	Yarn日志。

文件目录	日志内容
/var/log/Bigdata/zookeeper	ZooKeeper日志。
/var/log/Bigdata/kafka	Kafka日志。
/var/log/Bigdata/storm	Storm日志。
/var/log/Bigdata/patch	补丁日志。

运行日志

运行日志记录的运行信息描述如表12-19所示。

表 12-19 运行信息一览表

运行日志	日志描述
服务安装前的准备日志	记录服务安装前的准备工作，如检测、配置和反馈操作的信息。
进程启动日志	记录进程启动过程中执行的命令信息。
进程启动异常日志	记录进程启动失败时产生异常的信息，如依赖服务错误、资源不足等
进程运行日志	记录进程运行轨迹信息及调试信息，如函数入口和出口打印、模块间接口消息等。
进程运行异常日志	记录导致进程运行时错误的错误信息，如输入对象为空、编解码失败等错误。
进程运行环境信息日志	记录进程运行环境的信息，如资源状态、环境变量等。
脚本日志	记录脚本执行的过程信息。
资源回收日志	记录资源回收的过程信息。
服务卸载时的清理日志	记录卸载服务时执行的步骤操作信息，如清除目录数据、执行时间等

审计日志

审计日志记录的审计信息包含Manager审计信息和组件审计信息。

表 12-20 Manager 审计信息一览表

审计日志	操作类型	操作
Manager审计日志	用户管理	创建用户 修改用户 删除用户 创建组 修改组 删除组 添加角色 修改角色 删除角色 密码策略修改 修改密码 密码重置 用户登录 用户注销 屏幕解锁 下载认证凭据 用户越权操作 用户账号解锁 用户账号锁定 屏幕锁定 导出用户 导出用户组 导出角色
	租户管理	保存静态配置 增加租户 删除租户 关联租户服务 删除租户服务 配置资源 创建资源 删除资源 增加资源池 修改资源池 删除资源池 恢复租户数据

审计日志	操作类型	操作
	集群管理	启动集群 停止集群 保存配置 同步集群配置 定制集群监控指标 保存监控阈值 下载客户端配置 北向接口配置 北向SNMP接口配置 创建阈值模板 删除阈值模板 应用阈值模板 保存集群监控配置数据 导出配置数据 导入集群配置数据 导出安装模板 修改阈值模板 取消阈值模板应用 屏蔽告警 发送告警 修改OMS数据库密码 修改组件数据库密码 启动集群的健康检查 更新健康检查的配置 导出集群健康检查的结果 导入证书文件 删除健康检查历史报告 导出健康检查历史报告 定制报表监控指标 导出报表监控数据 定制静态资源池监控指标 导出静态资源池监控数据

审计日志	操作类型	操作
	服务管理	启动服务 停止服务 同步服务配置 刷新服务队列 定制服务监控指标 重启服务 导出服务监控数据 导入服务配置数据 启动服务的健康检查 导出服务健康检查的结果 服务配置 上传配置文件 下载配置文件
	实例管理	同步实例配置 实例入服 实例退服 启动实例 停止实例 定制实例监控指标 重启实例 导出实例监控数据 导入实例配置数据
	主机管理	设置节点机架 启动所有角色 停止所有角色 隔离主机 取消隔离主机 定制主机监控指标 导出主机监控数据 启动主机的健康检查 导出主机健康检查的结果

审计日志	操作类型	操作
	维护管理	导出告警 清除告警 导出事件 批量清除告警 SNMP清除告警 SNMP添加trap目标 SNMP删除trap目标 SNMP检查告警 SNMP同步告警 修改审计转储配置 导出审计日志 采集日志文件 下载日志文件 上传文件 删除已上传的文件 创建备份任务 执行备份任务 停止备份任务 删除备份任务 修改备份任务 锁定备份任务 解锁备份任务 创建恢复任务 执行恢复任务 停止恢复任务 重试恢复任务 删除恢复任务

表 12-21 组件审计信息一览表

审计日志	操作类型	操作
DBService审计日志	维护管理	备份恢复操作

审计日志	操作类型	操作
HBase审计日志	DDL（数据定义）语句	创建表 删除表 修改表 增加列族 修改列族 删除列族 启用表 禁用表 用户信息修改 修改密码 用户登录
	DML（数据操作）语句	put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并put数据（针对hbase:meta表、_ctmeta_表和hbase:acl表） 检查并删除数据（针对hbase:meta表、_ctmeta_表和hbase:acl表）
	权限控制	给用户授权 取消用户授权
Hive审计日志	元数据操作	元数据定义，如创建数据库、表等 元数据删除，如删除数据库、表等 元数据修改，如增加列、重命名表等 元数据导入/导出
	数据维护	向表中加载数据 向表中插入数据
	权限管理	创建/删除角色 授予/回收角色 授予/回收权限
HDFS审计日志	权限管理	文件/文件夹访问权限 文件/文件夹owner信息

审计日志	操作类型	操作
	文件操作	创建文件夹 创建文件 打开文件 追加文件内容 修改文件名称 删除文件/文件夹 设置文件时间属性 设置文件副本个数 多文件合并 文件系统检查 文件链接
Mapreduce审计日志	程序运行	启动Container请求 停止Container请求 Container结束，状态为成功 Container结束，状态为失败 Container结束，状态为中止 提交任务 结束任务
LdapServer审计日志	维护管理	添加操作系统用户 添加组 添加用户到组 删除用户 删除组
KrbServer审计日志	维护管理	修改kerberos账号密码 添加kerberos账号 删除kerberos账号 用户认证
Loader审计日志	安全管理	用户登录
	元数据管理	查询connector 查询framework 查询step

审计日志	操作类型	操作
	数据源连接管理	查询数据源连接 增加数据源连接 更新数据源连接 删除数据源连接 激活数据源连接 禁用数据源连接
	作业管理	查询作业 创建作业 更新作业 删除作业 激活作业 禁用作业 查询作业所有执行记录 查询作业最近执行记录 提交作业 停止作业
Hue审计日志	服务启动	启动Hue
	用户操作	用户登录 用户退出
	任务操作	创建任务 修改任务 删除任务 提交任务 保存任务 任务状态更新
Zookeeper审计日志	权限管理	设置ZNODE访问权限
	ZNODE操作	创建ZNODE 删除ZNODE 设置ZNODE数据
Storm审计日志	Nimbus	提交拓扑 中止拓扑 重分配拓扑 去激活拓扑 激活拓扑

审计日志	操作类型	操作
	UI	中止拓扑 重分配拓扑 去激活拓扑 激活拓扑

MRS的审计日志保存在数据库中，可通过“审计管理”页面查看及导出审计日志。

组件审计日志的文件信息见下表。部分组件审计日志文件保存在“/var/log/Bigdata/audit”，例如HDFS、HBase、Mapreduce、Hive、Hue、Yarn、Storm和ZooKeeper。每天凌晨3点自动将组件审计日志压缩备份到“/var/log/Bigdata/audit/bk”，最多保留最近的90个压缩备份文件，不支持修改备份时间。

其他组件审计日志文件保存在组件日志目录中。

表 12-22 组件审计日志目录

组件名称	审计日志目录
DBService	/var/log/Bigdata/audit/dbservice/dbservice_audit.log
HDFS	/var/log/Bigdata/audit/hdfs/nn/hdfs-audit-namenode.log /var/log/Bigdata/audit/hdfs/dn/hdfs-audit-datanode.log /var/log/Bigdata/audit/hdfs/jn/hdfs-audit-journalnode.log /var/log/Bigdata/audit/hdfs/zkfc/hdfs-audit-zkfc.log /var/log/Bigdata/audit/hdfs/httpfs/hdfs-audit-httpfs.log /var/log/Bigdata/audit/hdfs/router/hdfs-audit-router.log
Mapreduce	/var/log/Bigdata/audit/mapreduce/jobhistory/mapred-audit-jobhistory.log
Hive	/var/log/Bigdata/audit/hive/hiveserver/hive-audit.log /var/log/Bigdata/audit/hive/metastore/metastore-audit.log /var/log/Bigdata/audit/hive/webhcat/webhcat-audit.log
Loader	/var/log/Bigdata/loader/audit/default.audit
Hue	/var/log/Bigdata/audit/hue/hue-audits.log
ZooKeeper	/var/log/Bigdata/audit/zookeeper/quorumpeer/zk-audit-quorumpeer.log

组件名称	审计日志目录
Spark	/var/log/Bigdata/audit/spark/jdbcserver/jdbcserver-audit.log /var/log/Bigdata/audit/spark/jobhistory/jobhistory-audit.log
Yarn	/var/log/Bigdata/audit/yarn/rm/yarn-audit-resourcemanager.log /var/log/Bigdata/audit/yarn/nm/yarn-audit-nodemanager.log
Storm	/var/log/Bigdata/audit/storm/nimbus/audit.log /var/log/Bigdata/audit/storm/ui/audit.log

12.6.2 Manager 日志清单

日志描述

日志存储路径：Manager相关日志的默认存储路径为“/var/log/Bigdata/*Manager组件*”。

- ControllerService：/var/log/Bigdata/controller/（OMS安装、运行日志）
- Httpd：/var/log/Bigdata/httpd（httpd安装、运行日志）
- logman：/var/log/Bigdata/logman（日志打包工具日志）
- NodeAgent：/var/log/Bigdata/nodeagent（NodeAgent安装、运行日志）
- okerberos：/var/log/Bigdata/okerberos（okerberos安装、运行日志）
- oldapserver：/var/log/Bigdata/oldapserver（oldapserver安装、运行日志）
- MetricAgent：/var/log/Bigdata/metric_agent（MetricAgent运行日志）
- omm：/var/log/Bigdata/omm（omm安装、运行日志）
- timestamp：/var/log/Bigdata/timestamp（NodeAgent启动时间日志）
- tomcat：/var/log/Bigdata/tomcat（Web进程日志）
- patch：/var/log/Bigdata/patch（补丁安装日志）
- Sudo：/var/log/Bigdata/sudo（sudo脚本执行日志）
- OS：/var/log/*message文件*（OS系统日志）
- OS Performance：/var/log/osperf（OS性能统计日志）
- OS Statistics：/var/log/osinfo/statistics（OS参数配置信息日志）

日志归档规则：

Manager的日志启动了自动压缩归档功能，缺省情况下，当日志大小超过10MB的时候，会自动压缩，压缩后的日志文件名规则为：“<原有日志名>-<yyyy-mm-dd_hh-mm-ss>.[编号].log.zip”。最多保留最近的20个压缩文件。

表 12-23 Manager 日志列表

日志类型	日志文件名	描述
Controller运行日志	controller.log	记录组件安装、升级、补丁、配置、监控、告警和日常运维操作日志。
	controller_client.log	Rest接口运行日志。
	acs.log	Acs运行日志。
	acs_spnego.log	acs中spnego用户日志
	aos.log	Aos运行日志。
	plugin.log	Aos插件日志
	backupplugin.log	备份恢复进程运行日志
	controller_config.log	配置运行日志
	controller_nodesetup.log	Controller加载任务日志
	controller_root.log	Controller进程系统日志
	controller_trace.log	Controller与NodeAgent之间RPC通信日志
	controller_monitor.log	监控日志
	controller_fsm.log	状态机日志
	controller_alarm.log	Controller发送告警日志
	controller_backup.log	Controller备份恢复日志
	install.log，distributeAdapterFiles.log，install_os_optimization.log	oms安装日志
	oms_ctl.log	oms启停日志
	installntp.log	ntp安装日志
	modify_manager_param.log	修改Manager参数日志
	backup.log	OMS备份脚本运行日志
	supressionAlarm.log	告警脚本运行日志
	om.log	生成om证书日志
	backupplugin_ctl.log	备份恢复插件进程启动日志
	getLogs.log	采集日志脚本运行日志

日志类型	日志文件名	描述
	backupAuditLogs.log	审计日志备份脚本运行日志
	certStatus.log	证书定期检查日志
	distribute.log	证书分发日志
	ficertgenetrate.log	证书替换日志，包括生成二级证书、cas证书、httpd证书的日志。
	genPwFile.log	生成证书密码文件日志
	modifyproxyconf.log	修改HTTPD代理配置的日志
	importTar.log	证书导入信任库日志
Httpd	install.log	Httpd安装日志
	access_log, error_log	Httpd运行日志
logman	logman.log	日志打包工具日志。
NodeAgent	install.log，install_os_optimization.log	NodeAgent安装日志
	installntp.log	ntp安装日志
	start_ntp.log	ntp启动日志
	ntpChecker.log	ntp检查日志
	ntpMonitor.log	ntp监控日志
	heartbeat_trace.log	NodeAgent与Controller心跳日志
	alarm.log	告警日志
	monitor.log	监控日志
	nodeagent_ctl.log，start-agent.log	NodeAgent启动日志
	agent.log	NodeAgent运行日志
	cert.log	证书日志
	agentplugin.log	监控agent侧插件运行日志
	omaplugin.log	OMA插件运行日志
	diskhealth.log	磁盘健康检查日志
	supressionAlarm.log	告警脚本运行日志

日志类型	日志文件名	描述
	updateHostFile.log	更新主机列表日志
	collectLog.log	节点日志采集脚本运行日志
	host_metric_collect.log	主机指标采集运行日志
	checkfileconfig.log	文件权限配置检查运行日志
	entropycheck.log	熵值检查运行日志
	timer.log	节点定时调度日志
	pluginmonitor.log	组件监控插件日志
	agent_alarm_py.log	NodeAgent检查文件权限发送告警日志
okerberos	addRealm.log，modifyKerberosRealm.log	切域日志
	checkservice_detail.log	Okerberos健康检查日志
	genKeytab.log	生成keytab日志
	KerberosAdmin_genConfigDetail.log	启动kadmin进程时，生成kadmin.conf的运行日志
	KerberosServer_genConfigDetail.log	启动krb5kdc进程时，生成krb5kdc.conf的运行日志
	oms-kadmind.log	kadmin进程的运行日志
	oms_kerberos_install.log，postinstall_detail.log	okerberos安装日志
	oms-krb5kdc.log	krbkdc运行日志
	start_detail.log	okerberos启动日志
	realmDataConfigProcess.log	切域失败，回滚日志
	stop_detail.log	okerberos停止日志
oldapserver	ldapserver_backup.log	Oldapserver备份日志
	ldapserver_chk_service.log	Oldapserver健康检查日志
	ldapserver_install.log	Oldapserver安装日志
	ldapserver_start.log	Oldapserver启动日志

日志类型	日志文件名	描述
	ldapserver_status.log	Oldapserver进程状态检查日志。
	ldapserver_stop.log	Oldapserver停止日志
	ldapserver_wrap.log	Oldapserver服务管理日志。
	ldapserver_uninstall.log	Oldapserver卸载日志
	restart_service.log	Oldapserver重启日志
	ldapserver_unlockUser.log	记录解锁Ldap用户和管理账户的日志
omm	omsconfig.log	OMS配置日志
	check_oms_heartbeat.log	OMS心跳运行日志
	monitor.log	OMS监控日志
	ha_monitor.log	HA_Monitor操作日志
	ha.log	HA操作日志
	fms.log	告警日志
	fms_ha.log	告警的HA监控日志
	fms_script.log	告警控制日志
	config.log	告警配置日志
	iam.log	IAM日志
	iam_script.log	IAM控制日志
	iam_ha.log	IAM的HA监控日志
	config.log	IAM配置日志
	operatelog.log	IAM操作日志
	heartbeatcheck_ha.log	OMS心跳的HA监控日志
	install_oms.log	OMS安装日志
	pms_ha.log	监控的HA监控日志
	pms_script.log	监控控制日志
	config.log	监控配置日志
	plugin.log	监控插件运行日志
	pms.log	监控日志
	ha.log	HA运行日志

日志类型	日志文件名	描述
	cep_ha.log	CEP的HA监控日志
	cep_script.log	CEP控制日志
	cep.log	CEP日志
	config.log	CEP配置日志
	omm_gaussdba.log	gaussdb的HA监控日志
	gaussdb-<SERIAL>.log	gaussdb运行日志
	gs_ctl-<DATE>.log	gaussdb控制日志的归档日志
	gs_ctl-current.log	gaussdb控制日志
	gs_guc-current.log	gaussdb操作日志
	encrypt.log	omm加密日志
	omm_agent_ctl.log	OMA控制日志
	oma_monitor.log	OMA监控日志
	install_oma.log	OMA安装日志
	config_oma.log	OMA配置日志
	omm_agent.log	OMA运行日志
	acs.log	acs资源日志。
	aos.log	aos资源日志
	controller.log	controller资源日志
	feed_watchdog.log	feed_watchdog资源日志
	floatip.log	floatip资源日志
	ha_ntp.log	ntp资源日志
	httpd.log	httpd资源日志
	okerberos.log	okerberos资源日志
	oldap.log	oldap资源日志
	tomcat.log	tomcat资源日志
	send_alarm.log	管理节点HA告警发送脚本运行日志
timestamp	restart_stamp	NodeAgent启动时间
tomcat	cas.log，localhost_access_cas_log.log	cas运行日志

日志类型	日志文件名	描述
	catalina.log，catalina.out，host-manager.log，localhost.log，manager.log	tomcat运行日志
	localhost_access_web_log.log	记录访问FusionInsight Manager系统REST接口的日志
	web.log	web进程运行日志
	northbound_ftp_sftp.log，snmp.log	北向日志
watchdog	watchdog.log，feed_watchdog.log	watchdog.log运行日志
patch	oms_installPatch.log	OMS补丁安装日志
	agent_installPatch.log	Agent补丁安装日志
	agent_uninstallPatch.log	agent补丁卸载日志
	NODE_AGENT_restoreFile.log	agent补丁恢复文件日志
	NODE_AGENT_updateFile.log	agent补丁更新文件日志
	OMA_restoreFile.log	OMA补丁恢复文件日志
	OMA_updateFile.log	OMA补丁更新文件日志
	CONTROLLER_restoreFile.log	CONTROLLER补丁恢复文件日志
	CONTROLLER_updateFile.log	CONTROLLER补丁更新文件日志
	OMS_restoreFile.log	OMS补丁恢复文件日志
	oms_uninstallPatch.log	OMS补丁卸载日志
	OMS_updateFile.log	OMS补丁更新文件日志
	createStackConf.log，decompress.log，decompress_OMS.log ，distrExtractPatchOnOMS.log，slimReduction.log，switch_adapter.log	补丁安装日志
sudo	sudo.log	sudo脚本执行日志

日志级别

Manager中提供了如表12-24所示的日志级别。日志级别优先级从高到低分别是FATAL、ERROR、WARN、INFO、DEBUG。程序会打印高于或等于所设置级别的日志，设置的日志等级越高，打印出来的日志就越少。

表 12-24 日志级别

级别	描述
FATAL	FATAL表示当前事件处理出现严重错误信息，可能导致系统崩溃。
ERROR	ERROR表示当前事件处理出现错误信息，系统运行出错。
WARN	WARN表示当前事件处理存在异常信息，但认为是正常范围，不会导致系统出错。
INFO	INFO表示记录系统及各事件正常运行状态信息
DEBUG	DEBUG表示记录系统及系统的调试信息。

日志格式

Manager的日志格式如下所示：

表 12-25 日志格式

日志类型	组件	格式	示例
Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	Controller，Httpd，logman，NodeAgent，okerberos，oldapserver，omm，tomcat，upgrade	<yyyy-MM-dd HH:mm:ss,SSS>|<Log Level>|<产生该日志的线程名字>|<log中的message>|<日志事件的发生位置>	2015-06-30 00:37:09,067 INFO [pool-1-thread-1] Completed Discovering Node. com.xxx.hadoop.om.controller.tasks.nodesetup.DiscoverNodeTask.execute(DiscoverNodeTask.java:299)

12.6.3 查看及导出审计日志

操作场景

该任务指导用户在MRS Manager查看、导出审计日志工作，用于安全事件中事后追溯、定位问题原因及划分事故责任。

系统记录的日志信息包含：

- 用户活动信息，如用户登录与注销，系统用户信息变更，系统用户组信息变更等。
- 用户操作指令信息，如集群的启动、停止，软件升级等。

操作步骤

- 查看审计日志
  a. 在MRS Manager，单击“审计管理”，可直接查看默认的审计日志。
     若审计日志的审计内容长度大于256字符，请单击审计日志展开按钮展开审计详情。
     - 默认以“产生时间”列按降序排列，单击**操作类型**、**安全级别**、**产生时间**、**用户**、**主机**、**服务**、**实例或操作结果**可修改排列方式。
     - 支持在“安全级别”筛选相同级别的全部告警。结果包含已清除和未清除的告警。

     导出的审计日志文件，包含以下信息列：
     - “编号”：表示MRS Manager已生成的审计日志数量，每增加一条审计日志则编号自动加1。
     - “操作类型”：表示用户操作的操作类型，分为“告警”、“审计日志”、“备份恢复”、“集群”、“采集日志”、“主机”、“服务”、“多租户”和“用户管理”九种场景，其中“用户管理”仅在启用了Kerberos认证的集群中支持。每个场景中包含不同操作类型，例如“告警”中包含“导出告警”，“集群”中包含“启动集群”，“多租户”包含“增加租户”等。
     - “安全级别”：表示每条审计日志的安全级别，包含“高危”、“危险”、“一般”和“提示”四种。
     - “开始时间”：表示用户操作开始的时间，且时间为CET或CEST时间。
     - “结束时间”：表示用户操作结束的时间，且时间为CET或CEST时间。
     - “用户IP”：表示用户操作时所使用的IP地址。
     - “用户”：表示执行操作的用户名。
     - “主机”：表示用户操作发生在集群的哪个节点。如果操作不涉及节点则不保存信息。
     - “服务”：表示用户操作发生在集群的哪个服务。如果操作不涉及服务则不保存信息。
     - “实例”：表示用户操作发生在集群的哪个角色实例。如果操作不涉及角色实例则不保存信息。
     - “操作结果”：表示用户操作的结果，包含“成功”、“失败”和“未知”。
     - “内容”：表示用户操作的具体执行信息。
  b. 单击“高级搜索”，在审计日志搜索区域中，设置查询条件，单击“搜索”，查看指定类型的审计日志。单击“重置”清除输入的搜索条件。

说明

“开始时间”和“结束时间”表示时间范围的开始时间和结束时间，可以搜索此时间段内产生的告警。

- 导出审计日志
  a. 在审计日志列表中，单击“导出全部”，导出所有的日志。
  b. 在审计日志列表中，勾选日志信息前的复选框，单击“导出”，导出指定日志。

12.6.4 导出服务日志

操作场景

该任务指导用户从MRS Manager导出各个服务角色生成的日志。

前提条件

- 用户已经获取账号对应的Access Key ID（AK）和Secret Access Key（SK）。
- 用户已经在账号的对象存储服务（OBS）中创建了并行文件系统。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 单击“维护”下方的“日志导出”。

步骤3 “服务”选择服务，“主机”填写服务所部署主机的IP，“开始时间”与“结束时间”选择对应的开始与结束时间。

步骤4 在“导出类型”选择一个日志保存的位置。只有启用了Kerberos认证的集群支持选择。

- “下载到本地”：表示将日志保存到用户当前的本地环境。然后执行**步骤8**。
- “上传到OBS”：表示将日志保存到OBS中。默认值。然后执行**步骤5**。

步骤5 在“OBS路径”填写服务日志在OBS保存的路径。

需要填写完整路径且不能以“/”开头，路径可以不存在，系统将自动创建。OBS的完整路径最大支持900个字节。

步骤6 在“桶名”输入已创建的OBS文件系统名称。

步骤7 在“AK”和“SK”输入用户的Access Key ID和Secret Access Key。

步骤8 单击“确定”完成日志下载。

----结束

12.6.5 配置审计日志转储参数

操作场景

Manager的审计日志长期保留在数据库中，可能引起数据目录的磁盘空间不足问题，故通过设置转储参数及时将审计日志自动存储到指定服务器的指定目录下，便于管理审计日志信息。

若用户未配置审计日志转储，当审计日志达到十万条，系统自动将这十万条审计日志保存到文件中。保存路径为主管理节点“${BIGDATA_HOME}/OMSV100R001C00x8664/workspace/conf/data/operatelog”，保存的文件名格式为“OperateLog_store_YY_MM_DD_HH_MM_SS.csv”，保存的审计日志历史文件数最大为50。第一次转存审计日志时自动生成目录。

前提条件

转储服务器对应的弹性云服务器需要和MRS集群的Master节点在相同的VPC，且Master节点可以访问转储服务器的IP地址和指定端口。转储服务器的SFTP服务正常。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“维护”下单击“审计日志转储配置”。

表 12-26 审计日志转储参数

参数名	参数值	参数解释
审计日志转储	• 打开 • 关闭	必选参数，指定是否打开审计日志转储 • 打开：启用审计日志转储 • 关闭：禁用审计日志转储
转储模式	• 按数量 • 按时间	必选参数，指定转储模式 • “按数量”：日志到达指定条数（默认10万条）时开始转储 • “按时间”：指定某一日期开始转储
SFTP IP	192.168.10.51（举例）	必选参数，指定审计日志转储后存放的SFTP服务器
SFTP端口	22（举例）	必选参数，指定审计日志转储后存放的SFTP服务器连接端口
保存路径	/opt/omm/oms/auditLog（举例）	必选参数，指定SFTP服务器上保存审计日志的路径
SFTP用户名	root（举例）	必选参数，指定登录SFTP服务器的用户名
SFTP密码	Root_123（举例）	必选参数，指定登录SFTP服务器的密码
SFTP公共密钥	-	可选参数，指定SFTP服务器的公共密钥，建议配置SFTP的公共密钥，否则可能存在安全风险。

参数名	参数值	参数解释
转储日期	11月06（举例）	必选参数，指定选择“按时间”转储模式时，指定转储日期，系统将在此日期开始转储。转储的日志范围为当前年份1月1日0时之前的所有审计日志。

说明

审计日志转储文件关键字段参考：

- “USERTYPE”表示用户类型，“0”表示“人机”用户，“1”表示“机机”用户。
- “LOGLEVEL”表示安全级别，“0”表示高危，“1”表示危险，“2”表示一般，“3”表示提示。
- “OPERATERESULT”表示操作结果，“0”表示成功，“1”表示失败。

----结束

12.7 健康检查管理

12.7.1 执行健康检查

操作场景

该任务指导用户在日常运维中完成集群进行健康检查的工作，以保证集群各项参数、配置以及监控没有异常、能够长时间稳定运行。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

操作步骤

- 手动执行所有服务的健康检查
  a. 单击“服务管理”。
  b. 选择“更多 > 启动服务健康检查”，启动服务健康检查。

说明

- 集群健康检查包含了Manager、服务与主机状态的检查。
- 在MRS Manager界面，选择“系统设置 > 维护 > 健康检查 > 集群健康检查”，也可以执行集群健康检查。
- 手动执行健康检查的结果可直接在检查列表左上角单击“导出报告”，选择导出结果。

- 手动执行单个服务的健康检查
  a. 选择“服务管理”，在服务列表中单击服务指定名称。
  b. 选择“更多 > 启动服务健康检查”启动指定服务健康检查。
- 手动执行主机健康检查
  a. 单击“主机管理”。
  b. 勾选待检查主机前的复选框。
  c. 选择“更多 > 启动主机健康检查”启动指定主机健康检查。
- 自动执行健康检查
  a. 单击“系统设置”。
  b. 单击“维护”下方的“健康检查”。
  c. 单击“健康检查配置”，配置自动执行健康检查。

     “健康检查报告文件最大份数”：配置健康检查报告文件最大份数。该值必须是整数，取值范围1～100。

     “定期健康检查”配置是否启用自动执行健康检查。“定期健康检查”的开关默认为关闭，单击可启用该功能，根据管理需要选择“每天”、“每周”或“每月”。
  d. 单击“确定”保存配置。系统右上角弹出提示“健康检查配置保存成功。”。

12.7.2 查看并导出检查报告

操作场景

为了满足对健康检查结果的进一步具体分析，您可以在MRS Manager中查看以及导出健康检查的结果。

说明

系统健康检查的范围包含Manager、服务级别和主机级别的健康检查：

- Manager关注集群统一管理平台是否提供管理功能。
- 服务级别关注组件是否能够提供正常的服务。
- 主机级别关注主机的一系列指标是否正常。

系统健康检查可以包含三方面检查项：各检查对象的“健康状态”、相关的告警和自定义的监控指标，检查结果并不能等同于界面上显示的“健康状态”。

前提条件

已执行健康检查。

操作步骤

步骤1 单击“服务管理”。

步骤2 选择“更多 > 查看集群健康检查报告”，查看集群健康检查的报告。

步骤3 在健康检查的报告面板上单击“导出报告”导出健康检查报告，可查看检查项的完整信息。

说明

对于存在问题的检查项，请参见**DBService健康检查指标项说明**~**ZooKeeper健康检查指标项说明**进行修复。

----结束

12.7.3 配置健康检查报告保存数

操作场景

在不同时间、不同使用场景下，MRS集群、服务和主机产生的健康检查报告结果不完全相同。如果需要保存更多的报告用于比较时，可以在MRS Manager修改健康检查报告保存的文件数。

健康检查报告保存的文件数不区分集群、服务或主机类型的健康检查报告。健康检查完成后，报告文件默认保存在主管理节点的“$BIGDATA_DATA_HOME/Manager/healthcheck”，备管理节点将自动同步。

前提条件

用户已明确业务需求，并规划好保存的时间跨度与健康检查频率，检查主备管理节点磁盘空间使用率。

操作步骤

步骤1 选择“系统设置 > 维护 > 健康检查 > 健康检查配置”。

步骤2 “健康检查报告文件最大份数”参数填写健康检查报告的保存个数。默认值为“50”，取值范围为1~100。

步骤3 单击“确定”保存配置。系统右上角弹出提示“健康检查配置保存成功。”。

----结束

12.7.4 管理健康检查报告

操作场景

用户可以在MRS Manager对已保存的历史健康检查报告进行管理，即查看、下载和删除历史健康检查报告。

操作步骤

- 下载指定的健康检查报告
  a. 选择“系统设置 > 维护 > 健康检查”。

b. 在目标健康检查报告所在行，单击“下载”，下载报告文件。

- 批量下载指定的健康检查报告
  a. 选择“系统设置 > 维护 > 健康检查”。
  b. 勾选多个目标健康检查报告，单击“下载文件”，下载多个报告文件。
- 删除指定的健康检查报告
  a. 选择“系统设置 > 维护 > 健康检查”。
  b. 在目标健康检查报告所在行，单击“删除”，删除报告文件。
- 批量删除指定的健康检查报告
  a. 选择“系统设置 > 维护 > 健康检查”。
  b. 勾选多个目标健康检查报告，单击“删除文件”，删除多个报告文件。

12.7.5 DBService 健康检查指标项说明

服务健康检查

指标项名称：服务状态

指标项含义：检查DBService服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-27001进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.6 Flume 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查Flume服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-24000进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.7 HBase 健康检查指标项说明

运行良好的 RegionServer 数

指标项名称：运行良好的RegionServer数

指标项含义：检查HBase集群中运行良好的RegionServer数。

恢复指导：如果该指标项异常，请检查RegionServer的状态是否正常并处理，然后建议检查网络是否正常。

服务健康状态

指标项名称：服务状态

指标项含义：检查HBase服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，请检查HMaster和RegionServer的状态是否正常并先处理，然后检查ZooKeeper服务的状态是否为故障并处理。使用客户端，确认是否可以正确读取HBase表中的数据，排查读数据失败的原因。最后参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.8 Host 健康检查指标项说明

Swap 使用率

指标项名称：Swap使用率

指标项含义：系统Swap使用率，计算方法：已用Swap大小/总共Swap大小。当前阈值设置为75.0%，如果使用率超过阈值，则认为不健康。

恢复指导：

1. 确认节点Swap使用率。

   登录检查结果不健康的节点，执行**free -m**查看swap总量和已使用量，如果swap使用率已超过阈值，则执行2。
2. 如果Swap使用率超过阈值，建议对系统进行扩容，如：增加节点。

主机文件句柄使用率

指标项名称：主机文件句柄使用率

指标项含义：系统中的文件句柄的使用率，主机文件句柄使用率=已用句柄数/总共句柄数。如果使用率超过阈值，则认为不健康。

恢复指导：

1. 确认主机文件句柄使用率。

   登录检查结果不健康的节点，执行**cat /proc/sys/fs/file-nr**，输出结果的第一列和第三列分别表示系统已使用的句柄数和总句柄数，如果使用率超过阈值，则执行2。
2. 如果主机文件句柄使用率超过阈值，建议对系统进行排查，具体分析文件句柄的使用情况。

NTP 偏移量

指标项名称：NTP偏移量

指标项含义：NTP时间偏差。如果时间偏差超过阈值，则认为不健康。

恢复指导：

1. 确认NTP时间偏差。

   登录检查结果不健康的节点，执行**/usr/sbin/ntpq -np**查看信息，其中offset列表示时间偏差。如果时间偏差大于阈值，则执行**2**。
2. 如果该指标项异常，则需要检查时钟源配置是否正确，请联系运维人员处理。

平均负载

指标项名称：平均负载

指标项含义：系统平均负载，表示特定时间段内运行队列中的平均进程数。这里系统平均负载是通过uptime命令中得到的负载值计算得到。计算方法：（1分钟负载 + 5分钟负载 + 15分钟负载）/(3*CPU个数)。当前阈值设置为2，如果超过阈值，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点，执行**uptime**命令，命令输出的最后三列分别表示1分钟负载、5分钟负载和15分钟负载。根据系统平均负载的计算方法，如果负载超过阈值，则执行**2**。
2. 如果系统平均负载超过阈值，建议对系统进行扩容，如增加节点等。

D 状态进程

指标项名称：D状态进程

指标项含义：不可中断的睡眠进程，即D状态进程。D状态通常是进程在等待IO，比如磁盘IO，网络IO等，但是此时IO出现异常。如果系统中出现D状态进程，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12028进行处理。

硬件状态

指标项名称：硬件状态

指标项含义：检查系统硬件状态，包括CPU、内存、磁盘、电源、风扇等。该检查项通过ipmitool sdr elist获取相关硬件信息。如果相关硬件状态异常，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点。执行**ipmitool sdr elist**查看系统硬件状态，命令输出的最后一列表示对应的硬件状态。如果提示的状态在下面的故障描述表中，则任务不健康。

模块	故障描述
Processor	IERR Thermal Trip FRB1/BIST failure FRB2/Hang in POST failure FRB3/Processor startup/init failure Configuration Error SM BIOS Uncorrectable CPU-complex Error Disabled Throttled Uncorrectable machine check exception
Power Supply	Failure detected Predictive failure Power Supply AC lost AC lost or out-of-range AC out-of-range, but present Config Error: Vendor Mismatch Config Error: Revision Mismatch Config Error: Processor Missing Config Error: Power Supply Rating Mismatch Config Error: Voltage Rating Mismatch Config Error
Power Unit	240VA power down Interlock power down AC lost Soft-power control failure Failure detected Predictive failure
Memory	Uncorrectable ECC Parity Memory Scrub Failed Memory Device Disabled Correctable ECC logging limit reached Configuration Error Throttled Critical Overtemperature

模块	故障描述
Drive Slot	Drive Fault Predictive Failure Parity Check In Progress In Critical Array In Failed Array Rebuild In Progress Rebuild Aborted
Battery	Low Failed

2. 如果该指标项异常，建议联系运维人员解决处理。

主机名

指标项名称：主机名

指标项含义：检查是否设置了主机名。如果没有设置主机名，则认为不健康。如果该指标项异常，建议正确设置hostname。

恢复指导：

1. 登录检查结果不健康的节点。
2. 执行以下命令修改主机名，使节点主机名与规划的主机名保持一致：

   hostname *主机名*。例如，将主机名改为“Bigdata-OM-01”，请执行命令hostname Bigdata-OM-01。
3. 修改主机名配置文件。

   执行**vi /etc/HOSTNAME**命令编辑文件，修改文件内容为“Bigdata-OM-01”，并保存退出。

Umask

指标项名称：Umask

指标项含义：检查omm用户的umask设置是否正确。如果umask设置不等于0077，则认为不健康。

恢复指导：

1. 如果该指标异常，建议将omm用户的umask设置为0077。登录检查结果不健康的节点，执行**su - omm**切换到omm用户。
2. 执行**vi ${BIGDATA_HOME}/.om_profile**，修改**umask=0077**，保存并退出。

OMS 的 HA 状态

指标项名称：OMS的HA状态

指标项含义：检查OMS的双机资源是否正常。OMS双机资源状态的详细信息可使用**${CONTROLLER_HOME}/sbin/status-oms.sh**查看。如果有模块状态异常，认为不健康。

恢复指导：

1. 登录主管理节点，执行**su - omm**切换到**omm**用户，然后执行**${CONTROLLER_HOME}/sbin/status-oms.sh**查看OMS状态。
2. 如果floatip、okerberos、oldap等异常，可参见告警ALM-12002、ALM-12004、ALM-12005分别进行处理。
3. 如果是其他资源异常，建议查看相关异常模块的日志。

   controller资源异常：查看异常节点的/var/log/Bigdata/controller/controller.log。

   cep资源异常：查看异常节点的/var/log/Bigdata/omm/oms/cep/cep.log。

   aos资源异常：查看异常节点的/var/log/Bigdata/controller/aos/aos.log。

   feed_watchdog资源异常：查看异常节点的/var/log/Bigdata/watchdog/watchdog.log。

   httpd资源异常：查看异常节点的/var/log/Bigdata/httpd/error_log。

   fms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/fms/fms.log。

   pms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/pms/pms.log。

   iam资源异常：查看异常节点的/var/log/Bigdata/omm/oms/iam/iam.log。

   gaussDB资源异常：查看异常节点的/var/log/Bigdata/omm/oms/db/omm_gaussdba.log。

   ntp资源异常：查看异常节点的/var/log/Bigdata/omm/oms/ha/scriptlog/ha_ntp.log。

   tomcat资源异常：查看异常节点的/var/log/Bigdata/tomcat/catalina.log。
4. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

安装目录及数据目录检查

指标项名称：安装目录及数据目录检查

指标项含义：该指标项首先检查安装目录（默认为“/opt/Bigdata”）所在磁盘分区根目录下的lost+found目录。如果该目录下有**omm**用户的文件，则认为异常。节点异常时，会把相关的文件放入到“lost+found”目录。该检查主要是针对这类场景，检查文件是否丢失。然后，对安装目录（如：“/opt/Bigdata”）和数据目录（如：“/srv/BigData”）进行检查。如果目录下出现非**omm**用户的文件，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点，执行**su - omm**切换到**omm**用户。检查lost+found目录是否存在omm用户的文件或文件夹。

   如果有**omm**用户文件，建议对其进行恢复后重新检查；如果没有**omm**用户文件，则执行**2**。
2. 分别对安装目录和数据目录进行排查。查看目录下是否存在非**omm**用户是文件或文件夹。如果确认这些文件是手工生成的临时文件，建议对清理后重新检查。

CPU 使用率

指标项名称：CPU使用率

指标项含义：检查CPU使用率是否超过当前设定的阈值。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12016进行处理。

内存使用率

指标项名称：内存使用率

指标项含义：检查内存使用率是否超过当前设定的阈值。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12018进行处理。

主机磁盘使用率

指标项名称：主机磁盘使用率

指标项含义：检查主机磁盘使用率是否超过当前设定的阈值。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，系统中会产生对应的告警，建议参见告警ALM-12017进行处理。

主机磁盘写速率

指标项名称：主机磁盘写速率

指标项含义：检查主机磁盘写速率。根据业务场景不同，主机磁盘写速率大小可能存在差异，所以该指标项只反映具体的数值大小，用户需根据业务场景具体判断该指标是否健康。

恢复指导：用户根据具体的业务场景，判断当前磁盘写速率是否正常。

主机磁盘读速率

指标项名称：主机磁盘读速率

指标项含义：检查主机磁盘读速率。根据业务场景不同，主机磁盘读速率大小可能存在差异，所以该指标项只反映具体的数值大小，用户需根据业务场景具体判断该指标是否健康。

恢复指导：用户根据具体的业务场景，判断当前磁盘读速率是否正常。

主机业务平面网络状态

指标项名称：主机业务平面网络状态

指标项含义：检查集群主机业务平面网络连通性。如果出现无法连通的情况，则认为不健康。

恢复指导：如果是单平面组网，对应需检查单平面的IP。双平面组网排查恢复步骤如下：

1. 检查主备管理节点业务平面IP的网络连通性。
   如果网络异常，执行**3**。
   如果网络正常，执行**2**。

2. 检查主管理节点IP到集群内异常节点IP的网络连通性。
3. 如果网络不通，请联系运维人员排查网络问题，以保证满足业务使用。

主机状态

指标项名称：主机状态

指标项含义：检查主机状态是否正常。如果节点有故障，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-12006进行处理。

检查告警

指标项名称： 检查告警

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.9 HDFS 健康检查指标项说明

发送包的平均时间统计

指标项名称：发送包的平均时间统计

指标项含义：HDFS文件系统中DataNode每次执行SendPacket的平均时间统计，如果大于2000000纳秒，则认为不健康。

恢复指导： 如果该指标项异常，则需要检查集群的网络速度是否正常、内存或CPU使用率是否过高。同时检查集群中HDFS负载是否过高。

服务健康状态

指标项名称：服务状态

指标项含义：检查HDFS服务状态是否正常。如果节点有故障，则认为不健康。

恢复指导：如果该指标项异常，建议检查KrbServer、LdapServer、ZooKeeper三个服务的状态是否为异常并处理。然后再检查是否是HDFS SafeMode ON导致的写文件失败，并使用客户端，确认是否无法在HDFS中写入数据，排查HDFS写数据失败的原因。最后参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查HDFS服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，请参见告警进行修复。

12.7.10 Hive 健康检查指标项说明

HiveServer 允许的最大 session 数量

指标项名称：Hive允许连接的最大session数量

指标项含义：检查Hive允许连接的最大session数量。

恢复指导：如果该指标项异常，建议参见告警进行处理。

已经连接到 HiveServer 的 session 数量

指标项名称：已经连接到HiveServer的session数量

指标项含义：检查Hive连接数。

恢复指导：如果该指标项异常，建议参见告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Hive服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查主机是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.11 Kafka 健康检查指标项说明

Broker 可用节点数

标项名称：Broker数目

指标项含义：检查集群中可用的Broker节点数，若集群中可用的Broker节点数小于2，则认为不健康。

恢复指导：如果该指标项异常，进入Kafka服务实例页面，单击不可用Broker实例的“主机名”，在“概要信息”中查看主机的健康状态，若为“良好”，则参见“进程故障”告警进行处理；若不为“良好”，则参见“节点故障”告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Kafka服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见“Kafka服务不可用”告警进行处理。

检查告警

指标项名称：告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.12 KrbServer 健康检查指标项说明

KerberosAdmin 服务可用性检查

指标项名称：KerberosAdmin服务可用性

指标项含义：系统对KerberosAdmin服务状态进行检查，如果检查结果不正常，则KerberosAdmin服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是KerberosAdmin服务所在节点故障，或者SlapdServer服务不可用。操作人员进行KerberosAdmin服务恢复时，请尝试如下操作：

1. 检查KerberosAdmin服务所在节点是否故障。
2. 检查SlapdServer服务是否不可用。

KerberosServer 服务可用性检查

指标项名称： KerberosServer服务可用性

指标项含义：系统对KerberosServer服务状态进行检查，如果检查结果不正常，则KerberosServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是KerberosServer服务所在节点故障，或者SlapdServer服务不可用。操作人员进行KerberosServer服务恢复时，请尝试如下操作：

1. 检查KerberosServer服务所在节点是否故障。
2. 检查SlapdServer服务是否不可用。

服务健康状态

指标项名称：服务状态

指标项含义：系统对KrbServer服务状态进行检查，如果检查结果不正常，则KrbServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是KrbServer服务所在节点故障或者LdapServer服务不可用。详细操作请参见告警ALM-25500处理。

检查告警

指标项名称：告警信息

指标项含义：系统对KrbServer服务的告警信息进行检查。如果存在告警信息，则KrbServer服务可能存在异常。

恢复指导：如果该指标项检查结果不正常，建议根据告警内容，查看对应的告警资料，并进行相应的处理。

12.7.13 LdapServer 健康检查指标项说明

SlapdServer 服务可用性检查

指标项名称：SlapdServer服务可用性

指标项含义：系统对SlapdServer服务状态进行检查。如果检查结果不正常，则SlapdServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是SlapdServer服务所在节点故障或者SlapdServer进程故障。操作人员进行SlapdServer服务恢复时，请尝试如下操作：

1. 检查SlapdServer服务所在节点是否故障。详细操作请参见告警ALM-12006处理。
2. 检查SlapdServer进程是否正常。详细操作请参见告警ALM-12007处理。

服务健康状态

指标项名称：服务状态

指标项含义：系统对LdapServer服务状态进行检查。如果检查结果不正常，则LdapServer服务不可用。

恢复指导：如果该指标项检查结果不正常，原因可能是主LdapServer服务所在节点故障或者主LdapServer进程故障。详细操作请参见告警ALM-25000处理。

检查告警

指标项名称：告警信息

指标项含义：系统对LdapServer服务的告警信息进行检查。如果存在告警信息，则LdapServer服务可能存在异常。

恢复指导：如果该指标项检查结果不正常，建议根据告警内容，查看对应的告警资料，并进行相应的处理。

12.7.14 Loader 健康检查指标项说明

ZooKeeper 健康状态

指标项名称：ZooKeeper健康状态

指标项含义：检查ZooKeeper健康状态是否正常。如果ZooKeeper服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

HDFS 健康状态

指标项名称：HDFS健康状态

指标项含义：检查HDFS健康状态是否正常。如果HDFS服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

DBService 健康状态

指标项名称：DBService健康状态

指标项含义：检查DBService健康状态是否正常。如果DBService服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

Yarn 健康状态

指标项名称：Yarn健康状态

指标项含义：检查Yarn健康状态是否正常。如果Yarn服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

MapReduce 健康状态

指标项名称：MapReduce健康状态

指标项含义：检查MapReduce健康状态是否正常。如果MapReduce服务状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

Loader 进程状态

指标项名称：Loader进程状态

指标项含义：检查Loader进程状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Loader服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

告警检查

指标项名称：告警信息

指标项含义：检查Loader服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.15 MapReduce 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查MapReduce服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.16 OMS 健康检查指标项说明

OMS 状态检查

指标项名称：OMS状态检查

指标项含义：OMS状态检查包括HA状态检查和资源状态检查。 HA状态取值为active、standby和NULL，分别表示主节点、备节点和未知。资源状态取值为normal、abnormal和NULL，分别表示正常、异常和未知。HA状态为NULL时，认为不健康；资源状态为NULL或abnormal时，认为不健康。

表 12-27 OMS 状态说明表

名称	说明
HA状态	active表示主节点 standby表示备节点 NULL表示未知
资源状态	normal表示所有资源都正常 abnormal表示有异常资源 NULL表示未知

恢复指导：

1. 登录主管理节点，执行**su - omm**切换到omm用户。执行**${CONTROLLER_HOME}/sbin/status-oms.sh**查看OMS状态。
2. 如果HA状态为NULL，可能是系统在重启，这个一般是中间状态，HA后续会自动调整为正常状态。
3. 如果资源状态异常，则说明有Manager的某些组件资源异常， 可具体查看acs、aos、cep、controller、feed_watchdog、fms、gaussDB、httpd、iam、ntp、okerberos、oldap、pms、tomcat等组件状态是否正常。
4. 如果Manager组件资源异常，参见Manager组件状态检查进行处理。

Manager 组件状态检查

指标项名称：Manager组件状态检查

指标项含义：Manager组件状态检查包括组件资源运行状态和资源HA状态。资源运行状态，取值为Normal、Abnormal等；资源HA状态，取值为Normal、Exception等。Manager组件包含acs、aos、cep、controller、feed_watchdog、floatip、fms、gaussDB、heartBeatCheck、httpd、iam、ntp、okerberos、oldap、pms、tomcat等。当运行状态和HA状态不是Normal时，认为指标不健康。

表 12-28 Manager 组件状态说明表

名称	说明
资源运行状态	Normal表示正常运行 Abnormal表示运行异常 Stopped表示停止 Unknown表示状态未知 Starting表示正在启动 Stopping表示正在停止 Active_normal表示主正常运行 Standby_normal表示备正常运行 Raising_active表示正在升主 Lowing_standby表示正在降备 No_action表示没有该动作 Repairing表示正在修复 NULL表示未知
资源HA状态	Normal表示正常 Exception表示故障 Non_steady表示非稳态 Unknown表示未知 NULL表示未知

恢复指导：

1. 登录主管理节点，执行**su - omm**切换到omm用户。执行**${CONTROLLER_HOME}/sbin/status-oms.sh**查看OMS状态。
2. 如果floatip、okerberos、oldap等异常，可参见告警ALM-12002、ALM-12004、ALM-12005分别进行处理。
3. 如果是其他资源异常，建议查看相关异常模块的日志。

   controller资源异常：查看异常节点的/var/log/Bigdata/controller/controller.log。

   cep资源异常：查看异常节点的/var/log/Bigdata/omm/oms/cep/cep.log。

   aos资源异常：查看异常节点的/var/log/Bigdata/controller/aos/aos.log。

   feed_watchdog资源异常：查看异常节点的/var/log/Bigdata/watchdog/watchdog.log。

   httpd资源异常：查看异常节点的/var/log/Bigdata/httpd/error_log。

   fms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/fms/fms.log。

   pms资源异常：查看异常节点的/var/log/Bigdata/omm/oms/pms/pms.log。

   iam资源异常：查看异常节点的/var/log/Bigdata/omm/oms/iam/iam.log。

   gaussDB资源异常：查看异常节点的/var/log/Bigdata/omm/oms/db/omm_gaussdba.log。

ntp资源异常：查看异常节点的/var/log/Bigdata/omm/oms/ha/scriptlog/ha_ntp.log。

tomcat资源异常：查看异常节点的/var/log/Bigdata/tomcat/catalina.log。

4. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

OMA 运行状态

指标项名称：OMA运行状态

指标项含义：检查OMA的运行状态，状态结果包括运行和停止两种状态，如果OMA状态为停止，则认为不健康。

恢复指导：

1. 登录检查结果不健康的节点，然后执行**su - omm**切换到**omm**用户。
2. 执行**${OMA_PATH}/restart_oma_app**，手工启动OMA，然后重新检查。如果检查结果仍然不健康，则执行**3**。
3. 如果手工启动OMA无法恢复，建议查看分析OMA日志“/var/log/Bigdata/omm/oma/omm_agent.log”。
4. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

各节点与主管理节点之间 SSH 互信

指标项名称：各节点与主管理节点之间SSH互信

指标项含义：检查SSH互信是否正常。如果使用omm用户，在主管理节点可以通过SSH登录其他节点且不需要输入密码，则认为健康；否则，不健康。或者主管理节点SSH可以直接登录其他节点，但在其他节点无法通过SSH登录主管理节点，则也认为不健康。

恢复指导：

1. 如果该指标项检查异常，表示各节点与主管理节点之间SSH互信异常。SSH互信异常时，首先检查“/home/omm”目录的权限是否为**omm**。非**omm**的目录权限可能导致SSH互信异常，建议执行**chown omm:wheel**修改权限后重新检查。如果“/home/omm”目录权限正常，则执行**2**。
2. SSH互信异常一般会导致Controller和NodeAgent之间心跳异常，进而出现节点故障的告警。这时可参见告警ALM-12006进行处理。

进程运行时间

指标项名称：NodeAgent运行时间、Controller运行时间和Tomcat运行时间

指标项含义：检查NodeAgent、Controller、Tomcat进程的运行时间。如果小于半小时（即1800s），则进程可能重启过，建议半小时后再检查。如果多次检查，进程的运行时间都小于半小时，说进程状态异常。

恢复指导：

1. 登录检查结果不健康的节点，执行**su - omm**切换到**omm**用户。
2. 根据进程名称查看进程pid，执行命令：

   ps -ef | grep NodeAgent

3. 根据pid查看进程启动时间，执行命令：

   ps -p pid -o lstart

4. 判断进程启动时间是否正常。如果进程一直反复重启，执行**5**
5. 查看对应模块日志，分析重启原因。

   NodeAgent运行时间异常，检查相关日志**/var/log/Bigdata/nodeagent/agentlog/agent.log**。

   Controller运行时间异常，检查相关日志**/var/log/Bigdata/controller/controller.log**。

   Tomcat运行时间异常，检查相关日志**/var/log/Bigdata/tomcat/web.log**。

6. 如果通过日志无法排除问题，请联系运维人员处理，并发送已收集的故障日志信息。

账户及密码过期检查

指标项名称：账户及密码过期检查

指标项含义：该指标项检查MRS的两个操作系统用户**omm**和**ommdba**。对操作系统用户，同时检查账户及密码的过期时间。如果账户或密码有效期小于等于15天，则认为不健康。

恢复指导：如果账户或密码有效期小于等于15天，建议及时联系运维人员处理。

12.7.17 Spark 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查Spark服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警ALM-28001进行处理。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.18 Storm 健康检查指标项说明

工作节点数

指标项名称：Supervisor数

指标项含义：检查集群中可用的Supervisor数，若集群中可用的Supervisor数小于1，则认为不健康。

恢复指导：如果该指标项异常，进入Streaming服务实例页面，单击不可用Supervisor实例的“主机名”，在“概要信息”中查看主机的健康状态，若为“良好”，则参见“ALM-12007 进程故障”告警进行处理；若不为“良好”，则参见“ALM-12006 节点故障”告警进行处理。

空闲 Slot 数

指标项名称：空闲Slot数

指标项含义：检查集群中空闲的Slot数，若集群中空闲slot数目小于1，则认为不健康。

恢复指导：如果该指标项异常，进入Storm服务实例页面，查看Supervisor实例的“健康状态”，若均为“良好”，则需要扩容集群Core节点；若不为良好，则参见“ALM-12007 进程故障”告警进行处理。

服务健康状态

指标项名称：服务状态

指标项含义：检查Storm服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议参见“ALM-26051 Storm服务不可用”告警进行处理。

检查告警

指标项名称：告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.19 Yarn 健康检查指标项说明

服务健康状态

指标项名称：服务状态

指标项含义：检查Yarn服务状态是否正常。如果当前无法获取NodeManager节点数时，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理并确认网络无异常。

检查告警

指标项名称： 告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.7.20 ZooKeeper 健康检查指标项说明

ZooKeeper 服务处理请求平均延时

指标项名称：ZooKeeper服务处理请求平均延时

指标项含义：检查ZooKeeper服务处理请求的平均延时，如果大于300毫秒，则认为不健康。

恢复指导： 如果该指标项异常，则需要检查集群的网络速度是否正常、内存或CPU使用率是否过高。

ZooKeeper 连接数使用率

指标项名称：ZooKeeper连接数使用率

指标项含义：检查ZooKeeper内存使用率是否超过80%。如果超过阈值，则认为不健康。

恢复指导：如果该指标项异常，建议增加ZooKeeper服务可以使用的内存。可以通过ZooKeeper服务配置中的“GC_OPTS”配置项参数-Xmx来修改，修改完成需重启ZooKeeper服务。

服务健康状态

指标项名称：服务状态

指标项含义：检查ZooKeeper服务状态是否正常。如果状态不正常，则认为不健康。

恢复指导：如果该指标项异常，建议检查KrbServer、LdapServer两个服务的健康状态是否为故障并进行处理。然后登录ZooKeeper客户端，确认ZooKeeper是否无法写入数据，根据错误提示排查ZooKeeper写数据失败的原因。最后参告警ALM-13000进行处理。

检查告警

指标项名称：告警信息

指标项含义：检查服务是否存在未清除的告警。如果存在，则认为不健康。

恢复指导：如果该指标项异常，建议参见告警进行处理。

12.8 静态服务池管理

12.8.1 查看静态服务池状态

操作场景

MRS Manager支持通过静态服务资源池对没有运行在YARN上的服务资源进行管理和隔离。支持动态地管理HDFS和YARN在部署节点可使用的CPU、I/O和内存总量。系统支持基于时间的静态服务资源池自动调整策略，使集群在不同的时间段自动调整参数值，从而更有效地利用资源。

用户可以在MRS Manager查看静态服务池各个服务使用资源的监控指标结果，包含监控指标如下：

- 服务总体CPU使用率
- 服务总体磁盘I/O读速率
- 服务总体磁盘I/O写速率
- 服务总体内存使用大小

操作步骤

步骤1 在MRS Manager，单击“系统设置”，在“资源管理”区域单击“静态服务池”。

步骤2 单击“状态”。

步骤3 查看系统资源调整基数。

- “系统资源调整基数”表示集群中每个节点可以被集群服务使用的最大资源。如果节点只有一个服务，则表示此服务独占节点可用资源。如果节点有多个服务，则表示所有服务共同使用节点可用资源。
- “CPU(%)”表示节点中服务可使用的最大CPU。
- “Memory(%)”表示节点中服务可使用的最大内存。

步骤4 查看集群服务资源使用状态。

在图表区域的服务选择框中选择“所有服务”，则“图表”中会显示服务池所有服务的资源使用状态。

说明

“生效的配置组”表示集群服务当前使用的资源控制配置组。默认情况下每天所有时间均使用“default”配置组，表示集群服务可以使用节点全部CPU，以及70%的内存。

步骤5 查看单个服务资源使用状态。

在图表区域的服务选择框中选择指定服务，“图表”中会显示服务池此服务的资源使用状态。

步骤6 用户可以选择页面自动刷新间隔的设置。

支持三种参数值：

- “每30秒刷新一次”：刷新间隔30秒。
- “每60秒刷新一次”：刷新间隔60秒。
- “停止刷新”：停止刷新。

步骤7 在“时间区间”选择需要查看服务资源的时间段。可供选择的选项如下：

- 实时
- 最近3小时
- 最近6小时
- 最近24小时
- 最近一周
- 最近一个月
- 最近三个月
- 最近六个月
- 自定义：选择自定义时，在时间范围内自行选择需要查看的时间。

步骤8 单击“查看”可以查看相应时间区间的服务资源数据。

步骤9 自定义服务资源报表。

1. 单击“定制”，勾选需要显示的服务源指标。
   - 服务总体磁盘I/O读速率

- 服务总体内存使用大小
- 服务总体磁盘I/O写速率
- 服务总体CPU使用率

2. 单击“确定”保存并显示所选指标。

说明

单击“清除”可批量取消全部选中的指标项。

步骤10 导出监控指标报表。

单击“导出”，Manager将生成指定时间范围内、已勾选的服务资源指标报表文件，请选择一个位置保存，并妥善保管该文件。

说明

如果需要查看指定时间范围的监控指标对应的分布曲线图，请单击“查看”，界面将显示用户自定义时间范围内选定指标的分布曲线图。

----结束

12.8.2 配置静态服务池

操作场景

当需要控制集群服务可以使用节点的资源情况，或者在不同时间段集群服务使用节点的CPU不同，用户可以在MRS Manager调整资源基数，并自定义资源配置组。

前提条件

- 配置静态服务池后，HDFS和YARN服务需要重启，重启期间服务不可用。
- 配置静态服务池后，各服务及角色实例使用的最大资源将不能超过限制。

操作步骤

步骤1 修改系统资源调整基数。

1. 在MRS Manager界面，单击“系统设置”，在“资源管理”区域单击“静态服务池”。
2. 单击“配置”，显示服务池配置组管理页面。
3. 在“系统资源调整基数”分别修改参数“CPU(%)”和“Memory(%)”。

   修改“系统资源调整基数”将限制Flume、HBase、HDFS、Impala和YARN服务能够使用节点的最大物理CPU和内存资源百分比。如果多个服务部署在同一节点，则所有服务使用的最大物理资源百分比不能超过此参数值。
4. 单击“下一步”完成编辑。

   需要重新修改参数，可单击页面下方的“上一步”。

步骤2 修改服务池默认“default”配置组。

1. 在“服务池配置”表格中Flume、HBase、HDFS、Impala和YARN服务对应的“CPU LIMIT(%)”、“CPU SHARE(%)”、“I/O(%)”和“Memory(%)”填写各服务的资源使用百分比数量。

说明

- 所有服务使用的“CPU LIMIT(%)”资源配置总和可以大于100%。
- 所有服务使用的“CPU SHARE(%)”和“I/O(%)”资源配置总和需为100%。例如为HDFS和Yarn服务分配使用的CPU资源，2个服务分配到的CPU资源总和为100%。
- 所有服务使用的“Memory(%)”资源配置总和可以小于或等于100%，也可以大于100%。
- “Memory(%)”不支持动态生效，仅在“default”配置组中可以修改。

2. 单击页面空白处完成编辑，MRS Manager将根据集群硬件资源与分配情况，在“详细配置”生成服务池参数的正确配置值。
3. 如果根据业务需要，可以单击“详细配置”右侧的修改服务池的参数值。

   在“服务池配置”单击指定的服务名，“详细配置”将只显示此服务的参数。手工修改参数值并不会刷新服务使用资源的百分比显示。支持动态生效的参数，在新增加的配置组中显示名将包含配置组的编号，例如“HBase : RegionServer : dynamic-config1.RES_CPUSET_PERCENTAGE”，参数作用与“default”配置组中的参数相同。

表 12-29 静态服务池参数一览

参数名	参数描述
– RES_CPUSET_PERCENTAGE – dynamic-configX.RES_CPUSET_PERCENTAGE	配置服务使用CPU PERCENTAGE。
– RES_CPU_SHARE – dynamic-configX.RES_CPU_SHARE	配置服务使用CPU share。
– RES_BLKIO_WEIGHT – dynamic-configX.RES_BLKIO_WEIGHT	配置服务占用I/O的权重。
HBASE_HEAPSIZE	配置RegionServer的JVM最大内存。
HADOOP_HEAPSIZE	配置DataNode的JVM最大内存。
yarn.nodemanager.resource.memory-mb	配置当前节点上NodeManager可使用的内存大小。
dfs.datanode.max.locked.memory	配置DataNode用做HDFS缓存的最大内存。
FLUME_HEAPSIZE	配置每个flume实例能使用的最大JVM内存。
IMPALAD_MEM_LIMIT	配置impalad实例可使用的最大内存。

步骤3 添加自定义资源配置组。

1. 是否需要根据时间自动调整资源配置?

   是，执行**步骤3.2**。

否，执行**步骤4**。

2. 单击增加新的资源配置组。在“调度时间”，单击显示时间策略配置页面。

   根据业务需要修改以下参数，并单击“确定”保存：

   - “重复”：当勾选“重复”时表示此资源配置组按调度周期重复运行。不勾选时请设置一个资源配置组应用的日期与时间。
   - “重复策略”：支持“每天”、“每周”和“每月”。仅在“重复”模式中生效。
   - “介于”：表示资源配置组应用的开始与结束时间。请设置一个唯一的时间区间，如果与已有配置组的时间区间有重叠，则无法保存。仅在“重复”模式中生效。

   说明

   - “default”配置组会在所有未定义的时间段内生效。
   - 新增加的配置组属于动态生效的配置项集合，在配置组应用的时间区间内可直接生效。
   - 新增加的配置组可以被删除。最多增加4个动态生效的配置组。
   - 选择任一种“重复策略”，如果结束时间小于开始时间，默认标识为第二天的时间。例如“22:00”到“6:00”表示调度时间为当天22点到第二天6点。
   - 若多个配置组的“重复策略”类型不相同，则时间区间可以重叠，且生效的策略优先级从低到高的顺序为“每天”、“每周”、“每月”。例如，有“每月”与“每天”的调度配置组，时间区间分别为4:00到7:00，6:00到8:00，此时以每月的配置组为准。
   - 若多个配置组的“重复策略”类型相同，当日期不相同时，则时间区间可以重叠。例如，有两个“每周”的调度配置组，可以分别指定时间区间为周一和周三的4:00到7:00。

3. 在“服务池配置”修改各服务资源配置，并单击页面空白处完成编辑，然后执行**步骤4**。

   用户可单击“服务池配置”右侧的重新修改参数。如果根据业务需要，在“详细配置”单击，手动更新由系统生成的参数值。

步骤4 保存配置。

单击“保存”，在“保存配置”窗口勾选“重新启动受影响的服务或实例。”，单击是保存并重启相关服务。

界面提示“操作成功。”，单击“完成”，服务成功启动。

----结束

12.9 租户管理

12.9.1 租户简介

定义

MRS集群拥有的不同资源和服务支持多个组织、部门或应用共享使用。集群提供了一个逻辑实体来统一使用不同资源和服务，这个逻辑实例就是租户。多个不同的租户统称多租户。当前仅分析集群支持租户。

原理

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

计算资源指租户Yarn任务队列资源，可以修改任务队列的配额，并查看任务队列的使用状态和使用统计。

存储资源目前支持HDFS存储，可以添加删除租户HDFS存储目录，设置目录的文件数量配额和存储空间配额。

MRS Manager作为MRS集群的统一租户管理平台，可以为企业提供成熟的多租户管理模型，实现集中式的租户和业务管理。租户可以在界面上根据业务需要，在集群中创建租户、管理租户。

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。默认情况下，新的计算资源和存储资源的全部权限将分配给租户的角色。
- 默认情况下，查看当前租户的资源、在当前租户中添加子租户并管理子租户资源的权限将分配给租户的角色。
- 修改租户的计算资源或存储资源，对应的角色关联权限将自动更新。

MRS Manager中最多支持512个租户。系统默认创建的租户包含“default”。和默认租户同处于最上层的租户，可以统称为一级租户。

资源池

YARN任务队列支持一种调度策略，称为标签调度（Label Based Scheduling）。通过此策略，YARN任务队列可以关联带有特定节点标签（Node Label）的NodeManager，使YARN任务在指定的节点运行，实现任务的调度与使用特定硬件资源的需求。例如，需要使用大量内存的YARN任务，可以通过标签关联具有大量内存的节点上运行，避免性能不足影响业务。

在MRS集群中，租户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个资源池。YARN任务队列通过配置队列容量策略，与指定的资源池进行关联，可以更有效地使用资源池中的资源，且互不影响。

MRS Manager中最多支持50个资源池。系统默认包含一个“Default”资源池。

12.9.2 添加租户

操作场景

当租户需要根据业务需求指定资源使用情况时，可以在MRS Manager创建租户。

前提条件

- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的完整存储路径在HDFS目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“添加租户”，打开添加租户的配置页面，参见以下表格内容为租户配置属性。

表 12-30 租户参数一览表

参数名	描述
“名称”	指定当前租户的名称，长度为1到20，可包含数字、字母和下划线。
“租户类型”	可选参数值为“叶子租户”和“非叶子租户”。当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
“动态资源”	为当前租户选择动态计算资源。系统将自动在Yarn中以租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。
“默认资源池容量 (%)”	配置当前租户在“default”资源池中使用的计算资源百分比。
“默认资源池最大容量 (%)”	配置当前租户在“default”资源池中使用的最大计算资源百分比。
“储存资源”	为当前租户选择存储资源。系统将自动在“/tenant”目录中以租户名称创建文件夹。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。
“存储空间配额 (MB)”	配置当前租户使用的HDFS存储空间配额。取值范围为“1”到“8796093022208”。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间表示所有副本文件在HDFS中占用的磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
“存储路径”	配置租户在HDFS中的存储目录。系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“ta1”，默认HDFS存储目录为“tenant/ta1”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
“服务”	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
“描述”	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配Manager_tenant角色以及租户对应的角色。具体操作请参见**创建用户**。

----结束

相关任务

查看已添加的租户

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，单击已添加租户的名称。

默认在右侧显示“概述”页签。

步骤3 查看当前租户的“基本信息”、“资源配额”和“统计”。

如果HDFS处于“已停止”状态，“资源配额”中“Space”的“可用”和“已使用”会显示为“unknown”。

----结束

12.9.3 添加子租户

操作场景

当租户需要根据业务需求，将当前租户的资源进一步分配时，可以在MRS Manager添加子租户。

前提条件

- 已添加上级租户。
- 根据业务需求规划租户的名称，不得与当前集群中已有的角色或者Yarn队列重名。
- 如果子租户需要使用存储资源，则提前根据业务需要规划好存储路径，分配的存储目录在父租户的存储目录中不存在。
- 规划当前租户可分配的资源，确保每一级别父租户下，直接子租户的资源百分比之和不能超过100%。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，将光标移动到需要添加子租户的租户节点上，单击“添加子租户”，打开添加子租户的配置页面，参见以下表格内容为租户配置属性。

表 12-31 子租户参数一览表

参数名	描述
“父租户”	显示上级父租户的名称。
“名称”	指定当前租户的名称，长度为1到20，可包含数字、字母和下划线。
“租户类型”	可选参数值为“叶子租户”和“非叶子租户”，当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
“动态资源”	为当前租户选择动态计算资源。系统将自动在Yarn父租户队列中以子租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。如果父租户未选择动态资源，子租户也无法使用动态资源。
“默认资源池容量 (%)”	配置当前租户使用的资源百分比，基数为父租户的资源总量。
“默认资源池最大容量 (%)”	配置当前租户使用的最大计算资源百分比，基数为父租户的资源总量。
“储存资源”	为当前租户选择存储资源。系统将自动在HDFS父租户目录中，以子租户名称创建文件夹。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。如果父租户未选择存储资源，子租户也无法使用存储资源。
“存储空间配额 (MB)”	配置当前租户使用的HDFS存储空间配额。最小值值为“1”，最大值为父租户的全部存储配额。单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。若此配额大于父租户的配额，实际存储量受父租户配额影响。 **说明** 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“父租户中分配资源”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
“存储路径”	配置租户在HDFS中的存储目录。系统默认将自动在父租户目录中以子租户名称创建文件夹。例如子租户“ta1s”，父目录为“tenant/ta1”，系统默认自动配置此参数值为“tenant/ta1/ta1s”，最终子租户的存储目录为“/tenant/ta1/ta1s”。支持在父目录中自定义存储路径。存储路径的父目录必需是父租户的存储目录。

参数名	描述
“服务”	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
“描述”	配置当前租户的描述信息。

步骤3 单击“确定”保存，完成子租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配租户对应的角色。具体操作请参见**创建用户**。

----结束

12.9.4 删除租户

操作场景

当租户需要根据业务需求，将当前不再使用的租户删除时，可以在MRS Manager完成操作。

前提条件

- 已添加租户。
- 检查待删除的租户是否存在子租户，如果存在，需要先删除全部子租户，否则无法删除当前租户。
- 待删除租户的角色，不能与任何一个用户或者用户组存在关联关系。该任务对应取消角色与用户的绑定，请参见**修改用户信息**。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，将光标移动到需要删除的租户节点上，单击“删除”。

界面显示“删除租户”对话框。根据业务需求，需要保留租户已有的数据时请同时勾选“保留该租户的数据”，否则将自动删除租户对应的存储空间。

步骤3 单击“确定”保存，删除租户。

保存配置需要等待一段时间，租户成功删除。租户对应的角色、存储空间将删除。

说明

- 租户删除后，Yarn中对应的租户任务队列不会被删除。
- 删除父租户时选择不保留数据，如果存在子租户且子租户使用了存储资源，则子租户的数据也会被删除。

----结束

12.9.5 管理租户目录

操作场景

用户根据业务需求，可以在MRS Manager对指定租户使用的HDFS存储目录，进行管理操作。支持用户对租户添加目录、修改目录文件数量配额、修改存储空间配额和删除目录。

前提条件

已添加关联了HDFS存储资源的租户。

操作步骤

- 查看租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击目标的租户。
  c. 单击“资源”页签。
  d. 查看“HDFS存储”表格。
     - 指定租户目录的“文件目录数上限”列表示文件和目录数量配额。
     - 指定租户目录的“存储空间配额 (MB)”列表示租户目录的存储空间大小。
- 添加租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击需要添加HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，单击“添加目录”。
     - “父目录”选择一个父租户的存储目录。
       该参数仅适用于子租户。如果父租户有多个目录，请选择其中任何一个。
     - “路径”填写租户目录的路径。

       说明

       - 如果当前租户不是子租户，新路径将在HDFS的根目录下创建。
       - 如果当前租户是一个子租户，新路径将在指定的目录下创建。

       完整的HDFS存储目录最多包含1023个字符。HDFS目录名称包含数字、大小写字母、空格和下划线。空格只能在HDFS目录名称的中间使用。

- “文件\目录数上限”填写文件和目录数量配额。
  “文件\目录数上限”为可选参数，取值范围从1到9223372036854775806。
- “存储空间配额 (MB)”填写租户目录的存储空间大小。
  “存储空间配额 (MB)”的取值范围从1到8796093022208。

  说明

  为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

e. 单击“确定”完成租户目录添加，系统将在HDFS根目录下创建租户的目录。

- 修改租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击需要修改HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，指定租户目录的“操作”列，单击“修改”。
     - “文件\目录数上限”填写文件和目录数量配额。
       “文件\目录数上限”为可选参数，取值范围从1到9223372036854775806。
     - “存储空间配额”填写租户目录的存储空间大小。
       “存储空间配额”的取值范围从1到8796093022208。

       说明

       为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间球所有副本文件在HDFS中占用磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。

  e. 单击“确定”完成租户目录修改。
- 删除租户目录
  a. 在MRS Manager，单击“租户管理”。
  b. 在左侧租户列表，单击需要删除HDFS存储目录的租户。
  c. 单击“资源”页签。
  d. 在“HDFS存储”表格，指定租户目录的“操作”列，单击“删除”。
     创建租户时设置的默认HDFS存储目录不支持删除，仅支持删除新添加的HDFS存储目录。
  e. 单击“确定”完成租户目录删除。

12.9.6 恢复租户数据

操作场景

租户的数据默认在Manager和集群组件中保存相关数据，在组件故障恢复或者卸载重新安装的场景下，所有租户的部分配置数据可能状态不正常，需要手动恢复。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 在左侧租户列表，单击某个租户节点。

步骤3 检查租户数据状态。

1. 在“概述”，查看“基本信息”左侧的圆圈，绿色表示租户可用，灰色表示租户不可用。
2. 单击“资源”，查看“Yarn”或者“HDFS存储”左侧的圆圈，绿色表示资源可用，灰色表示资源不可用。
3. 单击“服务关联”，查看关联的服务表格的“状态”列，“良好”表示组件可正常为关联的租户提供服务，“故障”表示组件无法为租户提供服务。
4. 任意一个检查结果不正常，需要恢复租户数据，请执行**步骤4**。

步骤4 单击“恢复租户数据”。

步骤5 在“恢复租户数据”窗口，选择一个或多个需要恢复数据的组件，单击“确定”，等待系统自动恢复租户数据。

----结束

12.9.7 添加资源池

操作场景

在MRS集群中，用户从逻辑上对YARN集群的节点进行分区，使多个NodeManager形成一个YARN资源池。每个NodeManager只能属于一个资源池。系统中默认包含了一个名为“Default”的资源池，所有未加入用户自定义资源池的NodeManager属于此资源池。

该任务指导用户通过MRS Manager添加一个自定义的资源池，并将未加入自定义资源池的主机加入此资源池。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 单击“添加资源池”。

步骤4 在“添加资源池”设置资源池的属性。

- “名称”：填写资源池的名称。不支持创建名称为“Default”的资源池。

  资源池的名称，长度为1到20位，可包含数字、字母和下划线，且不能以下划线开头。

- “可用主机”：在界面左边主机列表，选择指定的主机名称，单击 >> ，将选中的主机加入资源池。只支持选择本集群中的主机。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

步骤6 完成资源池创建后，用户可以在资源池的列表中查看资源池的“名称”、“成员”、“类型”、“虚拟核数”与“内存”。已加入自定义资源池的主机，不再是“Default”资源池的成员。

----结束

12.9.8 修改资源池

操作场景

该任务指导用户通过MRS Manager，修改已有资源池中的成员。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“修改”。

步骤4 在“编辑资源池”修改“已添加主机”。

- 增加主机：在界面左边主机列表，选择指定的主机名称，单击

，将选中的主机加入资源池。
- 删除主机：在界面右边主机列表，选择指定的主机名称，单击

，将选中的主机移出资源池。资源池中的主机列表可以为空。

步骤5 单击“确定”保存。

----结束

12.9.9 删除资源池

操作场景

该任务指导用户通过MRS Manager，删除已有资源池。

前提条件

- 集群中任何一个队列不能使用待删除资源池为默认资源池，删除资源池前需要先取消默认资源池，请参见**配置队列**。
- 集群中任何一个队列不能在待删除资源池中配置过资源分布策略，删除资源池前需要先清除策略，请参见**清除队列配置**。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“资源池”页签。

步骤3 在资源池列表指定资源池所在行的“操作”列，单击“删除”。

在弹出窗口中单击“确定”。

----结束

12.9.10 配置队列

操作场景

用户根据业务需求，可以在MRS Manager修改指定租户的队列配置。

前提条件

已添加关联Yarn并分配了动态资源的租户。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“动态资源计划”页签。

步骤3 单击“队列配置”页签。

步骤4 在租户队列表格，指定租户队列的“操作”列，单击“修改”。

说明

在“租户管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击 也能打开修改队列配置页面。

表 12-32 队列配置参数

参数名	描述
“最大应用数量”	表示最大应用程序数量。取值范围从“1”到“2147483647”。
“AM最大资源百分比”	表示集群中可用于运行application master的最大资源占比。取值范围从“0”到“1”。
“用户资源最小上限百分比 (%)”	表示用户使用的最小资源上限百分比。取值范围从“0”到“100”。
“用户资源上限因子”	表示用户使用的最大资源限制因子，与当前租户在集群中实际资源百分比相乘，可计算出用户使用的最大资源百分比。最小值为“0”。
“状态”	表示资源计划当前的状态，“运行”为运行状态，“停止”为停止状态。
“默认资源池”	表示队列使用的资源池。默认为“Default”，如果需要修改为其他资源，需要先配置队列容量，请参见**配置资源池的队列容量策略**。

----结束

12.9.11 配置资源池的队列容量策略

操作场景

添加资源池后，需要为YARN任务队列配置在此资源池中可使用资源的容量策略，队列中的任务才可以正常在这个资源池中执行。每个队列只能配置一个资源池的队列容量策略。用户可以在任何一个资源池中查看队列并配置队列容量策略。配置队列策略后，YARN任务队列与资源池形成关联关系。

该任务指导用户通过MRS Manager配置队列策略。

前提条件

- 已添加资源池。
- 任务队列与其他资源池无关联关系。默认情况下，所有队列与“Default”资源池存在关联关系。

操作步骤

步骤1 在MRS Manager，单击“租户管理”。

步骤2 单击“动态资源计划”页签。

步骤3 在“资源池”选择指定的资源池。

“可用资源配额”：表示每个资源池默认所有资源都可分配给队列。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“修改”。

步骤5 在“修改资源分配”窗口设置任务队列在此资源池中的资源容量策略。

- “资源容量 (%)”：表示当前租户计算资源使用的资源百分比。
- “最大资源容量 (%)”：表示当前租户计算资源使用的最大资源百分比。

步骤6 单击“确定”保存配置。

----结束

12.9.12 清除队列配置

操作场景

当队列不再需要某个资源池的资源，或资源池需要与队列取消关联关系时，用户可以在MRS Manager清除队列配置。清除队列配置即取消队列在此资源池中的资源容量策略。

前提条件

如果队列需要清除与某个资源池的绑定关系，该资源池不能作为队列的默认资源池，需要先将队列的默认资源池更改为其他资源池，请参见**配置队列**。

操作步骤

步骤1 在MRS Manager界面，单击“租户管理”。

步骤2 单击“动态资源计划”页签。

步骤3 在“资源池”选择指定的资源池。

步骤4 在“资源分配”列表指定队列的“操作”列，单击“清除”。

在“清除队列配置”中单击“确定”，清除队列在当前资源池的配置。

说明

如果用户未配置队列的资源容量策略，则清除功能默认不可用。

----结束

12.10 备份与恢复

12.10.1 备份与恢复简介

概述

MRS Manager提供对系统内的用户数据及系统数据的备份恢复能力，备份功能按组件提供，支持备份管理系统Manager的数据（需要同时备份OMS和LdapServer）、Hive用户数据、DBService中保存的组件元数据和HDFS元数据备份。

备份恢复任务的使用场景如下：

- 用于日常备份，确保系统及组件的数据安全。
- 当系统故障导致无法工作时，使用已备份的数据完成恢复操作。
- 当主集群完全故障，需要创建一个与主集群完全相同的镜像集群，可以使用已备份的数据完成恢复操作。

表 12-33 根据业务需要备份元数据

备份类型	备份内容
OMS	默认备份集群管理系统中的数据库数据（不包含告警数据）以及配置数据。
LdapServer	备份用户信息，包括用户名、密码、密钥、密码策略、组信息。
DBService	备份DBService管理的组件（Hive）的元数据。
NameNode	备份HDFS元数据。

原理

任务

在进行备份恢复之前，需要先创建备份恢复任务，并指定任务的参数，例如任务名称、备份数据源和备份文件保存的目录类型等等。通过执行备份恢复任务，用户可完成数据的备份恢复需求。在使用Manager执行恢复HDFS、Hive和NameNode数据时，无法访问集群。

每个备份任务可同时备份不同的数据源，每个数据源将生成独立的备份文件，每次备份的所有备份文件组成一个备份文件集，可用于恢复任务。备份任务支持将备份文件保存在Linux本地磁盘、本集群HDFS与备集群HDFS中。备份任务提供全量备份或增量备份的策略，增量备份策略支持HDFS和Hive备份任务，OMS、LdapServer、DBService和NameNode备份任务默认只应用全量备份策略。

说明

任务运行规则：

- 某个任务已经处于执行状态，则当前任务无法重复执行，其他任务也无法启动。
- 周期任务自动执行时，距离该任务上次执行的时间间隔需要在120秒以上，否则任务推迟到下个周期启动。手动启动任务无时间间隔限制。
- 周期任务自动执行时，当前时间不得晚于任务开始时间120秒以上，否则任务推迟到下个周期启动。
- 周期任务锁定时无法自动执行，需要手动解锁。
- OMS、LdapServer、DBService和NameNode备份任务开始执行前，若主管理节点“LocalBackup”分区可用空间小于20GB，则无法开始执行。
- 用户在规划备份恢复任务时，请严格根据业务逻辑、数据存储结构、数据库或表关联关系，选择需要备份或者恢复的数据。系统默认创建了一个间隔为24小时的周期备份任务“default”，支持全量备份OMS、LdapServer、DBService和NameNode数据到Linux本地磁盘。

规格

表 12-34 备份恢复特性规格

项目	规格
备份或恢复任务最大数量（个）	100
同时运行的任务数量（个）	1
等待运行的任务最大数量（个）	199
Linux本地磁盘最大备份文件大小（GB）	600

表 12-35 “default”任务规格

项目	OMS	LdapServer	DBService	NameNode
备份周期	1小时			
最大备份数	2个			
单个备份文件最大大小	10MB	20MB	100MB	1.5GB
最大占用磁盘大小	20MB	40MB	200MB	3GB
备份数据保存位置	主备管理节点“*数据存放路径*/LocalBackup/”			

说明

“default”任务保存的备份数据，请用户根据企业运维要求，定期转移并保存到集群外部。

12.10.2 备份元数据

操作场景

为了确保元数据信息安全，或者用户需要对元数据功能进行重大操作（如扩容缩容、安装补丁包、升级或迁移等）前后，需要对元数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。元数据包含OMS数据、LdapServer数据、DBService数据和NameNode数据。备份Manager数据包含同时备份OMS数据和LdapServer数据。

默认情况下，元数据备份由“default”任务支持。该任务指导用户通过MRS Manager创建备份任务并备份元数据。支持创建任务自动或手动备份数据。

前提条件

- 需要准备一个用于备份数据的备集群，且网络连通。每个集群的安全组，需分别添加对端集群的安全组入方向规则，允许安全组中所有弹性云服务器全部协议全部端口的访问请求。
- 根据业务需要，规划备份的类型、周期和策略等规格，并检查主备管理节点“数据存放路径/LocalBackup/”是否有充足的空间。

操作步骤

步骤1 创建备份任务。

1. 在MRS Manager，选择“系统设置 > 备份管理”。
2. 单击“创建备份任务”。

步骤2 设置备份策略。

1. 在“任务名称”填写备份任务的名称。
2. 在“备份类型”选择备份任务的运行类型，“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

   创建周期备份任务，还需要填写以下参数：

   - “开始时间”：表示任务第一次启动的时间。
   - “周期”：表示任务下次启动，与上一次运行的时间间隔，支持“按小时”或“按天”。
   - “备份策略”：表示任务每次启动时备份的数据量。支持“首次全量备份，后续增量备份”、“每次都全量备份”和“每n次进行一次全量备份”。选择“每n次进行一次全量备份”时，需要指定**n**的值。

步骤3 选择备份源。

在“备份配置”，勾选元数据选项，例如“OMS”和“LdapServer”。

步骤4 设置备份参数。

1. 在“OMS”和“LdapServer”的“路径类型”，选择一个备份目录的类型。

   备份目录支持以下类型：

- “LocalDir”：表示将备份文件保存在主管理节点的本地磁盘上，备管理节点将自动同步备份文件。默认保存目录为“数据存放路径/LocalBackup/”。选择此参数值，还需要配置“最大备份数”，表示备份目录中可保留的备份文件集数量。
- “LocalHDFS”：表示将备份文件保存在当前集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “目的端路径”：填写备份文件在HDFS中保存的目录。不支持填写HDFS中的隐藏目录，例如快照或回收站目录；也不支持默认的系统目录。
  - “最大备份数”：填写备份目录中可保留的备份文件集数量。
  - “目标实例名称”：选择备份目录对应的NameService名称。默认值为“hacluster”。

2. 单击“确定”保存。

步骤5 执行备份任务。

在备份任务列表中已创建任务的“操作”列，若“备份类型”选择“周期备份”请单击“即时备份”，若“备份类型”选择“手动备份”请单击“启动”，开始执行备份任务。

备份任务执行完成后，系统自动在备份目录中为每个备份任务创建子目录，目录名为备份任务名_任务创建时间，用于保存数据源的备份文件。备份文件的名称为版本号_数据源_任务执行时间.tar.gz。

----结束

12.10.3 恢复元数据

操作场景

在用户意外修改删除、数据需要找回，对元数据组件进行重大操作（如升级、重大数据调整等）后系统数据出现异常或未达到预期结果，模块全部故障完全无法使用，或者迁移数据到新集群的场景中，需要对元数据进行恢复操作。

该任务指导用户通过MRS Manager创建恢复元数据任务。只支持创建任务手动恢复数据。

须知

- 只支持进行数据备份时的系统版本与当前系统版本一致时的数据恢复。
- 当业务正常时需要恢复数据，建议手动备份最新管理数据后，再执行恢复数据操作。否则会丢失从备份时刻到恢复时刻之间的元数据。
- 必须使用同一时间点的OMS和LdapServer备份数据进行恢复，否则可能造成业务和操作失败。
- MRS集群默认使用DBService保存Hive的元数据。

对系统的影响

- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。

前提条件

- 检查OMS和LdapServer备份文件是否是同一时间点备份的数据。
- 检查OMS资源状态是否正常，检查LdapServer实例状态是否正常。如果不正常，不能执行恢复操作。
- 检查集群主机和服务的状态是否正常。如果不正常，不能执行恢复操作。
- 检查恢复数据时集群主机拓扑结构与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份。
- 检查恢复数据时集群中已添加的服务与备份数据时是否相同。如果不相同，不能执行恢复操作，必须重新备份
- 检查DBService主备实例状态是否正常。如果不正常，不能执行恢复操作。
- 停止依赖MRS集群运行的上层业务应用。
- 在MRS Manager停止所有待恢复数据的NameNode角色实例，其他的HDFS角色实例必须保持正常运行，恢复数据后重启NameNode。NameNode角色实例重启前无法访问。
- 检查NameNode备份文件是否保存在主管理节点“数据存放路径/LocalBackup/”。

操作步骤

步骤1 查看备份数据位置。

1. 在MRS Manager，选择“系统设置 > 备份管理”。
2. 在任务列表指定任务的“操作”列，单击“更多 > 查询历史”，打开备份任务执行历史记录。 在弹出的窗口中，在指定一次执行成功记录的“备份路径”列，单击“查看”，打开此次任务执行的备份路径信息，查找以下信息：
   - “备份对象”表示备份的数据源。
   - “备份路径”表示备份文件保存的完整路径。
3. 选择正确的项目，在“备份路径”手工选中备份文件的完整路径并复制。

步骤2 创建恢复任务。

1. 在MRS Manager，选择“系统设置 > 恢复管理”。
2. 单击“创建恢复任务”。
3. 在“任务名称”填写恢复任务的名称。

步骤3 选择恢复源。

在“恢复配置”，勾选待恢复数据的元数据组件。

步骤4 设置恢复参数。

1. 在“路径类型”，选择一个备份目录的类型。
2. 选择不同的备份目录时，对应设置如下：
   - “LocalDir”：表示备份文件保存在主管理节点的本地磁盘上。选择此参数值，还需要配置“源端路径”，表示备份文件保存位置的完整路径。例如，

“数据存放路径/LocalBackup/备份任务名_任务创建时间/数据源_任务执行时间/版本号_数据源_任务执行时间.tar.gz”。

- “LocalHDFS”：表示备份文件保存在当前集群的HDFS目录。选择此参数值，还需要配置以下参数：
  - “源端路径”：表示备份文件在HDFS中保存的完整路径。例如“备份路径/备份任务名_任务创建时间/版本号_数据源_任务执行时间.tar.gz”。
  - “源实例名称”：选择恢复任务执行时备份目录对应的NameService名称。默认值为“hacluster”。

3. 单击“确定”保存。

步骤5 执行恢复任务。

在恢复任务列表已创建任务的“操作”列，单击“启动”，开始执行恢复任务。

- 恢复成功后进度显示为绿色。
- 恢复成功后此恢复任务不支持再次执行。
- 如果恢复任务在第一次执行时由于某些原因未执行成功，在排除错误原因后单击“启动”，重试恢复任务。

步骤6 恢复了哪个元数据?

- 恢复了OMS和LdapServer元数据，执行**步骤7**。
- 恢复了DBService数据，任务结束。
- 恢复NameNode数据，在MRS Manager，选择“服务管理 > HDFS > 更多 > 重启服务”，任务结束。

步骤7 重启Manager使恢复数据生效。

1. 在MRS Manager，选择“LdapServer > 更多 > 重启服务”，单击“确定”，等待LdapServer服务重启成功。
2. 登录主管理节点，详情请参见**如何确认Manager的主备管理节点**。
3. 执行以下命令，重新启动OMS。

   sh ${BIGDATA_HOME}/om-0.0.1/sbin/restart-oms.sh

   提示以下信息表示命令执行成功：

   ```
   start HA successfully.
   ```

4. 在MRS Manager，选择“KrbServer > 更多 > 同步配置”，不勾选“重启配置过期的服务或实例”，单击“确定”，等待KrbServer服务配置同步及重启成功。
5. 选择“服务管理 > 更多 > 同步配置”，不勾选“重启配置过期的服务或实例”，单击“确定”，等待集群配置同步成功。
6. 选择“服务管理 > 更多 > 停止集群”。待停止集群的操作生效后，选择“服务管理 > 更多 > 启动集群”，等待集群启动成功。

----结束

12.10.4 修改备份任务

操作场景

该任务指导用户通过MRS Manager修改已创建的备份任务的配置参数，以适应业务需求的变化。不支持修改任何恢复任务配置参数，只能查看恢复任务的配置参数。

对系统的影响

修改备份任务后，新的参数在下一次执行任务时生效。

前提条件

- 已创建备份任务。
- 已根据业务实际需求，规划新的备份任务策略。

操作步骤

步骤1 在MRS Manager，选择“系统设置 > 备份管理”。

步骤2 在任务列表指定任务的“操作”列，单击“修改”，打开修改配置页面。

步骤3 在新页面中修改任务参数。

- 手动备份支持修改的参数项如下：
  - 目的端路径
  - 最大备份数
- 周期备份支持修改的参数项如下：
  - 开始时间
  - 周期
  - 目的端路径
  - 最大备份数

  说明

  - 当备份任务的“路径类型”为“LocalHDFS”时，修改备份任务时参数“目的端路径”有效。
  - 修改某个备份任务参数“目的端路径”后，第一次执行此任务默认为全量备份。

步骤4 单击“确定”保存。

----结束

12.10.5 查看备份恢复任务

操作场景

该任务指导用户通过MRS Manager查看已创建的备份恢复任务，以及任务的运行情况。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 单击“备份管理”或“恢复管理”。

步骤3 在任务列表中，查看“当次任务进度”列获取上一次任务运行的结果。绿色表示运行成功，红色表示运行失败。

步骤4 在任务列表指定任务的“操作”列，单击“更多 > 查询历史”，打开备份恢复任务运行记录。

在弹出的窗口中，在指定一次执行记录的“详情”列，单击“查看”，打开此次任务运行的日志信息。

----结束

相关任务

- 修改备份任务

  参考**修改备份任务**。
- 查看恢复任务

  在任务列表指定任务的“操作”列，单击“查询详情”，查看恢复任务。恢复任务的参数只能查看但不能修改。
- 运行备份恢复任务

  在任务列表指定任务的“操作”列，单击“启动”，启动处于准备或失败状态的备份、恢复任务。已成功执行过的恢复任务不能重新运行。
- 停止备份任务

  在任务列表指定任务的“操作”列，单击“更多 > 停止”，停止处于运行状态的备份恢复任务。
- 删除备份恢复任务

  在任务列表指定任务的“操作”列，单击“更多 > 删除”，删除备份恢复任务。删除任务后备份的数据默认会保留。
- 挂起备份任务

  在任务列表指定任务的“操作”列，单击“更多 > 挂起”，挂起备份任务。仅支持周期备份的任务，挂起后周期备份任务不再自动执行。挂起正在执行的备份任务时，该任务会停止运行。需要取消任务的挂起状态时，单击“更多 > 重新执行”。

12.11 安全管理

12.11.1 未开启 Kerberos 认证集群中的默认用户清单

用户分类

MRS集群提供以下2类用户，请用户定期修改密码，不建议使用默认密码。

用户类型	使用说明
系统用户	用于OMS系统进程运行的用户。
数据库用户	• 用于OMS数据库管理和数据访问的用户。 • 用于业务组件（Hive、Loader和DBservice）数据库的用户。

系统用户

说明

- MRS集群需要使用操作系统中ldap用户，此账号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作用户维护。
- 首次修改“ommdba”和“omm”密码需执行重置密码操作。找回密码后建议定期修改。

类别	用户名称	初始密码	描述
MRS集群系统管理员	admin	在集群创建时由用户指定。	• 针对1.8.0之前集群，MRS Manager记录集群审计日志的默认用户。 • 针对1.8.0及之后集群，MRS Manager的管理员。 此外还具有以下权限： – 具有HDFS、ZooKeeper普通用户的权限。 – 具有提交、查询Mapreduce、YARN任务的权限，以及YARN队列管理权限和访问YARN WebUI的权限。 – Storm中，具有提交、查询、激活、去激活、重分配、删除拓扑的权限，可以操作所有拓扑。 – Kafka服务中，具有创建、删除、授权、Reassign、消费、写入、查询主题的权限。
MRS集群节点操作系统用户	omm	系统随机生成	MRS集群系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。
MRS集群节点操作系统用户	root	用户设置的密码。 **说明** 适用于MRS1.6.2版本（包含）及以后的集群。	MRS集群所属节点的登录用户。在全部节点生成，属于操作系统用户。

用户组信息

默认用户组	描述
supergroup	admin用户的主组，在关闭Kerberos认证的集群中没有额外的权限。
check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
Manager_tenant	租户系统用户组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
System_administrator	MRS集群系统管理员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_viewer	MRS Manager系统查看员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_operator	MRS Manager系统操作员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_auditor	MRS Manager系统审计员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
Manager_administrator	MRS Manager系统管理员组。系统内部组，仅限组件间内部使用，且仅在已启用Kerberos认证的集群中使用。
compcommon	MRS集群系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
default_1000	为租户创建的用户组。系统内部组，仅限组件间内部使用。
launcher-job	MRS系统内部组，用于使用V2接口提交作业。

操作系统用户组	描述
wheel	MRS集群系统内部运行用户“omm”的主组。
ficommon	MRS集群系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

数据库用户

MRS集群系统数据库用户包含OMS数据库用户、DBService数据库用户。

说明

数据库用户不能删除，否则可能导致集群或组件服务无法正常工作。

类别	默认用户	初始密码	描述
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作
	omm	ChangeMe@123456	OMS数据库数据访问用户
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户
	hive	HiveUser@	Hive连接DBService数据库用户
	hue	HueUser@123	Hue连接DBService数据库用户

12.11.2 开启 Kerberos 认证集群中的默认用户清单

用户分类

MRS集群提供以下3类用户，请用户定期修改密码，不建议使用默认密码。

用户类型	使用说明
系统用户	● 通过Manager创建，是MRS集群操作运维与业务场景中主要使用的用户，包含两种类型： – “人机”用户：用于在Manager的操作运维场景，以及在组件客户端操作的场景。 – “机机”用户：用于MRS集群应用开发的场景。 ● 用于OMS系统进程运行的用户。
系统内部用户	MRS集群提供的用于进程通信、保存用户组信息和关联用户权限的内部用户。
数据库用户	● 用于OMS数据库管理和数据访问的用户。 ● 用于业务组件（Hive、Hue、Loader和DBservice）数据库的用户。

系统用户

说明

- MRS集群需要使用操作系统中ldap用户，此账号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作用户维护。
- 首次修改“ommdba”和“omm”用户需要执行重置密码操作。找回密码后建议定期修改。

类别	用户名称	初始密码	描述
MRS集群系统管理员	admin	在集群创建时由用户指定。	Manager的管理员。 此外还具有以下权限： ● 具有HDFS、ZooKeeper普通用户的权限。 ● 具有提交、查询Mapreduce、YARN任务的权限，以及YARN队列管理权限和访问YARN WebUI的权限。 ● Storm中，具有提交、查询、激活、去激活、重分配、删除拓扑的权限，可以操作所有拓扑。 ● Kafka服务中，具有创建、删除、授权、Reassign、消费、写入、查询主题的权限。
MRS集群节点操作系统用户	omm	系统随机生成	MRS集群系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。
MRS集群节点操作系统用户	root	用户设置的密码。	MRS集群所属节点的登录用户。在全部节点生成，属于操作系统用户。

系统内部用户

说明

以下系统内部用户不能删除，否则可能导致集群或组件无法正常工作。

类别	默认用户	初始密码	描述
组件运行用户	hdfs	Hdfs@123	HDFS系统管理员，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看webUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统
	hbase	Hbase@123	HBase系统管理员，用户权限： ● 集群管理权限: 表的Enable、Disable操作，触发MajorCompact，ACL操作 ● 授权或回收权限，集群关闭等操作相关的权限 ● 表管理权限: 建表、修改表、删除表等操作权限 ● 数据管理权限：表级别、列族级别以及列级别的数据读写权限 ● 访问HBase WebUI的权限
	mapred	Mapred@123	MapReduce系统管理员，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 访问Yarn、MapReduce WebUI的权限
	spark	Spark@123	Spark系统管理员，用户权限： ● 访问Spark WebUI的权限 ● 提交Spark任务的权限

用户组信息

默认用户组	描述
hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。
hbase	普通用户组，将用户加入此用户组不会获得额外的权限。
hive	将用户加入此用户组，可以使用Hive。
spark	普通用户组，将用户加入此用户组不会获得额外的权限。
supergroup	将用户加入此用户组，可获得HBase、HDFS和Yarn的管理员权限，并可以使用Hive。
check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
Manager_tenant	租户系统用户组。系统内部组，仅限组件间内部使用。
System_administrator	MRS集群系统管理员组。系统内部组，仅限组件间内部使用。
Manager_viewer	MRS Manager系统查看员组。系统内部组，仅限组件间内部使用。
Manager_operator	MRS Manager系统操作员组。系统内部组，仅限组件间内部使用。
Manager_auditor	MRS Manager系统审计员组。系统内部组，仅限组件间内部使用。
Manager_administrator	MRS Manager系统管理员组。系统内部组，仅限组件间内部使用。
compcommon	MRS系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
default_1000	为租户创建的用户组。系统内部组，仅限组件间内部使用。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限,才能访问对应Topic。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。

默认用户组	描述
stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。
opentsdb	普通用户组，将用户加入此用户组不会获得额外的权限。
presto	普通用户组，将用户加入此用户组不会获得额外的权限。
flume	普通用户组，添加到该用户组的用户无任何额外权限。
launcher-job	MRS系统内部组，用于使用V2接口提交作业。

操作系统用户组	描述
wheel	MRS集群系统内部运行用户“omm”的主组。
ficommon	MRS集群系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

数据库用户

MRS集群系统数据库用户包含OMS数据库用户、DBService数据库用户。

说明

数据库用户不能删除，否则可能导致集群或组件服务无法正常工作。

类别	默认用户	初始密码	描述
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作
	omm	ChangeMe@123456	OMS数据库数据访问用户
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户
	hive	HiveUser@	Hive连接DBService数据库用户
	hue	HueUser@123	Hue连接DBService数据库用户
	sqoop	SqoopUser@	Loader连接DBService数据库的用户

类别	默认用户	初始密码	描述
	ranger	RangerUser@	Ranger连接DBService数据库的用户

12.11.3 修改操作系统用户密码

操作场景

该任务指导用户定期修改MRS集群节点操作系统用户“omm”、“ommdba”、"root"的登录密码，以提升系统运维安全性。

各节点“omm”、“ommdba”、"root"无需设置为统一的密码。

操作步骤

步骤1 登录Master1节点，然后登录要修改操作系统用户密码的其他节点。

步骤2 执行以下命令切换到用户。

sudo su - root

步骤3 执行如下命令，修改omm/ommdba/root用户密码。

passwd omm

passwd ommdba

passwd root

例如omm：passwd，系统显示：

```
Changing password for user omm.
New password:
```

输入用户的新密码。操作系统的密码修改策略由用户实际使用的操作系统类型决定。

```
Retype new password:
passwd: all authentication tokens updated successfully.
```

说明

MRS集群默认密码复杂度要求：

- 密码字符长度至少8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 重置的密码不能是最近5次使用过的密码。

----结束

12.11.4 修改 admin 密码

该任务指导用户定期修改集群用户“admin”的密码，以提升系统运维安全性。

修改该密码会导致已经下载的用户凭证不可用，请修改该密码后重新下载认证凭据并替换旧凭据。

在集群节点修改 admin 密码

步骤1 更新主管理节点客户端，具体请参看**更新客户端（3.x之前版本）**。

步骤2 登录主管理节点。

步骤3 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤4 执行以下命令切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤5 执行以下命令配置环境变量。

source bigdata_env

步骤6 执行以下命令，修改“admin”密码。此操作在整个集群中生效。

kpasswd admin

先输入旧密码，再输入两次新密码。

集群中，默认的密码复杂度要求：

- 密码字符长度至少8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

在 MRS Manager 页面修改 admin 密码

开启Kerberos认证的集群和开启弹性公网IP功能未开启Kerberos认证的集群支持通过MRS Manager界面修改admin密码。

步骤1 用admin账户登录MRS Manager页面。

步骤2 单击页面右上角用户名，选择“修改密码”。

步骤3 在修改密码页面，输入“旧密码”、“新密码”、“确认新密码”。

说明

默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

步骤4 单击“确定”完成密码修改，使用新密码重新登录MRS Manager页面。

----结束

重置 admin 密码

步骤1 登录Master1节点。

步骤2 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如重置admin密码：**cpw admin**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.11.5 修改 Kerberos 管理员密码

操作场景

该任务指导用户定期修改MRS集群Kerberos管理员“kadmin”的密码，以提升系统运维安全性。

修改该密码会导致已经下载的用户凭证不可用，请修改该密码后重新下载认证凭据并替换旧凭据。

前提条件

已在Master1节点准备客户端。

操作步骤

步骤1 登录Master1节点。

步骤2 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，修改**kadmin/admin**密码。此操作对所有服务器生效。

kpasswd kadmin/admin

集群中，默认的密码复杂度要求：

- 密码字符长度至少8位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.11.6 修改 LDAP 管理员和 LDAP 用户密码

操作场景

该任务指导用户定期修改MRS集群的LDAP管理员用户“rootdn:cn=root,dc=hadoop,dc=com”和LDAP用户“pg_search_dn:cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启全部服务，服务在重启时无法访问。

操作步骤

步骤1 在MRS Manager，选择“服务管理 > LdapServer > 更多”。

步骤2 单击“修改密码”。

步骤3 在“修改密码”对话框的“用户信息”选择要修改的用户。

步骤4 在“修改密码”对话框的“旧密码”输入旧密码，“新密码”和“确认密码”输入新密码。

默认密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符`~!@#$%^&*()-_=+\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序用户名相同。
- 不可与当前密码相同。

说明

LDAP管理员用户“rootdn:cn=root,dc=hadoop,dc=com”的默认密码为“LdapChangeMe@123”，LDAP用户“pg_search_dn:cn=pg_search_dn,ou=Users,dc=hadoop,dc=com”的默认密码为“pg_search_dn@123”，请定期修改密码并妥善保存。

步骤5 勾选“我已阅读此信息并了解其影响。”，单击“确定”确认修改并重启服务。

----结束

12.11.7 修改组件运行用户密码

操作场景

该任务指导用户定期修改MRS集群组件运行用户的密码，以提升系统运维安全性。

如果初始密码由系统随机生成，需要直接重置密码。

修改该密码会导致已经下载的用户凭证不可用，请修改该密码后重新下载认证凭据并替换旧凭据。

前提条件

已在Master1节点准备客户端。

操作步骤

步骤1 登录Master1节点。

步骤2 （可选）若想要使用omm用户修改密码，请执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如重置admin密码：**cpw admin**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。

- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.11.8 修改 OMS 数据库管理员密码

操作场景

该任务指导用户定期修改OMS数据库管理员的密码，以提升系统运维安全性。

操作步骤

步骤1 登录主管理节点。

说明

ommdba用户密码不支持在备管理节点修改，否则集群无法正常工作。只需在主管理节点执行修改操作，无需在备管理节点操作。

步骤2 执行以下命令，切换用户。

sudo su - omm

步骤3 执行以下命令，切换目录。

cd $OMS_RUN_PATH/tools

步骤4 执行以下命令，修改ommdba用户密码。

mod_db_passwd ommdba

步骤5 输入**ommdba**的原密码后，再输入两次新密码。

密码复杂度要求：

- 密码字符长度为16～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序用户名相同。
- 不可与前20个历史密码相同。

显示如下结果，说明修改成功：

```
Congratulations, update [ommdba] password successfully.
```

----结束

12.11.9 修改 OMS 数据库数据访问用户密码

操作场景

该任务指导用户定期修改OMS数据库访问用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启OMS服务，服务在重启时无法访问。

操作步骤

步骤1 在MRS Manager单击“系统设置”。

步骤2 在“权限配置”区域下，单击“OMS数据库密码修改”。

步骤3 在omm用户所在行，单击“操作”列下的“修改密码”，修改OMS数据库密码。

密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。
- 不可与前20个历史密码相同。

步骤4 单击“确定”，等待界面提示“操作成功”后单击“完成”。

步骤5 在omm用户所在行，单击“操作”列下的“重启OMS服务”，重启OMS数据库。

说明

如果修改了密码但未重启OMS数据库，则omm用户的状态变为“Waiting to restart”且无法再修改密码，直到重启OMS数据库

步骤6 在弹出的对话框中，勾选“我已阅读此信息并了解其影响。”，单击“确定”，重新启动OMS服务。

----结束

12.11.10 修改组件数据库用户密码

操作场景

该任务指导用户定期修改组件数据库用户的密码，以提升系统运维安全性。

对系统的影响

修改密码需要重启服务，服务在重启时无法访问。

操作步骤

步骤1 在MRS Manager单击“服务管理”，单击待修改数据库用户服务的名称。

步骤2 确定修改哪个组件数据库用户密码。

- 修改DBService数据库用户密码，直接执行**步骤3**。
- 修改Hive或者Hue或者Loader数据库用户密码，需要先停止服务再执行**步骤3**。

单击“停止服务”。

步骤3 选择“更多 > 修改密码”。

步骤4 根据界面信息，输入新旧密码。

密码复杂度要求：

- DBService数据库用户密码字符长度为16～32位。Hive或Hue或Loader数据库用户密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、特殊字符~`!@#$%^&*()-+_=\|[{}];:",<.>/?中的3种类型字符。
- 不能与用户名或倒序用户名相同。
- 不可与前20个历史密码相同。

步骤5 单击“确定”，系统自动重新启动对应的服务。界面提示“操作成功”，单击“完成”。

----结束

12.11.11 更换 HA 证书

操作场景

HA证书用于主备进程与高可用进程的通信过程中加密数据，实现安全通信。该任务指导用户为MRS Manager完成主备管理节点的HA证书替换工作，以确保产品安全使用。

证书文件和密钥文件可由用户生成。

对系统的影响

更换过程中MRS Manager需要重启，此时系统无法访问且无法提供服务。

前提条件

- 获取需要更换的HA根证书文件“root-ca.crt”和密钥文件“root-ca.pem”。
- 准备一个访问密钥文件的密码password，例如“Userpwd@123”用于访问密钥文件。

  密码复杂度要求如下，如果密码复杂度不满足如下要求，可能存在安全风险：
  - 密码字符长度至少为8个字符
  - 至少需要包含大写字母、小写字母、数字、特殊字符~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=中的4种类型字符

操作步骤

步骤1 登录主管理节点。

步骤2 执行以下命令切换用户：

sudo su - root

su - omm

步骤3 执行以下命令在主管理节点“${OMS_RUN_PATH}/workspace0/ha/local/cert”目录生成“root-ca.crt”和“root-ca.pem”：

sh ${OMS_RUN_PATH}/workspace/ha/module/hacom/script/gen-cert.sh --root-ca --country=*country* **--state=***state* **--city=***city* **--company=***company* **--**

organize=*organize* **--common-name=***commonname* **--email=***管理员邮箱* **--password=***password*

例如，执行以下命令：**sh ${OMS_RUN_PATH}/workspace/ha/module/hacom/script/gen-cert.sh --root-ca --country=XX--state=XX --city=XX --company=XX --organize=IT --common-name=HADOOP.COM --email=abc@XXX.com --password=Userpwd@123**

提示以下信息表示命令执行成功：

```
Generate root-ca pair success.
```

步骤4 在主管理节点以“omm”用户执行以下命令，复制“root-ca.crt”和“root-ca.pem”到“${BIGDATA_HOME}/om-0.0.1/security/certHA”目录。

cp -arp ${OMS_RUN_PATH}/workspace0/ha/local/cert/root-ca.* ${BIGDATA_HOME}/om-0.0.1/security/certHA

步骤5 使用“omm”用户将主管理节点生成的“root-ca.crt”和“root-ca.pem”复制到备管理节点“${BIGDATA_HOME}/om-0.0.1/security/certHA”目录。

步骤6 执行以下命令，生成HA用户证书并自动替换。

sh ${BIGDATA_HOME}/om-0.0.1/sbin/replacehaSSLCert.sh

根据提示信息输入password，并按回车键确认。

```
Please input ha ssl cert password:
```

界面提示以下信息表示HA用户证书替换成功：

```
[INFO] Succeed to replace ha ssl cert.
```

步骤7 执行以下命令，重启OMS。

sh ${BIGDATA_HOME}/om-0.0.1/sbin/restart-oms.sh

界面提示以下信息：

```
start HA successfully.
```

步骤8 登录备管理节点并切换到**omm**用户，重复**步骤6**～**步骤7**。

执行**sh ${BIGDATA_HOME}/om-0.0.1/sbin/status-oms.sh**，查看管理节点的“HAAllResOK”是否为“Normal”，并可以重新访问MRS Manager表示操作成功。

----结束

12.11.12 更新集群密钥

操作场景

在创建集群时，系统将自动生成加密密钥key值以对集群的部分安全信息（例如所有数据库用户密码、密钥文件访问密码等）进行加密存储。在集群安装成功后，建议用户定期通过以下操作手动更改密钥值。

对系统的影响

- 更新集群密钥后，集群中新增加一个随机生成的新密钥，用于加密解密新保存的数据。旧的密钥不会删除，用于解密旧的加密数据。在修改安全信息后，例如修改数据库用户密码，新密码将使用新的密钥加密。

- 更新集群密钥需要停止集群，集群停止时无法访问。

前提条件

停止依赖集群运行的上层业务应用。

操作步骤

步骤1 在MRS Manager，选择“服务管理 > 更多 > 停止集群”。

在弹出窗口勾选“我已阅读此信息并了解影响。”，单击“确定”，界面提示“操作成功。”，单击“完成”，集群成功停止。

步骤2 登录主管理节点。

步骤3 执行以下命令切换用户：

sudo su - omm

步骤4 执行以下命令，防止超时退出。

TMOUT=0

步骤5 执行以下命令，切换目录。

cd ${BIGDATA_HOME}/om-0.0.1/tools

步骤6 执行以下命令，更新集群密钥。

sh updateRootKey.sh

根据界面提示，输入**y**：

```
The root key update is a critical operation.
Do you want to continue?(y/n):
```

界面提示以下信息表示更新密钥成功：

```
...
Step 4-1: The key save path is obtained successfully.
...
Step 4-4: The root key is sent successfully.
```

步骤7 在MRS Manager界面，选择“服务管理 > 更多 > 启动集群”。

在弹出的提示框中单击“是”，开始启动集群。界面提示“操作成功。”，单击“完成”，集群成功启动。

----结束

12.12 权限管理

12.12.1 创建角色

操作场景

该任务指导用户在MRS Manager创建角色，并对Manager和组件进行授权管理。

MRS Manager支持的角色数为1000。

前提条件

用户已明确业务需求。

操作步骤

步骤1 在MRS Manager，选择“系统设置 > 角色管理”。

步骤2 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

“角色名称”为必选参数，字符长度为3到30，可以包含数字、字母和下划线。“描述”为可选参数。

步骤3 设置角色“权限”。

1. 单击“服务名称”，并选择一个“视图名称”。
2. 勾选一个或多个权限。

说明

- “权限”为可选参数。
- 在选择“视图”设置组件的权限时，可通过右上角的“搜索”框输入资源名称，然后单击 显示搜索结果。
- 搜索范围仅包含当前权限目录，无法搜索子目录。搜索关键字支持模糊搜索，不区分大小写。支持搜索下一页的结果。

表 12-36 Manager 权限描述

支持权限管理的资源	权限设置说明
“Alarm”	Manager告警功能授权，勾选“View”表示可以查看告警，勾选“Management”表示可以管理告警。
“Audit”	Manager审计日志功能授权，勾选“View”表示可以查看审计，勾选“Management”表示可以管理审计。
“Dashboard”	Manager概览功能授权，勾选“View”表示可以查看集群概览。
“Hosts”	Manager集群节点管理功能授权，勾选“View”表示可以查看节点，勾选“Management”表示可以管理节点。
“Services”	MRS集群服务管理功能授权，勾选“View”表示可以查看服务，勾选“Management”表示可以管理服务。
“System_cluster_management”	MRS集群管理授权，勾选“Management”表示可以使用MRS补丁管理功能。
“System_configuration”	MRS集群配置功能授权，勾选“Management”表示可以使用Manager配置MRS集群。
“System_task”	MRS集群任务功能授权，勾选“Management”表示可以使用Manager管理MRS集群的周期任务。

支持权限管理的资源	权限设置说明
“Tenant”	Manager多租户管理功能授权，勾选“Management”表示可以查看Manager的租户管理页面。

表 12-37 HBase 权限描述

支持权限管理的资源	权限设置说明
“SUPER_USER_GROUP”	选中时表示授予HBase管理员权限。
“Global”	HBase的一种资源类型，表示HBase整体组件。
“Namespace”	HBase的一种资源类型，表示命名空间，用来保存HBase表。具体权限： ● “Admin”：表示管理此命名空间的权限。 ● “Create”：表示在此命名空间创建HBase表的权限。 ● “Read”：表示访问此命名空间的权限。 ● “Write”：表示写入此命名空间数据的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“Table”	HBase的一种资源类型，表示数据表，用来保存数据。具体权限： ● “Admin”：表示管理此数据表的权限。 ● “Create”：表示在此数据表创建列族和列的权限。 ● “Read”：表示读取数据表的权限。 ● “Write”：表示写入数据到表的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“ColumnFamily”	HBase的一种资源类型，表示列族，用来保存数据。具体权限： ● “Create”：表示在此列族创建列的权限。 ● “Read”：表示读取列族的权限。 ● “Write”：表示写入数据到列族的权限。
“Qualifier”	HBase的一种资源类型，表示列，用来保存数据。具体权限： ● “Read”：表示读取列的权限。 ● “Write”：表示写入数据到列的权限。

HBase中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如命名空间“default”添加了“Read”和“Write”权限，则命名空间中

的表、列族和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 12-38 HDFS 权限描述

支持权限管理的资源	权限设置说明
“Folder”	HDFS的一种资源类型，表示HDFS目录，可以保存文件或子目录。具体权限： ● “Read”：表示访问此HDFS目录的权限。 ● “Write”：表示在此HDFS目录写入数据的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。
“Files”	HDFS的一种资源类型，表示HDFS中的文件。具体权限： ● “Read”：表示访问此文件的权限。 ● “Write”：表示写入此文件的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。

HDFS中每一级目录的权限默认不会传递到下级目录类型。例如目录“tmp”添加了“Read”和“Execute”，需要同时勾选“递归”才能为子目录添加权限。

表 12-39 Hive 权限描述

支持权限管理的资源	权限设置说明
“Hive Admin Privilege”	选中时表示授予Hive管理员权限。
“Database”	Hive的一种资源类型，表示Hive数据库，用来保存Hive表。具体权限： ● “Select”：表示查询Hive数据库的权限。 ● “Delete”：表示在Hive数据库执行删除操作的权限。 ● “Insert”：表示在Hive数据库执行插入操作的权限。 ● “Create”：表示在Hive数据库执行创建操作的权限。

支持权限管理的资源	权限设置说明
“Table”	Hive的一种资源类型，表示Hive表，用来保存数据。具体权限： ● “Select”：表示查询Hive表的权限。 ● “Delete”：表示在Hive表执行删除操作的权限。 ● “Update”：表示为角色添加Hive表的“Update”权限。 ● “Insert”：表示在Hive表执行插入操作的权限。 ● “Grant of Select”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Select”权限。 ● “Grant of Delete”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Delete”权限。 ● “Grant of Update”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Update”权限。 ● “Grant of Insert”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Insert”权限。

Hive中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如数据库“default”添加了“Select”和“Insert”权限，则数据库中的表和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 12-40 YARN 权限描述

支持权限管理的资源	权限设置说明
“Cluster Admin Operations”	选中时表示授予YARN管理员权限。
“root”	YARN的根队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Parent Queue”	YARN的一种资源类型，表示父队列，可以包含子队列。根队列也属于父队列的一种。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Leaf Queue”	YARN的一种资源类型，表示叶子队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。

YARN中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如队列“root”添加了“Submit”权限，则子队列自动添加该权限。子队列继承的权限不在“权限”表格显示被选中。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 12-41 Hue 权限描述

支持权限管理的资源	权限设置说明
“Storage Policy Admin”	选中时表示授予Hue中存储策略管理员权限。

步骤4 单击“确定”完成，返回“角色管理”。

----结束

相关任务

修改角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要修改角色所在的行，单击“修改”，修改角色信息。

说明

修改角色分配的权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要删除角色所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

12.12.2 创建用户组

操作场景

该任务指导用户通过MRS Manager创建新用户组并指定其操作权限，使用户组可以统一管理加入用户组的单个或多个用户。用户加入用户组后，可获得用户组具有的操作权限。

MRS Manager支持用户组数为100。

前提条件

用户已明确业务需求，并已创建业务场景需要的角色。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在组列表上方，单击“添加用户组”。

步骤4 填写“组名”和“描述”。

“组名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。“描述”为可选参数。

步骤5 在“角色”，单击“选择添加角色”选择指定的角色并添加。

如果不添加角色，则当前创建的用户组没有使用MRS集群的权限。

步骤6 单击“确定”完成用户组创建。

----结束

相关任务

修改用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要修改用户组所在的行，单击“修改”，修改用户组信息。

说明

为用户组修改分配的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要删除用户组所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

12.12.3 创建用户

操作场景

该任务指导用户根据实际业务场景需要，通过MRS Manager创建新用户并指定其操作权限以满足业务使用。

MRS Manager支持的用户数为1000。

如需对新创建用户的密码使用新的密码策略，请先**修改密码策略**，再参考本章节创建用户。

前提条件

已明确业务需求，并已创建业务场景需要的角色和用户组。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在用户列表上方，单击“添加用户”。

步骤4 根据界面提示配置参数，填写“用户名”。

说明

- 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。
- 使用已创建的用户时，请输入和用户名完全一样的大小写字符。
- “用户名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。
- “root”、“omm”和“ommdba”为系统保留用户，请选择其他用户名。

步骤5 设置“用户类型”，可选值包括“人机”和“机机”。

- “人机”用户：用于在MRS Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。
- “机机”用户：用于MRS应用开发的场景。选择该值用户密码随机生成，无需填写。

步骤6 在“用户组”，单击“选择添加的用户组”，选择对应用户组将用户添加进去。

说明

- 如果用户组添加了角色，则用户可获得对应角色中的权限。
- 为新用户分配Hive的权限，请将用户加入hive组。
- 如果用户需要管理租户资源，用户组必须分配了Manager_tenant角色以及租户对应的角色。

步骤7 在“主组”选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。

步骤8 根据业务实际需要在“分配角色权限”，单击“选择并绑定角色”为用户添加角色。

说明

- 创建用户时，如果用户从用户组获得的权限还不满足业务需要，则可以再分配其他已创建的角色。为新用户分配角色授权，最长可能需要3分钟时间生效。
- 创建用户时添加角色可细化用户的权限。
- 没有为新用户分配角色时，此用户可以访问HDFS、HBase、Yarn、Spark和Hue的WebUI。

步骤9 根据业务实际需要“描述”。

“描述”为可选参数。

步骤10 单击“确定”完成用户创建。

第一次在MRS集群中使用新创建的用户，例如登录Manager或者使用集群客户端，需要修改密码，具体请参见《修改操作用户密码》。

----结束

12.12.4 修改用户信息

操作场景

该任务指导用户在MRS Manager修改已创建的用户信息，包括修改用户组、主组、角色和描述。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要修改用户所在的行，单击“修改”，修改用户信息。

说明

为用户修改用户组或分配的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

12.12.5 锁定用户

该任务指导用户将MRS集群中的用户锁定。用户被锁定后，不能在MRS Manager重新登录或在集群中重新进行安全认证。

可通过以下两种方式锁定用户，锁定后的用户需要手动解锁或者等待锁定时间结束才能恢复使用：

- 自动锁定：通过设置密码策略中的“允许输入错误次数”，将超过登录失败次数的用户自动锁定。具体操作请参见**修改密码策略**。
- 手动锁定：通过手动的方式将用户锁定。

以下将具体介绍手动锁定。不支持锁定“机机”用户。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要锁定用户所在行，单击“锁定用户”，锁定用户。

步骤4 在弹出的提示窗口，单击“确定”完成锁定操作。

----结束

12.12.6 解锁用户

在用户输入错误密码次数大于允许输入错误次数，造成用户被锁定或者用户被手动锁定后需要解锁用户的场景下，用户可以通过MRS Manager为锁定的用户解锁。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要解锁用户所在行，选择“解锁用户”，解锁用户。

步骤4 在弹出的提示窗口，单击“确定”完成解锁操作。

----结束

12.12.7 删除用户

MRS集群用户不再需要使用时，用户可以在MRS Manager中删除此用户。

说明

如果删除用户A后，再次准备重新创建同名用户A，如果该用户A已经提交过作业（客户端提交或者MRS console页面提交），那么需要在删除该用户A的同时，删除该用户A残留的文件夹，否则使用重新创建的同名用户A提交作业会失败。

删除用户残留文件夹操作方法为：依次登录MRS集群的Core节点，在每个Core节点上执行如下两条命令，其中如下命令中“$user”为具体的以用户名命名的文件夹。

cd /srv/BigData/hadoop/data1/nm/localdir/usercache/

rm -rf $user

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要删除用户所在的行，选择“更多 > 删除”。

步骤4 单击“确定”完成删除操作。

----结束

12.12.8 修改操作用户密码

操作场景

出于MRS集群安全的考虑，“人机”类型系统用户密码必须定期修改。该任务指导用户通过MRS Manager完成修改自身密码工作。

如需对用户修改的密码使用新的密码策略，请先**修改密码策略**，再参考本章节修改密码。

对系统的影响

修改MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

前提条件

- 已获取当前的密码策略。
- 已获取MRS Manager访问地址。

操作步骤

步骤1 在MRS Manager，移动鼠标到界面右上角的。

在弹出菜单，选择“修改密码”。

步骤2 分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.12.9 初始化系统用户密码

操作场景

该任务指导在用户遗忘密码或公共账号密码需要定期修改时，通过MRS Manager初始化密码。初始化密码后用户首次使用需要修改密码。

对系统的影响

初始化MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

初始化“人机”用户密码

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在要初始化密码用户所在行，单击“更多 > 初始化密码”，按界面提示信息修改用户密码。

在弹出窗口中输入当前登录的用户密码确认身份，单击“确定”，然后在“初始化密码”单击“确定”。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。

- 不能与用户名或倒序的用户名相同。

----结束

初始化“机机”用户密码

步骤1 根据业务情况，准备好客户端，并登录安装客户端的节点。

步骤2 执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/client”。

cd /opt/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

kadmin -p kadmin/admin

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如：**cpw oms/manager**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.12.10 下载用户认证文件

操作场景

用户开发大数据应用程序并在支持Kerberos认证的MRS集群中运行程序时，需要准备访问MRS集群的用户认证文件。认证文件中的keytab文件可用于认证用户身份。

该任务指导用户通过MRS Manager下载用户认证文件并导出keytab文件。

说明

- 如果选择下载“人机”用户的认证文件，在下载前需要使用Manager修改过一次此用户的密码使设置的初始密码失效，否则导出的keytab文件无法使用。请参见**修改操作用户密码**。
- 修改用户密码后，之前导出的keytab将失效，需要重新导出。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户管理”。

步骤3 在需导出keytab文件用户所在的行，选择“更多 > 下载认证凭据”下载认证文件，待文件自动生成后指定保存位置，并妥善保管该文件。

步骤4 使用解压程序打开认证文件。

- “user.keytab”表示用户keytab文件，用于认证用户身份。
- “krb5.conf”表示认证服务器配置文件，应用程序在进行用户认证身份时根据该文件的配置信息连接认证服务器。

----结束

12.12.11 修改密码策略

操作场景

该任务指导用户设置密码安全规则、用户登录安全规则及用户锁定规则。由于“机机”用户密码随机生成，在MRS Manager设置密码策略只影响“人机”用户。

如需对新增用户的密码或用户修改的密码使用新的密码策略，请先参考本章节修改密码策略，再**创建用户**或**修改密码**。

须知

密码策略涉及用户管理的安全性，请根据业务安全要求谨慎修改，否则会有安全性风险。

操作步骤

步骤1 在MRS Manager，单击“系统设置”。

步骤2 单击“密码策略配置”。

步骤3 根据界面提示，修改密码策略，具体参数见下表。

表 12-42 密码策略参数说明

参数名称	描述
最小密码长度	密码包含的最小字符个数，取值范围是8～32。默认值为“8”。
字符类型的数目	密码字符包含大写字母、小写字母、数字、空格和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=）的最小种类。可选择数值为“3”和“4”。默认值“3”表示至少必须使用大写字母、小写字母、数字、特殊符号和空格中的任意3种。

参数名称	描述
密码有效期（天）	密码有效使用天数，取值范围0～90，0表示永久有效。默认值为“90”。
密码失效提前提醒天数	提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录MRS Manager时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒。默认值为“5”。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔（分钟），取值范围为0～1440。“0”表示永远有效，“1440”表示1天。默认值为“5”。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。默认值为“5”。
用户锁屏时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。默认值为“5”。

----结束

12.13 MRS 多用户权限管理

12.13.1 MRS 集群中的用户与权限

概述

- **MRS集群用户**
  Manager中的安全账号，包含用户名、密码等属性，MRS集群的使用者通过这类用户访问集群中的资源。每个启用Kerberos认证的MRS集群可以有多个用户。
- **MRS集群角色**
  用户在实际使用MRS集群时需根据业务场景获取访问资源的权限，访问资源的权限是定义到MRS集群对象上的。集群的角色就是包含一个或者多个权限的集合。例如，HDFS中某个目录的访问权限，需要在指定的目录上配置，并保存在角色中。

Manager支持MRS集群用户权限管理功能，使权限管理与用户管理更加直观、易用。

- 权限管理：使用RBAC（Role-Based Access Control）方式，即基于角色授予权限，形成权限的集合。用户通过分配一个或多个已授权的角色取得对应的权限。
- 用户管理：使用Manager统一管理MRS集群用户，并通过Kerberos协议认证用户，LDAP协议存储用户信息。

权限管理

MRS集群提供的权限包括Manager和各组件（例如HDFS、HBase、Hive和Yarn等）的操作维护权限，在实际应用时需根据业务场景为各用户分别配置不同权限。为了提升

权限管理的易用性，Manager引入角色的功能，通过选取指定的权限并统一授予角色，以权限集合的形式实现了权限集中查看和管理，提升了权限管理的易用性和用户体验。

角色可以理解为集中一个或多个权限的逻辑体，角色被授予指定的权限，用户通过绑定角色获得对应的权限。

一个角色可以有多个权限，一个用户可以绑定多个角色。

- 角色1：授予操作权限A和B，用户a和用户b通过绑定角色1取得对应的权限。
- 角色2：授予操作权限C，用户c和用户d通过绑定角色2取得对应的权限。
- 角色3：授予操作权限D和F，用户a通过绑定配角色3取得对应的权限。

例如，MRS集群用户绑定了管理员角色，那么这个用户成为MRS集群的管理员用户。

Manager界面显示系统默认创建的角色如**表12-43**所示。

表 12-43 Manager 默认角色与描述

默认角色	角色描述
default	为租户创建的角色。
Manager_administrator	Manager管理员，具有Manager的管理权限。
Manager_auditor	Manager审计员，具有查看和管理审计信息的权限。
Manager_operator	Manager操作员，具有除租户管理、配置管理和集群管理功能以外的权限。
Manager_viewer	Manager查看员，具有查看系统概览，服务，主机，告警，审计日志等信息的权限。
System_administrator	系统管理员，具有Manager的管理权限及所有服务管理员的所有权限。
Manager_tenant	Manager租户管理页面查看角色，具有Manager“租户管理”页面查看权限。

通过Manager创建角色时支持对Manager和组件进行授权管理，如**表12-44**所示。

表 12-44 Manager 与组件授权管理

授权类型	授权描述
Manager	Manager访问与登录权限。
HBase	HBase管理员权限设置和表、列族授权。
HDFS	HDFS中的目录和文件授权。

授权类型	授权描述
Hive	• Hive Admin Privilege Hive管理员权限。 • Hive Read Write Privileges Hive数据表管理权限，可设置与管理已创建的表的数据操作权限。
Hue	存储策略管理员权限。
Yarn	• Cluster Admin Operations Yarn管理员权限。 • Scheduler Queue 队列资源管理。

用户管理

支持Kerberos认证的MRS集群使用Kerberos协议和LDAP（Lightweight Directory Access Protocol）协议来配合工作，实现用户管理：

- Kerberos用于在用户登录Manager与使用组件客户端时认证用户身份，未启用Kerberos认证的集群则不认证用户身份。
- LDAP用于存储用户记录、用户组信息与权限信息等用户信息。

MRS集群支持在Manager执行创建用户或者修改用户等任务时，系统自动完成更新Kerberos和LDAP的用户数据，用户登录Manager或使用组件客户端时，系统自动完成认证用户身份和获取用户信息。这样一方面保证了用户管理的安全性，另一方面简化了用户管理的操作任务。Manager还提供了用户组功能，可对单个或多个用户进行分类管理：

- 用户组为一批用户的集合，可对用户进行分类管理。系统中的用户可以单独存在也可以加入到某个用户组中。
- 对已分配角色的用户组来说，当用户添加到该组时，用户组分配的角色权限将授权给用户。

MRS 3.x之前版本集群MRS Manager界面显示系统默认创建的用户组如**表12-45**所示。

表 12-45 Manager 默认用户组与描述

用户组名称	描述
hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。
hbase	普通用户组，将用户加入此用户组不会获得额外的权限。
hive	将用户加入此用户组，可以使用Hive。
spark	普通用户组，将用户加入此用户组不会获得额外的权限。

用户组名称	描述
supergroup	将用户加入此用户组，可获得HBase、HDFS和Yarn的管理员权限，并可以使用Hive。
flume	普通用户组。添加到该用户组的用户无任何额外权限。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限,才能访问对应Topic。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。
stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。

启用Kerberos认证的MRS集群默认创建“admin”用户账号，用于集群管理员维护集群。

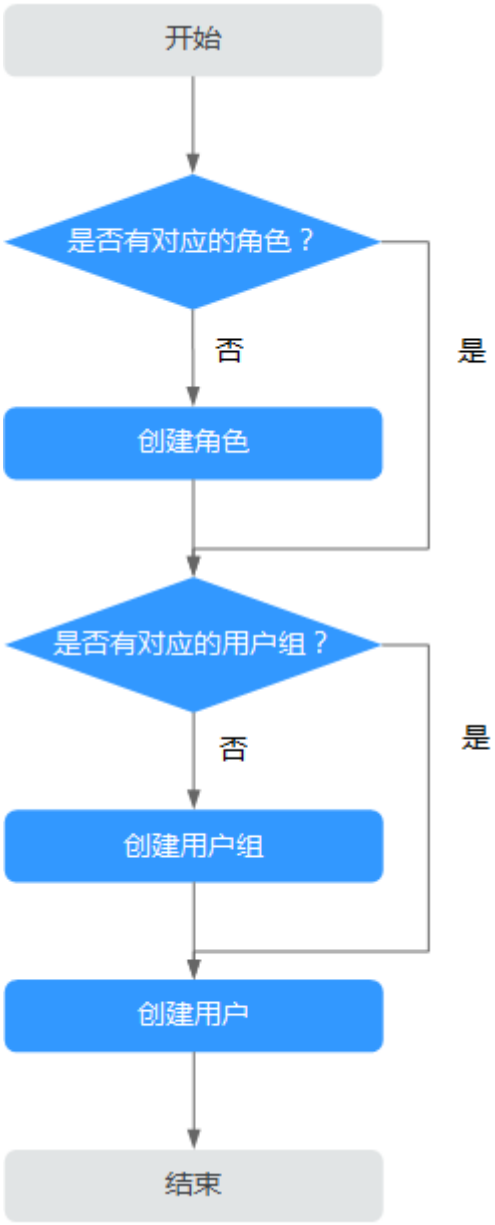
流程概述

在实际业务中，MRS集群用户需要先明确大数据的业务场景，规划集群用户对应的权限。然后在Manager界面创建角色，并设置角色包含的权限以匹配业务的需求。如果需要统一管理单个或多个相同业务场景中的用户，Manager提供了用户组的功能，管理员可以创建用户组。

说明

如果角色设置HDFS、HBase、Hive或Yarn各组件的权限，仅可以使用组件自身功能。如果还需要使用Manager，请在角色中添加对应的Manager权限。

图 12-1 创建用户流程示意



12.13.2 开启 Kerberos 认证集群中的默认用户清单

用户分类

MRS集群提供以下3类用户，请用户定期修改密码，不建议使用默认密码。

用户类型	使用说明
系统用户	● 通过Manager创建，是MRS集群操作运维与业务场景中主要使用的用户，包含两种类型： – “人机”用户：用于在Manager的操作运维场景，以及在组件客户端操作的场景。 – “机机”用户：用于MRS集群应用开发的场景。 ● 用于OMS系统进程运行的用户。
系统内部用户	MRS集群提供的用于进程通信、保存用户组信息和关联用户权限的内部用户。
数据库用户	● 用于OMS数据库管理和数据访问的用户。 ● 用于业务组件（Hive、Hue、Loader和DBservice）数据库的用户。

系统用户

说明

- MRS集群需要使用操作系统中ldap用户，此账号不能删除，否则可能导致集群无法正常工作。密码管理策略由操作用户维护。
- 首次修改“ommdba”和“omm”用户需要执行重置密码操作。找回密码后建议定期修改。

类别	用户名称	初始密码	描述
MRS集群系统管理员	admin	在集群创建时由用户指定。	Manager的管理员。 此外还具有以下权限： ● 具有HDFS、ZooKeeper普通用户的权限。 ● 具有提交、查询Mapreduce、YARN任务的权限，以及YARN队列管理权限和访问YARN WebUI的权限。 ● Storm中，具有提交、查询、激活、去激活、重分配、删除拓扑的权限，可以操作所有拓扑。 ● Kafka服务中，具有创建、删除、授权、Reassign、消费、写入、查询主题的权限。
MRS集群节点操作系统用户	omm	系统随机生成	MRS集群系统的内部运行用户。在全部节点生成，属于操作系统用户，无需设置为统一的密码。

类别	用户名称	初始密码	描述
MRS集群节点操作系统用户	root	用户设置的密码。	MRS集群所属节点的登录用户。在全部节点生成，属于操作系统用户。

系统内部用户

说明

以下系统内部用户不能删除，否则可能导致集群或组件无法正常工作。

类别	默认用户	初始密码	描述
组件运行用户	hdfs	Hdfs@123	HDFS系统管理员，用户权限： 1. 文件系统操作权限： ● 查看、修改、创建文件 ● 查看、创建目录 ● 查看、修改文件属组 ● 查看、设置用户磁盘配额 2. HDFS管理操作权限： ● 查看webUI页面状态 ● 查看、设置HDFS主备状态 ● 进入、退出HDFS安全模式 ● 检查HDFS文件系统
	hbase	Hbase@123	HBase系统管理员，用户权限： ● 集群管理权限: 表的Enable、Disable操作，触发MajorCompact，ACL操作 ● 授权或回收权限，集群关闭等操作相关的权限 ● 表管理权限: 建表、修改表、删除表等操作权限 ● 数据管理权限：表级别、列族级别以及列级别的数据读写权限 ● 访问HBase WebUI的权限

类别	默认用户	初始密码	描述
	mapred	Mapred@123	MapReduce系统管理员，用户权限： ● 提交、停止和查看MapReduce任务的权限 ● 修改Yarn配置参数的权限 ● 访问Yarn、MapReduce WebUI的权限
	spark	Spark@123	Spark系统管理员，用户权限： ● 访问Spark WebUI的权限 ● 提交Spark任务的权限

用户组信息

默认用户组	描述
hadoop	将用户加入此用户组，可获得所有Yarn队列的任务提交权限。
hbase	普通用户组，将用户加入此用户组不会获得额外的权限。
hive	将用户加入此用户组，可以使用Hive。
spark	普通用户组，将用户加入此用户组不会获得额外的权限。
supergroup	将用户加入此用户组，可获得HBase、HDFS和Yarn的管理员权限，并可以使用Hive。
check_sec_ldap	用于内部测试主LDAP是否工作正常。用户组随机存在，每次测试时创建，测试完成后自动删除。系统内部组，仅限组件间内部使用。
Manager_tenant	租户系统用户组。系统内部组，仅限组件间内部使用。
System_administrator	MRS集群系统管理员组。系统内部组，仅限组件间内部使用。
Manager_viewer	MRS Manager系统查看员组。系统内部组，仅限组件间内部使用。
Manager_operator	MRS Manager系统操作员组。系统内部组，仅限组件间内部使用。
Manager_auditor	MRS Manager系统审计员组。系统内部组，仅限组件间内部使用。

默认用户组	描述
Manager_administrator	MRS Manager系统管理员组。系统内部组，仅限组件间内部使用。
compcommon	MRS系统内部组，用于访问集群公共资源。所有系统用户和系统运行用户默认加入此用户组。
default_1000	为租户创建的用户组。系统内部组，仅限组件间内部使用。
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限,才能访问对应Topic。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
storm	Storm的普通用户组，属于该组的用户拥有提交拓扑和管理属于自己的拓扑的权限。
stormadmin	Storm的管理员用户组，属于该组的用户拥有提交拓扑和管理所有拓扑的权限。
opentsdb	普通用户组，将用户加入此用户组不会获得额外的权限。
presto	普通用户组，将用户加入此用户组不会获得额外的权限。
flume	普通用户组，添加到该用户组的用户无任何额外权限。
launcher-job	MRS系统内部组，用于使用V2接口提交作业。

操作系统用户组	描述
wheel	MRS集群系统内部运行用户“omm”的主组。
ficommon	MRS集群系统公共组，对应“compcommon”，可以访问集群在操作系统中保存的公共资源文件。

数据库用户

MRS集群系统数据库用户包含OMS数据库用户、DBService数据库用户。

说明

数据库用户不能删除，否则可能导致集群或组件服务无法正常工作。

类别	默认用户	初始密码	描述
OMS数据库	ommdba	dbChangeMe@123456	OMS数据库管理员用户，用于创建、启动和停止等维护操作
	omm	ChangeMe@123456	OMS数据库数据访问用户
DBService数据库	omm	dbserverAdmin@123	DBService组件中GaussDB数据库的管理员用户
	hive	HiveUser@	Hive连接DBService数据库用户
	hue	HueUser@123	Hue连接DBService数据库用户
	sqoop	SqoopUser@	Loader连接DBService数据库的用户
	ranger	RangerUser@	Ranger连接DBService数据库的用户

12.13.3 创建角色

操作场景

该任务指导用户在Manager创建角色，并对Manager和组件进行授权管理。

Manager支持的角色数为1000。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

前提条件

- 用户已明确业务需求。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，选择“系统设置 > 角色管理”。

步骤3 单击“添加角色”，然后在“角色名称”和“描述”输入角色名字与描述。

“角色名称”为必选参数，字符长度为3到30，可以包含数字、字母和下划线。“描述”为可选参数。

步骤4 设置角色“权限”。

1. 单击“服务名称”，并选择一个“视图名称”。
2. 勾选一个或多个权限。

说明

- “权限”为可选参数。
- 在选择“视图”设置组件的权限时，可通过右上角的“搜索”框输入资源名称，然后单击 显示搜索结果。
- 搜索范围仅包含当前权限目录，无法搜索子目录。搜索关键字支持模糊搜索，不区分大小写。支持搜索下一页的结果。

表 12-46 Manager 权限描述

支持权限管理的资源	权限设置说明
“Alarm”	Manager告警功能授权，勾选“View”表示可以查看告警，勾选“Management”表示可以管理告警。
“Audit”	Manager审计日志功能授权，勾选“View”表示可以查看审计，勾选“Management”表示可以管理审计。
“Dashboard”	Manager概览功能授权，勾选“View”表示可以查看集群概览。
“Hosts”	Manager集群节点管理功能授权，勾选“View”表示可以查看节点，勾选“Management”表示可以管理节点。
“Services”	MRS集群服务管理功能授权，勾选“View”表示可以查看服务，勾选“Management”表示可以管理服务。
“System_cluster_management”	MRS集群管理授权，勾选“Management”表示可以使用MRS补丁管理功能。
“System_configuration”	MRS集群配置功能授权，勾选“Management”表示可以使用Manager配置MRS集群。
“System_task”	MRS集群任务功能授权，勾选“Management”表示可以使用Manager管理MRS集群的周期任务。
“Tenant”	Manager多租户管理功能授权，勾选“Management”表示可以查看Manager的租户管理页面。

表 12-47 HBase 权限描述

支持权限管理的资源	权限设置说明
“SUPER_USER_GROUP”	选中时表示授予HBase管理员权限。
“Global”	HBase的一种资源类型，表示HBase整体组件。

支持权限管理的资源	权限设置说明
“Namespace”	HBase的一种资源类型，表示命名空间，用来保存HBase表。具体权限： ● “Admin”：表示管理此命名空间的权限。 ● “Create”：表示在此命名空间创建HBase表的权限。 ● “Read”：表示访问此命名空间的权限。 ● “Write”：表示写入此命名空间数据的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“Table”	HBase的一种资源类型，表示数据表，用来保存数据。具体权限： ● “Admin”：表示管理此数据表的权限。 ● “Create”：表示在此数据表创建列族和列的权限。 ● “Read”：表示读取数据表的权限。 ● “Write”：表示写入数据到表的权限。 ● “Execute”：表示可执行协处理器（Endpoint）的权限。
“ColumnFamily”	HBase的一种资源类型，表示列族，用来保存数据。具体权限： ● “Create”：表示在此列族创建列的权限。 ● “Read”：表示读取列族的权限。 ● “Write”：表示写入数据到列族的权限。
“Qualifier”	HBase的一种资源类型，表示列，用来保存数据。具体权限： ● “Read”：表示读取列的权限。 ● “Write”：表示写入数据到列的权限。

HBase中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如命名空间“default”添加了“Read”和“Write”权限，则命名空间中的表、列族和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 12-48 HDFS 权限描述

支持权限管理的资源	权限设置说明
“Folder”	HDFS的一种资源类型，表示HDFS目录，可以保存文件或子目录。具体权限： ● “Read”：表示访问此HDFS目录的权限。 ● “Write”：表示在此HDFS目录写入数据的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。
“Files”	HDFS的一种资源类型，表示HDFS中的文件。具体权限： ● “Read”：表示访问此文件的权限。 ● “Write”：表示写入此文件的权限。 ● “Execute”：表示执行操作的权限。在添加访问或写入权限必须同时勾选。

HDFS中每一级目录的权限默认不会传递到下级目录类型。例如目录“tmp”添加了“Read”和“Execute”，需要同时勾选“递归”才能为子目录添加权限。

表 12-49 Hive 权限描述

支持权限管理的资源	权限设置说明
“Hive Admin Privilege”	选中时表示授予Hive管理员权限。
“Database”	Hive的一种资源类型，表示Hive数据库，用来保存Hive表。具体权限： ● “Select”：表示查询Hive数据库的权限。 ● “Delete”：表示在Hive数据库执行删除操作的权限。 ● “Insert”：表示在Hive数据库执行插入操作的权限。 ● “Create”：表示在Hive数据库执行创建操作的权限。

支持权限管理的资源	权限设置说明
“Table”	Hive的一种资源类型，表示Hive表，用来保存数据。具体权限： ● “Select”：表示查询Hive表的权限。 ● “Delete”：表示在Hive表执行删除操作的权限。 ● “Update”：表示为角色添加Hive表的“Update”权限。 ● “Insert”：表示在Hive表执行插入操作的权限。 ● “Grant of Select”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Select”权限。 ● “Grant of Delete”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Delete”权限。 ● “Grant of Update”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Update”权限。 ● “Grant of Insert”：选中表示属于此角色的用户可以使用Hive语句为其他用户添加“Insert”权限。

Hive中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如数据库“default”添加了“Select”和“Insert”权限，则数据库中的表和列自动添加该权限。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 12-50 YARN 权限描述

支持权限管理的资源	权限设置说明
“Cluster Admin Operations”	选中时表示授予YARN管理员权限。
“root”	YARN的根队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Parent Queue”	YARN的一种资源类型，表示父队列，可以包含子队列。根队列也属于父队列的一种。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。
“Leaf Queue”	YARN的一种资源类型，表示叶子队列。具体权限： ● “Submit”：表示在队列提交作业的权限。 ● “Admin”：表示管理当前队列的权限。

YARN中每一级资源类型的权限默认会传递到下级资源类型，但“递归”选项没有默认勾选。例如队列“root”添加了“Submit”权限，则子队列自动添加该权限。子队列继承的权限不在“权限”表格显示被选中。若设置父资源后，再手动设置子资源，则子资源的权限取父资源与当前子资源设置的并集。

表 12-51 Hue 权限描述

支持权限管理的资源	权限设置说明
“Storage Policy Admin”	选中时表示授予Hue中存储策略管理员权限。

步骤5 单击“确定”完成，返回“角色管理”。

----结束

相关任务

修改角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要修改角色所在的行，单击“修改”，修改角色信息。

说明

修改角色分配的权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除角色

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“角色管理”。

步骤3 在要删除角色所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

12.13.4 创建用户组

操作场景

该任务指导用户通过Manager创建新用户组并指定其操作权限，使用户组可以统一管理加入用户组的单个或多个用户。用户加入用户组后，可获得用户组具有的操作权限。

Manager支持用户组数为100。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

前提条件

- 用户已明确业务需求，并已创建业务场景需要的角色。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户组管理”。

步骤4 在组列表上方，单击“添加用户组”。

步骤5 填写“组名”和“描述”。

“组名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。“描述”为可选参数。

步骤6 在“角色”，单击“选择添加角色”选择指定的角色并添加。

如果不添加角色，则当前创建的用户组没有使用MRS集群的权限。

步骤7 单击“确定”完成用户组创建。

----结束

相关任务

修改用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要修改用户组所在的行，单击“修改”，修改用户组信息。

说明

为用户组修改分配的角色权限，最长可能需要3分钟时间生效。

步骤4 单击“确定”完成修改操作。

----结束

删除用户组

步骤1 在MRS Manager，单击“系统设置”。

步骤2 在“权限配置”区域，单击“用户组管理”。

步骤3 在要删除用户组所在的行，单击“删除”。

步骤4 单击“确定”完成删除操作。

----结束

12.13.5 创建用户

操作场景

该任务指导用户根据实际业务场景需要，通过Manager创建新用户并指定其操作权限以满足业务使用。

Manager支持的用户数为1000。

如需对新创建用户的密码使用新的密码策略，请先**修改密码策略**，再参考本章节创建用户。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

前提条件

- 已明确业务需求，并已创建业务场景需要的角色和用户组。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在用户列表上方，单击“添加用户”。

步骤5 根据界面提示配置参数，填写“用户名”。

说明

- 不支持创建两个名称相同但大小写不同的用户。例如已创建用户“User1”，无法创建用户“user1”。
- 使用已创建的用户时，请输入和用户名完全一样的大小写字符。
- “用户名”为必选参数，字符长度为3到20，可以包含数字、字母和下划线。
- “root”、“omm”和“ommdba”为系统保留用户，请选择其他用户名。

步骤6 设置“用户类型”，可选值包括“人机”和“机机”。

- “人机”用户：用于在MRS Manager的操作运维场景，以及在组件客户端操作的场景。选择该值需同时填写“密码”和“确认密码”。
- “机机”用户：用于MRS应用开发的场景。选择该值用户密码随机生成，无需填写。

步骤7 在“用户组”，单击“选择添加的用户组”，选择对应用户组将用户添加进去。

说明

- 如果用户组添加了角色，则用户可获得对应角色中的权限。
- 为新用户分配Hive的权限，请将用户加入hive组。
- 如果用户需要管理租户资源，用户组必须分配了Manager_tenant角色以及租户对应的角色。
- 通过Manager创建的用户无法添加到通过IAM用户同步功能同步的用户组中。

步骤8 在“主组”选择一个组作为用户创建目录和文件时的主组。下拉列表包含“用户组”中勾选的全部组。

步骤9 根据业务实际需要在“分配角色权限”，单击“选择并绑定角色”为用户添加角色。

说明

- 创建用户时，如果用户从用户组获得的权限还不满足业务需要，则可以再分配其他已创建的角色。为新用户分配角色授权，最长可能需要3分钟时间生效。
- 创建用户时添加角色可细化用户的权限。
- 没有为新用户分配角色时，此用户可以访问HDFS、HBase、Yarn、Spark和Hue的WebUI。

步骤10 根据业务实际需要“描述”。

“描述”为可选参数。

步骤11 单击“确定”完成用户创建。

第一次在MRS集群中使用新创建的用户，例如登录Manager或者使用集群客户端，需要修改密码，具体请参见**修改操作用户密码**。

----结束

12.13.6 修改用户信息

操作场景

该任务指导用户在Manager修改已创建的用户信息，包括修改用户组、主组、角色和描述。

开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要修改用户所在的行，单击“修改”，修改用户信息。

说明

为用户修改用户组或分配的角色权限，最长可能需要3分钟时间生效。

步骤5 单击“确定”完成修改操作。

----结束

12.13.7 锁定用户

该任务指导用户将MRS集群中的用户锁定。用户被锁定后，不能在Manager重新登录或在集群中重新进行安全认证。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

可通过以下两种方式锁定用户，锁定后的用户需要手动解锁或者等待锁定时间结束才能恢复使用：

- 自动锁定：通过设置密码策略中的“允许输入错误次数”，将超过登录失败次数的用户自动锁定。具体操作请参见**修改密码策略**。
- 手动锁定：通过手动的方式将用户锁定。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

以下将具体介绍手动锁定。不支持锁定“机机”用户。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要锁定用户所在行，单击“锁定用户”，锁定用户。

步骤5 在弹出的提示窗口，单击“确定”完成锁定操作。

----结束

12.13.8 解锁用户

在用户输入错误密码次数大于允许输入错误次数，造成用户被锁定或者用户被手动锁定后需要解锁用户的场景下，用户可以通过Manager为锁定的用户解锁。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要解锁用户所在行，选择“解锁用户”，解锁用户。

步骤5 在弹出的提示窗口，单击“确定”完成解锁操作。

----结束

12.13.9 删除用户

MRS集群用户不再需要使用时，可以在MRS Manager中删除此用户。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持删除用户操作。

说明

- 如果删除用户A后，再次准备重新创建同名用户A，如果该用户A已经提交过作业（客户端提交或者MRS console页面提交），那么需要在删除该用户A的同时，删除该用户A残留的文件夹，否则使用重新创建的同名用户A提交作业会失败。
- 删除用户残留文件夹操作方法为：依次登录MRS集群的Core节点，在每个Core节点上执行如下两条命令，其中如下命令中“$user”为具体的以用户名命名的文件夹。

  cd /srv/BigData/hadoop/data1/nm/localdir/usercache/

  rm -rf $user
- 该章节操作仅适用于**MRS 3.x**之前版本集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要删除用户所在的行，选择“更多 > 删除”。

步骤5 单击“确定”完成删除操作。

----结束

12.13.10 修改操作用户密码

操作场景

出于MRS集群安全的考虑，“人机”类型系统用户密码必须定期修改。该任务指导用户通过MRS Manager完成修改自身密码工作。

如需对用户修改的密码使用新的密码策略，请先**修改密码策略**，再参考本章节修改密码。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

对系统的影响

修改MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

前提条件

- 已获取当前的密码策略。
- 已获取MRS Manager访问地址。
- 开启Kerberos认证的集群或开启弹性公网IP功能的普通集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，移动鼠标到界面右上角的。

在弹出菜单，选择“修改密码”。

步骤3 分别输入“旧密码”、“新密码”、“确认新密码”，单击“确定”完成修改。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.13.11 初始化系统用户密码

操作场景

该任务指导在用户遗忘密码或公共账号密码需要定期修改时，通过Manager初始化密码。初始化密码后用户首次使用需要修改密码。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

对系统的影响

初始化MRS集群用户密码后，如果以前下载过用户认证文件，则需要重新下载并获取keytab文件。

初始化“人机”用户密码

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在要初始化密码用户所在行，单击“更多 > 初始化密码”，按界面提示信息修改用户密码。

在弹出窗口中输入当前用户密码确认身份，单击“确定”，然后在“初始化密码”单击“确定”。

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

初始化“机机”用户密码

步骤1 根据业务情况，准备好客户端，并登录安装客户端的节点。

步骤2 执行以下命令切换用户。

sudo su - omm

步骤3 执行以下命令，切换到客户端目录，例如“/opt/Bigdata/client”。

cd /opt/Bigdata/client

步骤4 执行以下命令，配置环境变量。

source bigdata_env

步骤5 执行以下命令，使用kadmin/admin登录控制台。

说明

kadmin/admin的默认密码为“KAdmin@123”，首次登录后会提示该密码过期，请按照提示修改密码并妥善保存。

kadmin -p kadmin/admin

步骤6 执行以下命令，重置组件运行用户密码。此操作对所有服务器生效。

cpw *组件运行用户名*

例如：**cpw oms/manager**

集群中，默认的密码复杂度要求：

- 密码字符长度为8～32位。
- 至少需要包含大写字母、小写字母、数字、空格、特殊字符'~!@#$%^&*()-_=+\|[{}];:'",<.>/?中的3种类型字符。
- 不能与用户名或倒序的用户名相同。

----结束

12.13.12 下载用户认证文件

操作场景

用户开发大数据应用程序并在支持Kerberos认证的MRS集群中运行此程序时，需要准备访问MRS集群的“机机”用户认证文件。认证文件中的keytab文件可用于认证用户身份。

该任务指导用户通过Manager下载“机机”用户认证文件并导出keytab文件。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

说明

- 如果选择下载“人机”用户的认证文件，在下载前需要使用Manager修改过一次此用户的密码使设置的初始密码失效，否则导出的keytab文件无法使用。请参见**修改操作用户密码**。
- 该章节操作仅适用于**MRS 3.x**之前版本集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 在“权限配置”区域，单击“用户管理”。

步骤4 在需导出keytab文件用户所在的行，选择“更多 > 下载认证凭据”下载认证文件，待文件自动生成后指定保存位置，并妥善保管该文件。

步骤5 使用解压程序打开认证文件。

- “user.keytab”表示用户keytab文件，用于认证用户身份。
- “krb5.conf”表示认证服务器配置文件，应用程序在进行用户认证身份时根据该文件的配置信息连接认证服务器。

----结束

12.13.13 修改密码策略

操作场景

须知

密码策略涉及用户管理的安全性，请根据业务安全要求谨慎修改，否则会有安全性风险。

该任务指导用户设置密码安全规则、用户登录安全规则及用户锁定规则。由于“机机”用户密码随机生成，在MRS Manager设置密码策略只影响“人机”用户。开启Kerberos认证的集群或开启弹性公网IP功能的普通集群支持该操作。

如需对新增用户的密码或用户修改的密码使用新的密码策略，请先参考本章节修改密码策略，再**创建用户**或**修改密码**。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

操作步骤

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 在MRS Manager，单击“系统设置”。

步骤3 单击“密码策略配置”。

步骤4 根据界面提示，修改密码策略，具体参数见下表。

表 12-52 密码策略参数说明

参数名称	描述
最小密码长度	密码包含的最小字符个数，取值范围是8～32。默认值为“8”。
字符类型的数目	密码字符包含大写字母、小写字母、数字、空格和特殊符号（包含~`!?,.:;-_'(){}[]/<>@#$%^&*+|\=）的最小种类。可选择数值为“3”和“4”。默认值“3”表示至少必须使用大写字母、小写字母、数字、特殊符号和空格中的任意3种。
密码有效期（天）	密码有效使用天数，取值范围0～90，0表示永久有效。默认值为“90”。
密码失效提前提醒天数	提前一段时间提醒密码即将失效。设置后，若集群时间和该用户密码失效时间的差小于该值，则说明用户进入密码失效提醒期。用户登录MRS Manager时会提示用户密码即将过期，是否需要修改密码。取值范围为“0”-“X”，（“X”为密码有效期的一半，向下取整）。“0”表示不提醒。默认值为“5”。
认证失败次数重置时间间隔（分钟）	密码输入错误次数保留的时间间隔（分钟），取值范围为0～1440。“0”表示永远有效，“1440”表示1天。默认值为“5”。
密码连续错误次数	用户输入错误密码超过配置值后将锁定，取值范围为3～30。默认值为“5”。
用户锁屏时间（分钟）	满足用户锁定条件时，用户被锁定的时长，取值范围为5～120。默认值为“5”。

----结束

12.13.14 配置跨集群互信

操作场景

集群A需要访问另一个集群B的资源前，需要用户为这两个集群设置互信。

如果未配置跨集群互信，每个集群资源仅能被本集群用户访问。MRS自动为每个集群定义一个唯一且不重复的“域名”，用于表示用户的基本使用范围。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

对系统的影响

- 配置跨集群互信后，外部集群的用户可以在本集群中跨域使用，请根据业务与安全要求，定期检视集群中用户的权限。

- 安全集群配置跨集群互信，需要重启KrbServer服务，集群在重启期间无法使用。
- 配置跨集群互信后，互信的两个集群均会增加内部用户“krbtgt/*本集群域名@外部集群域名*”、“krbtgt/*外部集群域名@本集群域名*”，不支持删除。密码默认为“Crossrealm@123”。

操作步骤

步骤1 在MRS管理控制台，分别查看两个集群的所有安全组。

- 当两个集群的安全组相同，请执行**步骤3**。
- 当两个集群的安全组不相同，请执行**步骤2**。

步骤2 在VPC管理控制台，选择“访问控制 > 安全组”，在对应安全组名称的操作列单击“配置规则”，分别为每个安全组添加规则。

在入方向规则页签，单击“添加规则”，添加入方向规则。

- 优先级：优先级可选范围为1-100，默认值为1，即最高优先级。优先级数字越小，规则优先级级别越高。
- 策略：允许。
- 协议端口：基本协议/全部协议。
- 协议：根据实际情况选择“IPv4”或“IPv6”。
- “源地址”为“安全组”且是对端集群的安全组。
  - 为A集群的安全组添加入方向规则，源地址选择B集群（对端集群）的安全组。
  - 为B集群的安全组添加入方向规则，源地址选择A集群（对端集群）的安全组。

说明

未开启Kerberos认证的普通集群执行**步骤1**~**步骤2**即可完成跨集群互信配置，开启Kerberos认证的安全集群请继续执行后续步骤进行配置。

步骤3 参见**访问MRS Manager（MRS 2.x及之前版本）**分别登录两个集群MRS Manager，单击“服务管理”，查看全部组件的“健康状态”结果，是否全为“良好”？

- 是，执行**步骤4**。
- 否，任务结束，联系支持人员检查状态。

步骤4 查看配置信息。

1. 分别在两个集群MRS Manager，选择“服务管理 > KrbServer > 实例”，查看两个KerberosServer部署主机的“管理IP”。
2. 单击“服务配置”，将“基础配置”切换为“全部配置”并在左侧导航树上选择“KerberosServer > 端口”，查看“kdc_ports”的值，默认值为“21732”。
3. 单击“域”，查看“default_realm”的值。

步骤5 在其中一个集群的MRS Manager，修改配置参数“peer_realms”。

表 12-53 相关参数

参数名	描述
“realm_name”	填写互信集群的域名，即**步骤4**中获得的“default_realm”的值。
“ip_port”	填写互信集群的KDC地址，参数值格式为：*外部集群KerberosServer部署的节点IP地址:kdc_port*。 两个KerberosServer的IP地址使用逗号分隔，例如KerberosServer部署在10.0.0.1和10.0.0.2上，则对应参数值为“10.0.0.1:21732,10.0.0.2:21732”。

说明

- 如果需要配置与多个集群的互信关系，请单击添加新项目，并填写参数值。删除多余的配置项请单击。
- 最多支持与16个集群配置互信，且本集群的不同互信集群之间默认不存在互信关系，需要另外添加。

步骤6 单击“保存配置”，在弹出窗口中勾选“重新启动受影响的服务或实例。”，单击“确定”重启服务。若未勾选“重新启动受影响的服务或实例。”，请手动重启受影响的服务或实例。

界面提示“操作成功”，单击“完成”，服务成功启动。

步骤7 退出MRS Manager，重新登录正常表示配置已成功。

步骤8 在另外一个集群的MRS Manager，重复**步骤5**到**步骤7**。

----结束

后续操作

配置跨集群互信后，因在MRS Manager修改了服务配置参数并重启了服务，请重新准备好客户端配置文件并更新客户端。

场景1：

A集群和B集群（对端集群、互信集群）是同类型集群，例如均是分析集群或者流式集群，请参见**更新客户端（3.x之前版本）**分别更新客户端配置文件。

- 更新A集群的客户端配置文件。
- 更新B集群的客户端配置文件。

场景2：

A集群和B集群（对端集群、互信集群）是不同类型集群，请执行如下步骤分别更新对端集群的配置文件到本端集群和本端集群自身的配置文件。

- 将A集群的客户端配置文件更新到B集群上。
- 将B集群的客户端配置文件更新到A集群上。
- 更新A集群的客户端配置文件。

- 更新B集群的客户端配置文件。

步骤1 登录MRS Manager(A集群)。

步骤2 单击“服务管理”，然后单击“下载客户端”。

步骤3 “客户端类型”选择“仅配置文件”。

步骤4 “下载路径”选择“远端主机”。

步骤5 将“主机IP”设置为B集群的主Master节点IP地址，设置“主机端口”为“22”，并将“存放路径”设置为“/tmp”。

- 如果使用SSH登录B集群的默认端口“22”被修改，请将“主机端口”设置为新端口。
- “存放路径”最多可以包含256个字符。

步骤6 “登录用户”设置为“root”。

如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。

步骤7 在“SSH私钥”选择并上传创建B集群时使用的密钥文件。

步骤8 单击“确定”开始生成客户端文件。

若界面显示以下提示信息表示客户端包已经成功保存。单击“关闭”。

```
下载客户端文件到远端主机成功。
```

若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从**步骤2**执行重新开始下载客户端。

```
连接到服务器失败，请检查网络连接或参数设置。
```

步骤9 使用VNC方式，登录弹性云服务器（B集群）。参见 弹性云服务器用户指南的“实例 > 登录Windows弹性云服务器 > 远程登录（VNC方式）”。

请先参考**SSH密钥方式登录**登录弹性云服务器，设置云服务器密码，然后再采用VNC方式登录。

步骤10 执行以下命令切换到客户端目录，例如“/opt/Bigdata/client”：

cd /opt/Bigdata/client

步骤11 执行以下命令，将A集群的客户端配置更新到B集群上：

sh refreshConfig.sh *客户端安装目录 客户端配置文件压缩包完整路径*

例如，执行命令：

sh refreshConfig.sh /opt/Bigdata/client /tmp/MRS_Services_Client.tar

界面显示以下信息表示配置刷新更新成功：

```
ReFresh components client config is complete.
Succeed to refresh components client config.
```

说明

步骤**步骤1**~**步骤11**的操作也可以参考**更新客户端（3.x之前版本）**页面的方法二操作。

步骤12 参见**步骤1**~**步骤11**，将B集群的客户端配置文件更新到A集群上。

步骤13 参见**更新客户端（3.x之前版本）**，分别更新本端集群自身的客户端配置文件：

- 更新A集群的客户端配置文件。
- 更新B集群的客户端配置文件。

----结束

12.13.15 配置并使用互信集群的用户

操作场景

配置完跨集群互信后，需要在互信的集群上设置用户的权限，这样本集群中的用户才能访问互信集群中同名用户可访问的资源。

说明

该章节操作仅适用于**MRS 3.x**之前版本集群。

前提条件

已完成跨集群互信配置，然后刷新两个集群的客户端。

操作步骤

步骤1 在集群A的MRS Manager，选择“系统设置 > 用户管理”，检查互信集群B的用户，是否在A集群中已存在相同名字用户。

- 是，执行**步骤2**。
- 否，执行**步骤3**。

步骤2 单击用户名左侧的 展开用户的详细信息，检查该用户所在的用户组和角色分配的权限是否满足本次业务需求。

例如，集群A的“admin”用户拥有查看本集群HDFS中目录“/tmp”并创建文件的权限，然后执行**步骤4**。

步骤3 创建业务所需要使用的用户，同时关联业务所需要的用户组或者角色。然后执行**步骤4**。

步骤4 选择“服务管理 > HDFS > 实例”，查看“NameNode(主)”的“管理IP”。

步骤5 登录集群B的客户端节点。

例如在Master2节点更新客户端，则在该节点登录客户端，具体参见**使用MRS客户端**。

步骤6 执行以下命令，查看集群A中的目录“/tmp”。

hdfs dfs -ls hdfs://192.168.6.159:9820/tmp

其中，**192.168.6.159**是集群A中主NameNode的IP地址，9820是客户端与NameNode通信的默认端口。

说明

MRS 1.6.2及以前版本该默认端口为25000。详见**开源组件端口列表**

步骤7 执行以下命令，在集群A中的目录“/tmp”创建一个文件。

hdfs dfs -touchz hdfs://192.168.6.159:9820/tmp/mrstest.txt

访问集群A，在目录“/tmp”中可查询到mrstest.txt文件，则表示配置跨集群互信成功。

----结束

12.13.16 配置 MRS 多用户访问 OBS 细粒度权限

开启细粒度权限时，用户通过该指导配置访问OBS权限，实现MRS用户对OBS文件系统下的目录权限控制。

说明

该章节内容仅适用于MRS 2.x及之前版本（MRS 1.9.2版本不适用）。

如需对MRS的用户访问OBS的资源进行详细控制，可通过该功能实现。例如，您只允许用户组A访问某一OBS文件系统中的日志文件，您可以执行以下操作来实现：

1. 为MRS集群配置OBS访问权限的委托，实现使用ECS自动获取的临时AK/SK访问OBS。避免了AK/SK直接暴露在配置文件中的风险。
2. 在IAM中创建一个只允许访问某一OBS文件系统中的日志文件的策略，并创建一个绑定该策略权限的委托。
3. 在MRS集群中，新建的委托与MRS集群中的用户组A进行绑定，即可实现用户组A只拥有访问某一OBS文件系统中的日志文件的权限。

在以下场景运行作业时，提交作业的用户名为内置用户名，无法实现MRS多用户访问OBS：

- spark-beeline在安全集群中提交作业的内置用户名为spark，在普通集群中提交作业的内置用户名为omm。
- hbase shell在安全集群提交作业的内置用户名为hbase，在普通集群中提交作业的内置用户名为omm。
- Presto在安全集群提交作业的内置用户名为omm、hive，在普通集群提交作业的内置用户名为omm（当通过“组件管理 > Presto > 服务配置”，选择“全部配置”并搜索修改参数hive.hdfs.impersonation.enabled的值为true可以实现MRS多用户访问OBS细粒度权限功能）。

前提条件

- 开启细粒度权限控制的用户，权限管理请参考**创建MRS操作用户**。
- 需要对IAM的委托（请参考IAM服务用户指南 > 委托）和OBS细粒度策略有一定了解。

步骤一：给集群配置有 OBS 访问权限的委托

步骤1 请参考**配置存算分离集群（委托方式）**配置OBS访问权限的委托。

配置的委托对该集群上所有用户（包括内置用户）及用户组生效，如需对集群上的用户及用户组访问OBS的权限进行控制请继续执行后续步骤。

----结束

步骤二：在 IAM 服务创建策略及委托

创建拥有不同访问权限的策略，并将策略与委托进行绑定，具体操作请参考**在IAM服务创建策略及委托**。

步骤三：在 MRS 集群详情页面配置 OBS 权限控制映射关系

步骤1 在MRS控制台，选择“集群列表 > 现有集群”并单击集群名称。

步骤2 在“概览”页签的基本信息区域，单击“OBS权限控制”右侧的“单击管理”。

步骤3 单击“添加映射”，并参考**表12-54**配置相关参数。

表 12-54 OBS 权限控制参数

参数	说明
IAM委托	选择**步骤2**中创建的委托。
类型	• User：在用户级别进行映射 • Group：用户组级别进行映射 **说明** • 用户级别的映射优先级大于用户组级别的映射。若选择Group，建议在“MRS用户（组）”一栏，填写用户的主组名称。 • 请避免同个用户名（组）出现在多个映射记录上的情况。
MRS 用户（组）	MRS中的用户（组）的名称，以英文逗号进行分隔。 **说明** • 对于没有配置在OBS权限控制的用户，且没有配置AK、SK时，将以MRS_ECS_DEFAULT_AGENCY中的OBS Operator的权限访问OBS。对于组件内置用户不建议绑定在委托中。 • 如需对组件内置用户在以下场景提交作业时配置委托，要求如下： – 如需对spark-beeline的操作进行权限控制，安全集群时配置用户名“spark”，普通集群时配置用户名“omm”。 – 如需对hbase shell的操作进行权限控制，安全集群时配置用户名“hbase”，普通集群时配置用户名“omm”。 – 如需对Presto的操作进行权限控制，安全集群时配置用户名“omm”、“hive”和登录客户端的用户名，普通集群时配置用户名“omm”和登录客户端的用户名。 – 如需使用Hive在beeline模式下创建表时，配置内置用户“hive”。

步骤4 单击“确定”。

步骤5 勾选“我同意授权MRS用户（组）与IAM委托之间的信任关系。”，并单击“确定”，完成MRS用户与OBS权限的映射关系。

当集群详情页面“概览”页签的“OBS权限控制”后出现 或OBS权限控制的映射表已刷新，表示映射生效（过程大约需要1分钟）。

在关系列表的“操作”列可以对已添加的关系进行编辑和删除。

说明

- 对于没有配置在OBS权限控制的用户，且没有配置AK、SK时，将以集群配置的委托在“对象存储服务”项目下所拥有的权限访问OBS。
- 无论用户是否配置OBS权限控制，只要配置AK、SK时，将以AK、SK的权限访问OBS。
- 映射关系的修改、创建、删除需要用户有Security Administrator权限。
- 修改映射关系后，若想使之在spark-beeline中生效，需要重启Spark服务，若想使之在hive beeline中生效，需要退出beeline重新进入，若想使之在Presto服务中生效，需要重启Presto服务。

----结束

在开启 OBS 权限控制功能时各组件访问 OBS 的说明

步骤1 以root用户登录集群任意一个节点，密码为用户创建集群时设置的root密码。

步骤2 配置环境变量（MRS 3.x及之后版本客户端默认安装路径为“/opt/Bigdata/client”，MRS 3.x之前版本为“/opt/client”。具体以实际为准。）。

source /opt/Bigdata/client/bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

例如, **kinit admin**

步骤4 如果当前集群未启用Kerberos认证，执行如下命令登录执行操作的用户，该用户需要属于supergroup组，创建用户可参考创建用户，将XXXX替换成用户名。

mkdir /home/XXXX

chown XXXX /home/XXXX

su - XXXX

步骤5 访问OBS。无需再配置AK、SK和endpoint。OBS路径格式：obs://buck_name/XXX。

例如：**hadoop fs -ls "obs://obs-example/job/hadoop-mapreduce-examples-3.1.2.jar"**

说明

- 如需使用hadoop fs删除OBS上文件，请使用**hadoop fs -rm -skipTrash**来删除文件。
- spark-sql、spark-beeline在创建表时，若不涉及数据导入，则不会访问OBS。即若在一个无权限的OBS目录下创建表，CREATE TABLE仍会成功，但插入数据会报403 AccessDeniedException。

----结束

在 IAM 服务创建策略及委托

步骤1 在IAM服务创建策略。

1. 登录IAM服务控制台。
2. 单击“权限 > 创建自定义策略”。
3. 参考**表12-55**填写参数。常用的OBS自定义策略样例请参考。

表 12-55 策略参数

参数	说明
策略名称	只能包含如下字符：大小写字母、数字、空格和特殊字符（-_.,）。
作用范围	选择全局级服务，OBS为全局服务。
配置策略方式	选择可视化视图。
策略内容	1. “允许”选择“允许”。 2. “云服务”选择“对象存储服务(OBS)”。 3. “操作”勾选所有“写”、“列表”和“只读”权限。 4. “特定资源”选择： a. “object”选择“通过资源路径指定”，并单击“添加资源路径”分别输入路径***obs_bucket_name*/tmp/**和***obs_bucket_name*/tmp/***。此处以**/tmp**目录为例，如需其他目录权限请参考该步骤添加对应目录及该目录下所有对象的资源路径。 b. “bucket”选择“通过资源路径指定”，并单击“添加资源路径”输入路径***obs_bucket_name***。 5. （可选）请求条件，暂不添加。
策略描述	可选，对策略的描述。

说明

各个组件的写数据操作若通过rename的方式实现时，写数据时要配置删除对象的权限。

4. 单击“确定”保存策略。

步骤2 在IAM服务创建委托。

1. 登录IAM服务控制台。

2. 单击“委托 > 创建委托”。
3. 参考**表12-56**填写参数。

表 12-56 委托参数

参数	说明
委托名称	只能包含如下字符：大小写字母、数字、空格和特殊字符（ -_., ）。
委托类型	选择普通账号。
委托的账号	填写本用户的云账号，即使用手机号开通的账号，不能是联邦用户或者IAM用户。
持续时间	请根据需要选择。
描述	可选，对委托的描述。
权限选择	1. 在“项目”列对应的“对象存储服务”行，单击“操作”列的“修改”。 2. 勾选**步骤1**中创建的策略，使之出现在“已选择策略中”。 3. 单击“确定”。

4. 单击“确定”保存委托。

说明

当使用该委托访问过OBS后，再修改该委托及其绑定的策略时，最长需要等待15分钟，修改的内容才能生效。

----结束

12.14 补丁操作指导

12.14.1 MRS 1.7.0 前版本补丁操作指导

当您通过如下途径获知集群版本补丁信息，请根据您的实际需求进行补丁升级操作。

- 通过消息中心服务推送的消息获知MapReduce服务发布了补丁信息。
- 进入现有集群，查看“补丁信息”页面，呈现补丁信息。

安装补丁前准备

- 请参见**执行健康检查**检查集群状态，确认集群健康状态正常后再安装补丁。
- 已经将对应版本的集群补丁包上传至服务器，详见**上传补丁包**。
- 您根据“补丁内容”中的补丁信息描述，确认将要安装的目标补丁。

上传补丁包

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 单击“上传补丁”，分别设置如下参数

- 补丁文件路径:存储补丁包的OBS文件系统下创建的文件夹，例如MRS_1.6.2/MRS_1_6_2_11.tar.gz。
- 桶名:存储补丁包的OBS文件系统名称，例如mrs_patch。

  说明

  桶名和补丁文件路径可从管理面的“补丁信息”页签中获取，“补丁信息”页签中的“补丁路径”字段组成方式如下：【桶名】/【补丁文件路径】。
- AK：请参见“我的凭证 > 管理访问密钥”。
- SK：请参见“我的凭证 > 管理访问密钥”。

步骤4 单击“确定”，完成补丁的上传。

----结束

安装补丁

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“安装”。

步骤4 系统弹出提示框，单击“确定”，执行补丁的安装。

步骤5 补丁安装完成后，可在进度条中查看其安装状态，若安装失败，请联系MRS集群管理员处理。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“卸载”。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

12.14.2 支持滚动补丁

滚动补丁是指在补丁安装/卸载时，采用滚动重启服务（按批次重启服务或实例）的方式，在不中断或尽可能短地中断集群各个服务业务的前提下完成对集群中单个或多个服务的补丁安装/卸载操作。集群中的服务根据对滚动补丁的支持程度，分为三种：

- 支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的全部业务或部分业务（因服务而异，不同服务存在差别）不中断。
- 不支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的业务会中断。
- 部分角色支持滚动安装/卸载补丁的服务：在安装/卸载补丁过程中，服务的部分业务不中断。

当前MRS集群中，服务和实例是否支持滚动重启如**表12-57**所示。

表 12-57 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否

服务	实例	是否支持滚动重启
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	
	Logviewer	

安装补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁信息”页面，在操作列表中单击“安装”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动安装补丁功能开启：补丁安装前不会停止服务，补丁安装后滚动重启服务来完成补丁安装，可以最大程度减少对集群业务的影响，但相比普通方式安装耗时更久。
- 滚动安装补丁功能关闭：补丁安装前会停止服务，补丁安装后再重新启动服务来完成补丁安装，会造成集群和服务暂时中断，但相比滚动方式安装补丁耗时更短。
- 少于2个Master节点和少于3个Core节点的集群不支持滚动方式安装补丁。

步骤5 单击“确定”，安装目标补丁。

步骤6 查看补丁安装进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁安装进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

卸载补丁

步骤1 登录MRS管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中一集群并单击集群名，进入集群基本信息页面。

步骤3 进入“补丁信息”页面，在操作列表中单击“卸载”。

步骤4 进入“警告”页面，选择是否开启“滚动补丁”。

说明

- 滚动卸载补丁功能开启：补丁卸载前不会停止服务，补丁卸载后滚动重启服务来完成补丁卸载，可以最大程度减少对集群业务的影响，但相比普通方式卸载耗时更久。
- 滚动卸载补丁功能关闭：补丁卸载前会停止所有服务，补丁卸载后再重新启动所有服务来完成补丁卸载，会造成集群和服务暂时中断，但相比滚动方式卸载补丁耗时更短。
- 少于2个Master节点和少于3个Core节点的集群不支持滚动方式卸载补丁。

步骤5 单击“确定”，卸载目标补丁。

步骤6 查看补丁卸载进度。

1. 访问集群对应的MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。
2. 选择“系统设置 > 补丁管理”，进入补丁管理页面即可看到补丁卸载进度。

说明

对于集群中被隔离的主机节点，请参见**修复隔离主机补丁**进行补丁修复。

----结束

12.15 修复隔离主机补丁

若集群中存在主机被隔离的情况，集群补丁安装完成后，请参见本节操作对隔离主机进行补丁修复。修复完成后，被隔离的主机节点版本将与其他未被隔离的主机节点一致。

步骤1 访问MRS Manager，详细操作请参见**访问MRS Manager（MRS 2.x及之前版本）**。

步骤2 选择“系统设置 > 补丁管理”，进入补丁管理页面。

步骤3 在“操作”列表中，单击“详情”。

步骤4 在补丁详情界面，选中“Status”是“Isolated”的主机节点。

步骤5 单击“Select and Restore”，修复被隔离的主机节点。

----结束

12.16 支持滚动重启

在修改了大数据组件的配置项后，需要重启对应的服务来使得配置生效，使用普通重启方式会并发重启所有服务或实例，可能引起业务断服。为了确保服务重启过程中，尽量减少或者不影响业务运行，可以通过滚动重启来按批次重启服务或实例（对于有主备状态的实例，会先重启备实例，再重启主实例）。滚动重启方式的重启时间比普通重启时间久。

当前MRS集群中，服务和实例是否支持滚动重启如**表12-58**所示。

表 12-58 服务和实例是否支持滚动重启

服务	实例	是否支持滚动重启
HDFS	NameNode	是
	Zkfc	
	JournalNode	
	HttpFS	
	DataNode	
Yarn	ResourceManager	是
	NodeManager	
Hive	MetaStore	是
	WebHCat	
	HiveServer	
Mapreduce	JobHistoryServer	是
HBase	HMaster	是
	RegionServer	
	ThriftServer	
	RESTServer	
Spark	JobHistory	是
	JDBCServer	
	SparkResource	否
Hue	Hue	否
Tez	TezUI	否
Loader	Sqoop	否
Zookeeper	Quorumpeer	是
Kafka	Broker	是
	MirrorMaker	否
Flume	Flume	是
	MonitorServer	
Storm	Nimbus	是
	UI	
	Supervisor	

服务	实例	是否支持滚动重启
	Logviewer	

使用限制

- 请在低业务负载时间段进行滚动重启操作。
  - 例如：在滚动重启kafka服务时候， 如果kafka服务业务吞吐量很高（100M/s以上的情况下），会出现kafka服务滚动重启失败的情况。
  - 例如：在滚动重启HBase服务时候，如果原生界面上每个RegionServer上每秒的请求数超过1W，需要增大handle数来预防重启过程中负载过大导致的RegionServer重启失败。
- 重启前需要观察当前hbase的负载请求数（原生界面上每个rs的请求数如果超过1W，需要增大handle数来预防到时候负载不过来）
- 在集群Core节点个数小于6个的情况下，可能会出现业务短时间受影响的情况。
- 请优先使用滚动重启操作来重启实例或服务，并勾选“仅重启配置过期的实例”。

滚动重启服务

步骤1 在MRS Manager，单击“服务管理”，选择需要滚动重启的服务，进入服务页面。

步骤2 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤3 输入密码后，弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启服务。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启实例

步骤1 在MRS Manager，单击“服务管理”，选择需要滚动重启的服务，进入服务页面。

步骤2 在“实例”页签，勾选要重启的实例，单击“更多”，选择“滚动重启实例”。

步骤3 输入管理员密码后，弹出“滚动重启实例”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启实例。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启集群

步骤1 在MRS Manager，单击“服务管理”，进入服务管理页面。

步骤2 单击“更多”，选择“滚动重启集群”。

步骤3 输入管理员密码后，弹出“滚动重启集群”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启集群。

步骤4 滚动重启任务完成后，单击“完成”。

----结束

滚动重启参数说明

滚动重启参数说明如表12-59所示。

表 12-59 滚动重启参数说明

参数名称	描述
仅重启配置过期的实例	是否只重启集群内修改过配置的实例。
数据节点滚动重启并发数	采用分批并发滚动重启策略的数据节点实例每一个批次重启的实例数，默认为1，取值范围为1～20。只对数据节点有效。
批次时间间隔	滚动重启实例批次之间的间隔时间，默认为0，取值范围为0～2147483647，单位为秒。 说明：设置批次时间间隔参数可以增加滚动重启期间大数据组件进程的稳定性。建议设置该参数为非默认值，例如10。
批次容错阈值	滚动重启实例批次执行失败容错次数，默认为0，即表示任意一个批次的实例重启失败后，滚动重启任务终止。取值范围为0～214748364。

典型场景操作步骤

步骤1 在MRS Manager，单击“服务管理”，选择HBase，进入HBase服务页面。

步骤2 单击“服务配置”页签，修改HBase某个参数并保存配置，在出现如下弹窗后，单击“确定”进行保存。

说明

不要勾选“重新启动受影响的服务或实例”，该处重启是普通重启方式，会并发重启所有服务或实例，引起业务断服。

步骤3 保存配置完成后，单击“完成”。

步骤4 选择“服务状态”页签。

步骤5 在“服务状态”页签单击“更多”，选择“滚动重启服务”。

步骤6 输入管理员密码后，弹出“滚动重启服务”页面，勾选“仅重启配置过期的实例”，单击确认，开始滚动重启。

步骤7 滚动重启任务完成后，单击“完成”。

----结束

13 安全性说明

13.1 集群（未启用 Kerberos 认证）安全配置建议

Hadoop社区版本提供两种认证方式Kerberos认证（安全模式）和Simple认证（普通模式），在创建集群时，MRS支持配置是否启用Kerberos认证。

在安全模式下MRS集群统一使用Kerberos认证协议进行安全认证。

而普通模式下MRS集群各组件使用原生开源的认证机制，一般为Simple认证方式。而Simple认证，在客户端连接服务端的过程中，默认以客户端执行用户（例如操作系统用户“root”等）自动完成认证，MRS集群管理员或业务用户不显示感知认证。而且客户端在运行时，甚至可以通过注入UserGroupInformation来伪装成任意用户（包括superuser），集群资源管理接口和数据控制接口在服务端无认证和鉴权控制，很容易被黑客利用和攻击。

所以在普通模式下，必须通过严格限定网络访问权限来保障集群的安全。操作建议如下：

- 尽量将业务应用程序部署在同VPC和子网下的ECS中，避免通过外网访问MRS集群。
- 配置严格限制访问范围的安全组规则，禁止对MRS集群的入方向端口配置允许Any或0.0.0.0的访问规则。
- 如需从集群外访问集群内组件原生页面，请参考**创建连接MRS集群的SSH隧道并配置浏览器**进行配置。

13.2 安全认证原理和认证机制

功能

开启了 Kerberos认证的安全模式集群，进行应用开发时需要进行安全认证。

Kerberos作为安全认证的概念，使用Kerberos的系统在设计上采用“客户端/服务器”结构与AES等加密技术，并且能够进行相互认证（即客户端和服务器端均可对对方进行身份认证）。可以用于防止窃听、防止replay攻击、保护数据完整性等场合，是一种应用对称密钥体制进行密钥管理的系统。

结构

Kerberos的原理架构如图13-1所示，各模块的说明如表13-1所示。

图 13-1 原理架构

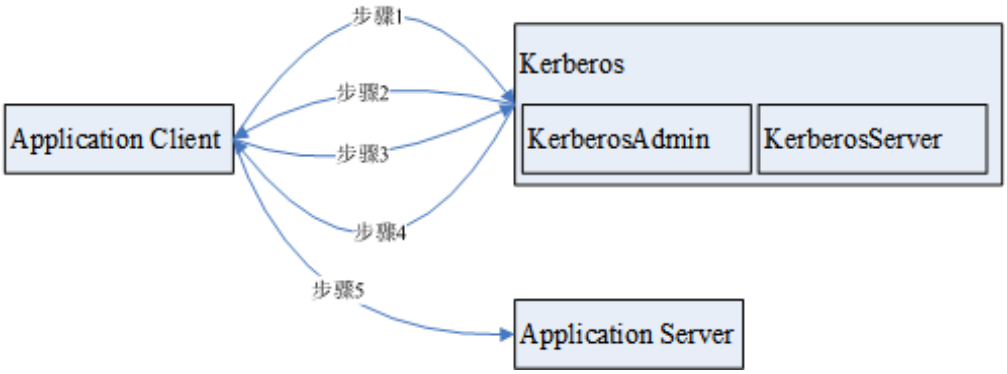


表 13-1 模块说明

模块	说明
Application Client	应用客户端，通常是需要提交任务（或者作业）的应用程序。
Application Server	应用服务端，通常是应用客户端需要访问的应用程序。
Kerberos	提供安全认证的服务。
KerberosAdmin	提供认证用户管理的进程。
KerberosServer	提供认证票据分发的进程。

步骤原理说明：

应用客户端（Application Client）可以是集群内某个服务，也可以是客户二次开发的一个应用程序，应用程序可以向应用服务提交任务或者作业。

1. 应用程序在提交任务或者作业前，需要向Kerberos服务申请TGT（Ticket-Granting Ticket），用于建立和Kerberos服务器的安全会话。
2. Kerberos服务在收到TGT请求后，会解析其中的参数来生成对应的TGT，使用客户端指定的用户名的密钥进行加密响应消息。
3. 应用客户端收到TGT响应消息后，解析获取TGT，此时，再由应用客户端（通常是rpc底层）向Kerberos服务获取应用服务端的ST（Server Ticket）。
4. Kerberos服务在收到ST请求后，校验其中的TGT合法后，生成对应的应用服务的ST，再使用应用服务密钥将响应消息进行加密处理。
5. 应用客户端收到ST响应消息后，将ST打包到发给应用服务的消息里面传输给对应的应用服务端（Application Server）。
6. 应用服务端收到请求后，使用本端应用服务对应的密钥解析其中的ST，并校验成功后，本次请求合法通过。

基本概念

以下为常见的基本概念，可以帮助用户减少学习Kerberos框架所花费的时间，有助于更好的理解Kerberos业务。以HDFS安全认证为例：

TGT

票据授权票据（ Ticket-Granting Ticket ），由Kerberos服务生成，提供给应用程序与Kerberos服务器建立认证安全会话，该票据的默认有效期为24小时，24小时后该票据自动过期。

TGT申请方式(以HDFS为例)：

1. 通过HDFS提供的接口获取。

```
/**
 * login Kerberos to get TGT, if the cluster is in security mode
 * @throws IOException if login is failed
 */
 private void login() throws IOException {
 // not security mode, just return
   if (! "kerberos".equalsIgnoreCase(conf.get("hadoop.security.authentication"))) {
       return;
   }

   //security mode
   System.setProperty("java.security.krb5.conf", PATH_TO_KRB5_CONF);

   UserGroupInformation.setConfiguration(conf);
   UserGroupInformation.loginUserFromKeytab(PRNCIPAL_NAME, PATH_TO_KEYTAB);
 }
```

2. 通过客户端shell命令以**kinit**方式获取。

ST

服务票据（ Server Ticket ），由Kerberos服务生成，提供给应用程序与应用服务建立安全会话，该票据一次性有效。

ST的生成在FusionInsight产品中，基于hadoop-rpc通信，由rpc底层自动向Kerberos服务端提交请求，由Kerberos服务端生成。

认证代码实例讲解

```
package com.xxx.bigdata.hdfs.examples;

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileStatus;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.security.UserGroupInformation;

public class KerberosTest {
    private static String PATH_TO_HDFS_SITE_XML = KerberosTest.class.getClassLoader().getResource("hdfs-site.xml")
            .getPath();
    private static String PATH_TO_CORE_SITE_XML = KerberosTest.class.getClassLoader().getResource("core-site.xml")
            .getPath();
    private static String PATH_TO_KEYTAB = KerberosTest.class.getClassLoader().getResource("user.keytab").getPath();
    private static String PATH_TO_KRB5_CONF = KerberosTest.class.getClassLoader().getResource("krb5.conf").getPath();
    private static String PRNCIPAL_NAME = "develop";
    private FileSystem fs;
```

```
    private Configuration conf;

    /**
     * initialize Configuration
     */
    private void initConf() {
        conf = new Configuration();

        // add configuration files
        conf.addResource(new Path(PATH_TO_HDFS_SITE_XML));
        conf.addResource(new Path(PATH_TO_CORE_SITE_XML));
    }

    /**
     * login Kerberos to get TGT, if the cluster is in security mode
     * @throws IOException if login is failed
     */
    private void login() throws IOException {
        // not security mode, just return
        if (! "kerberos".equalsIgnoreCase(conf.get("hadoop.security.authentication"))) {
            return;
        }

        //security mode
        System.setProperty("java.security.krb5.conf", PATH_TO_KRB5_CONF);

        UserGroupInformation.setConfiguration(conf);
        UserGroupInformation.loginUserFromKeytab(PRNCIPAL_NAME, PATH_TO_KEYTAB);
    }

    /**
     * initialize FileSystem, and get ST from Kerberos
     * @throws IOException
     */
    private void initFileSystem() throws IOException {
        fs = FileSystem.get(conf);
    }

    /**
     * An example to access the HDFS
     * @throws IOException
     */
    private void doSth() throws IOException {
        Path path = new Path("/tmp");
        FileStatus fStatus = fs.getFileStatus(path);
        System.out.println("Status of " + path + " is " + fStatus);
        //other thing
    }


    public static void main(String[] args) throws Exception {
        KerberosTest test = new KerberosTest();
        test.initConf();
        test.login();
        test.initFileSystem();
        test.doSth();
    }
}
```

说明

1. Kerberos认证时需要配置Kerberos认证所需要的文件参数，主要包含keytab路径，Kerberos认证的用户名称，Kerberos认证所需要的客户端配置krb5.conf文件。
2. 方法login()为调用hadoop的接口执行Kerberos认证，生成TGT票据。
3. 方法doSth()调用hadoop的接口访问文件系统，此时底层RPC会自动携带TGT去Kerberos认证，生成ST票据。

14 高危操作一览表

禁用操作

表14-1中描述了在集群操作与维护阶段，观察进行日常操作时应注意的禁用操作。

表 14-1 禁用操作

类别	操作风险
严禁删除ZooKeeper相关数据目录	ClickHouse/HDFS/Yarn/HBase/Hive等很多组件都依赖于ZooKeeper，在ZooKeeper中保存元数据信息。删除ZooKeeper中相关数据目录将会影响相关组件的正常运行。
严禁JDBCServer主备节点频繁倒换	频繁主备倒换将导致业务中断。
严禁删除Phoenix系统表或系统表数据(SYSTEM.CATALOG、SYSTEM.STATS、SYSTEM.SEQUENCE、SYSTEM.FUNCTION)	删除系统表将导致无法正常进行业务操作。
严禁手动修改Hive元数据库的数据（ hivemeta数据库 ）	修改Hive元数据可能会导致Hive数据解析错误，Hive无法正常提供服务。
禁止对Hive的元数据表手动进行**insert**和**update**操作	修改Hive元数据可能会导致Hive数据解析错误，Hive无法正常提供服务。
严禁修改Hive私有文件目录hdfs:///tmp/hive-scratch的权限	修改该目录权限可能会导致Hive服务不可用。
严禁修改Kafka配置文件中broker.id	修改Kafka配置文件中broker.id将会导致该节点数据失效。
严禁修改节点主机名	主机名修改后会导致该主机上相关实例和上层组件无法正常提供服务，且无法修复。
禁止重装节点OS	该操作会导致MRS集群进入异常状态，影响MRS集群使用。

类别	操作风险
禁止使用私有镜像	该操作会导致MRS集群进入异常状态，影响MRS集群使用。

以下各表分别描述了各组件在操作与维护阶段，进行日常操作时应注意的高危操作。

集群高危操作

表 14-2 集群高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
随意修改omm用户下的文件目录或者文件权限	该操作会导致MRS集群服务不可用	★ ★ ★ ★ ★	请勿执行该操作	观察MRS集群服务是否可用
绑定弹性公网IP	该操作会将集群的manager所在的master节点暴露在公网，会增大来自互联网的网络攻击风险可能性	★ ★ ★ ★ ★	请确认绑定的弹性公网IP为可信任的公网访问IP	无
开放集群22端口安全组规则	该操作会增大用户利用22进行漏洞攻击的风险	★ ★ ★ ★ ★	针对开放的22端口进行设置安全组规则，只允许可信的IP可以访问该端口，入方向规则不推荐设置允许0.0.0.0可以访问。	无
删除集群或删除集群数据	该操作会导致数据丢失	★ ★ ★ ★ ★	删除前请务必再次确认该操作的必要性，同时要保证数据已完成备份	无
缩容集群	该操作会导致数据丢失	★ ★ ★ ★ ★	缩容前请务必再次确认该操作的必要性，同时要保证数据已完成备份	无

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
卸载磁盘或格式化数据盘	该操作会导致数据丢失	★ ★ ★ ★ ★	操作前请请务必再次确认该操作的必要性，同时要保证数据已完成备份	无

Manager 高危操作

表 14-3 Manager 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改OMS密码	该操作会重启OMS各进程，影响集群的管理维护	★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常
导入证书	该操作会重启OMS进程和整个集群，影响集群的管理维护和业务	★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
升级	该操作会重启Manager和整个集群，影响集群的管理维护和业务 分配集群管理权限的用户，需要严格管控，以防范可能的安全风险	★ ★ ★	修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
恢复OMS	该操作会重启Manager和整个集群，影响集群的管理维护和业务	★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改IP	该操作会重启Manager和整个集群，影响集群的管理维护和业务	★ ★ ★	修改时确保同一时间无其它管理维护操作，且修改的IP填写正确无误	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改日志级别	如果修改为DEBUG，会导致Manager运行速度明显降低	★ ★	修改前确认操作的必要性，并及时修改回默认设定	无
更换控制节点	该操作会导致部署在该节点上的服务中断，且当该节点同时为管理节点时，更换节点会导致重启OMS各进程，影响集群的管理维护	★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
更换管理节点	该操作会导致部署在该节点上的服务中断，会导致重启OMS各进程，影响集群的管理维护	★ ★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
重启下层服务时，如果勾选同时重启上层服务	该操作会导致上层服务业务中断，影响集群的管理维护和业务	★ ★ ★ ★	操作前确认操作的必要性，操作时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改OLDAP端口	修改该参数时，会重启LdapServer和Kerberos服务和其关联的所有服务，会影响业务运行	★ ★ ★ ★ ★	操作前确认操作的必要性，操作时确保同一时间无其它管理维护操作	无
用户删除supergroup组	删除supergroup组导致相关用户权限变小，影响业务访问	★ ★ ★ ★ ★	修改前确认需要添加的权限，确保用户绑定的supergroup权限删除前，相关权限已经添加，不会对业务造成影响	无

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
重启服务	重启过程中会中断服务，如果勾选同时重启上层服务会导致依赖该服务的上层服务中断	★ ★ ★	操作前确认重启的必要性	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改节点SSH默认端口	修改默认端口（22）将导致创建集群、添加服务/实例、添加主机、重装主机等功能无法正常使用，并且会导致集群健康检查结果中节点互信、omm/ommdba用户密码过期等检查项不准确	★ ★ ★	执行相关操作前将SSH端口改回默认值	无

ClickHouse 高危操作

表 14-4 ClickHouse 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除数据目录	该操作将会导致业务信息丢失	★ ★ ★	请勿手动删除数据目录	观察数据目录是否正常
缩容ClickHouseServer实例	该操作需要关注同分片中的ClickHouseServer实例节点需要同时退服缩容，否则会造成逻辑集群拓扑信息错乱；该操作执行前需检查逻辑集群内各节点的数据库和数据表信息，进行缩容预分析，保证缩容退服过程中数据迁移成功，避免数据丢失	★ ★ ★ ★ ★	进行缩容操作前，提前收集信息进行ClickHouse逻辑集群及实例节点状态判断	观察ClickHouse逻辑集群拓扑信息，各ClickHouseServer中数据库和数据表信息，以及数据量

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
扩容ClickHouseServer实例	该操作需要关注新扩容节点是否需要创建老节点上同名的数据库或数据表，否则会造成后续数据迁移、数据均衡以及缩容退服失败	★ ★ ★ ★ ★	进行扩容操作前，确认新扩容ClickHouseServer实例作用和目的，是否需要同步创建相关数据库和数据表	观察ClickHouse逻辑集群拓扑信息，各ClickHouseServer中数据库和数据表信息，以及数据量
退服ClickHouseServer实例	该操作需要关注同分片中的ClickHouseServer实例节点需要同时退服，否则会造成逻辑集群拓扑信息错乱；该操作执行前需检查逻辑集群内各节点的数据库和数据表信息，进行预分析，保证退服过程中数据迁移成功，避免数据丢失	★ ★ ★ ★ ★	进行退服操作前，提前收集信息进行Clickhouse逻辑集群及实例节点状态判断	观察ClickHouse逻辑集群拓扑信息，各ClickHouseServer中数据库和数据表信息，以及数据量
入服ClickHouseServer实例	该操作需要关注入服时必须选择原有分片中的所有节点入服，否则会造成逻辑集群拓扑信息错乱	★ ★ ★ ★ ★	进行入服操作前，对于待入服节点的分片归属信息需要确认	观察ClickHouse逻辑集群拓扑信息
修改数据目录下内容（创建文件、文件夹）	该操作将会导致该节点上的ClickHouse的实例故障	★ ★ ★	请勿手动在数据目录下创建或修改文件及文件夹	观察数据目录是否正常
单独启停基础组件	该操作将会影响服务的一些基础功能导致业务失败	★ ★ ★	请勿单独启停ZooKeeper/Kerberos/LDAP等基础组件，启停基础组件请勾选关联服务	观察服务状态是否正常
重启/停止服务	该操作将会导致业务中断	★ ★	确保在必要时重启/停止服务	观察服务是否运行正常

DBService 高危操作

表 14-5 DBService 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改DBService密码	修改密码需要重启服务，服务在重启过程中无法访问。	★ ★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作。	观察是否有未恢复的告警产生，观察集群的管理维护是否正常
恢复DBService数据	数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。 数据恢复后，依赖DBService的组件可能配置过期，需要重启配置过期的服务。	★ ★ ★ ★	恢复前确认操作的必要性，恢复时确保同一时间无其它管理维护操作。	观察是否有未恢复的告警产生，观察集群的管理维护是否正常
DBService主备倒换	倒换DBServer过程中，DBService无法提供服务。	★ ★	操作前确认该操作的必要性，操作时确保同一时间无其它管理维护操作。	无
修改DBService浮动IP配置	需要重启DBService服务使配置生效，服务在重启无法访问。 如果浮动IP已被使用过，将会导致配置失败，DBService启动失败。	★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效。	观察服务能否正常启动

Flink 高危操作

表 14-6 Flink 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改日志级别	如果修改为DEBUG，会影响任务运行性能	★ ★	修改前确认操作的必要性，并及时修改回默认设定	无
修改文件权限	该操作可能导致任务运行失败	★ ★ ★	修改前确认操作的必要性	观察相关业务操作是否正常

Flume 高危操作

表 14-7 Flume 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Flume实例的启动参数GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
修改HDFS的副本数目dfs.replication，将默认值由3改为1	导致： 1. 存储可靠性下降，磁盘故障时，会发生数据丢失 2. NameNode重启失败，HDFS服务不可用	★ ★ ★ ★	修改相关配置项时，请仔细查看参数说明。保证数据存储的副本数不低于2	观察默认的副本值是否不为1，HDFS服务是否可以正常提供服务

HBase 高危操作

表 14-8 HBase 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改加密的相关配置项 ● hbase.regionserver.wal.encryption ● hbase.crypto.keyprovider.parameters.uri ● hbase.crypto.keyprovider.parameters.encryptedtext	导致服务启动异常	★ ★ ★ ★	修改相关配置项时请严格按照提示描述，加密相关配置项是有关联的，确保修改后的值有效	观察服务能否正常启动
已使用加密的情况下关闭或者切换加密算法，关闭主要指修改hbase.regionserver.wal.encryption为false，切换主要指AES和SMS4的切换	导致服务启动失败，数据丢失	★ ★ ★ ★	加密HFile和WAL内容的时候，如果已经使用一种加密算法加密并且已经建表，请不要随意关闭或者切换加密算法。 未建加密表（ENCRYPTION=>AES/SMS4）的情况下可以切换，否则禁止操作	无

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改HBase实例的启动参数GC_OPTS、HBASE_HEAPSIZE	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效，且GC_OPTS与HBASE_HEAPSIZE参数值无冲突	观察服务能否正常启动
使用OfflineMetaRepair工具	导致服务启动异常	★ ★ ★ ★	必须在HBase下线的情况下才可以使用该命令，而且不能在数据迁移的场景中使用该命令	观察HBase服务是否可以正常启动。

HDFS 高危操作

表 14-9 HDFS 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改HDFS的NameNode的数据存储目录dfs.namenode.name.dir、DataNode的数据配置目录dfs.datanode.data.dir	导致服务启动异常	★ ★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
执行**hadoop distcp**命令时，使用**-delete**参数	Distcp拷贝时，源集群没有而目的集群存在的文件，会在目的集群删除	★ ★	在使用Distcp的时候，确保是否保留目的集群多余的文件，谨慎使用**-delete**参数	Distcp数据拷贝后，查看目的的数据是否按照参数配置保留或删除
修改HDFS实例的启动参数GC_OPTS、HADOOP_HEAPSIZE和GC_PROFILE	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效，且GC_OPTS与HADOOP_HEAPSIZE参数值无冲突	观察服务能否正常启动
修改HDFS的副本数目dfs.replication，将默认值由3改为1	导致： 1. 存储可靠性下降，磁盘故障时，会发生数据丢失 2. NameNode重启失败，HDFS服务不可用	★ ★ ★ ★	修改相关配置项时，请仔细查看参数说明。保证数据存储的副本数不低于2	观察默认的副本值是否不为1，HDFS服务是否可以正常提供服务
修改Hadoop中各模块的RPC通道的加密方式hadoop.rpc.protection	导致服务故障及业务异常	★ ★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察HDFS及其他依赖HDFS的服务能否正常启动，并提供服务

Hive 高危操作

表 14-10 Hive 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Hive实例的启动参数GC_OPTS	修改该参数可能会导致Hive实例无法启动	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
删除MetaStore所有实例	Hive元数据丢失，Hive无法提供服务	★ ★ ★	除非确定丢弃Hive所有表信息，否则不要执行该操作	观察服务能否正常启动
使用HDFS文件系统接口或者HBase接口删除或修改Hive表对应的文件	该操作会导致Hive业务数据丢失或被篡改	★ ★	除非确定丢弃这些数据，或者确保该修改操作符合业务需求，否则不要执行该操作	观察Hive数据是否完整
使用HDFS文件系统接口或者HBase接口修改Hive表对应的文件或目录访问权限	该操作可能会导致相关业务场景不可用	★ ★ ★	请勿执行该操作	观察相关业务操作是否正常
使用HDFS文件系统接口删除或修改文件hdfs:///apps/templeton/hive-3.1.0.tar.gz	该操作可能会导致WebHCat无法正常执行业务	★ ★	请勿执行该操作	观察相关业务操作是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
导出表数据覆盖写入本地目录，例如将**t1**表中数据导出，覆盖到“/opt/dir”路径下： **insert overwrite local directory '/opt/dir' select * from t1;**	该操作会删除目标目录，如果设置错误，会导致软件或者操作系统无法启动	★ ★ ★ ★ ★	确认需要写入的路径下不要包含任何文件；或者不要使用overwrite关键字	观察目标路径是否有文件丢失
将不同的数据库、表或分区文件指定至相同路径，例如默认仓库路径“/user/hive/warehouse”。	执行创建操作后数据可能会紊乱，如果删除其中一个数据库、表或分区，会导致其他对象数据丢失	★ ★ ★ ★ ★	请勿执行该操作	观察目标路径是否有文件丢失

Kafka 高危操作

表 14-11 Kafka 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除Topic	该操作将会删除已有的主题和数据	★ ★ ★	采用Kerberos认证，保证合法用户具有操作权限，并确保主题名称正确	观察主题是否正常处理

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除数据目录	该操作将会导致业务信息丢失	★ ★ ★	请勿手动删除数据目录	观察数据目录是否正常
修改数据目录下内容（创建文件、文件夹）	该操作将会导致该节点上的Broker实例故障	★ ★ ★	请勿手动在数据目录下创建或修改文件及文件夹	观察数据目录是否正常
修改磁盘自适应功能“disk.adapter.enable”参数	该操作会在磁盘使用空间达到阈值时调整Topic数据保存周期，超出保存周期的历史数据可能被清除	★ ★ ★	若个别Topic不能做保存周期调整，将该Topic配置在“disk.adapter.topic.blacklist”参数中	在KafkaTopic监控页面观察数据的存储周期
修改数据目录“log.dirs”配置	该配置不正确将会导致进程故障	★ ★ ★	确保所修改或者添加的数据目录为空目录，且权限正确	观察数据目录是否正常
减容Kafka集群	该操作将会导致部分Topic数据副本数量减少，可能会导致Topic无法访问	★ ★	请先做好数据副本转移工作，然后再进行减容操作	观察分区所在备份节点是否都存活，确保数据安全
单独启停基础组件	该操作将会影响服务的一些基础功能导致业务失败	★ ★ ★	请勿单独启停ZooKeeper/Kerberos/LDAP等基础组件，启停基础组件请勾选关联服务	观察服务状态是否正常
重启/停止服务	该操作将会导致业务中断	★ ★	确保在必要时重启/停止服务	观察服务是否运行正常
修改配置参数	该操作将需要重启服务使得配置生效	★ ★	确保在必要时修改配置	观察服务是否运行正常
删除/修改元数据	修改或者删除ZooKeeper上Kafka的元数据可能导致Topic或者Kafka服务不可用	★ ★ ★	请勿删除或者修改Kafka在ZooKeeper上保存的元数据信息	观察Topic或者Kafka服务是否可用

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改元数据备份文件	修改Kafka元数据备份文件，并被使用进行Kafka元数据恢复成功后，可能导致Topic或者Kafka服务不可用	★ ★ ★	请勿修改Kafka元数据备份文件	观察Topic或者Kafka服务是否可用

KrbServer 高危操作

表 14-12 KrbServer 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改KrbServer的参数KADMIN_PORT	修改该参数后，若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
修改KrbServer的参数kdc_ports	修改该参数后，若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
修改KrbServer的参数KPASSWD_PORT	修改该参数后，若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
修改Manager系统域名	若没有及时重启KrbServer服务和其关联的所有服务，会导致集群内部KrbClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启KrbServer服务和其关联的所有服务	无
配置跨集群互信	该操作会重启KrbServer服务和其关联的所有服务，影响集群的管理维护和业务	★ ★ ★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常

LdapServer 高危操作

表 14-13 LdapServer 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改LdapServer的参数LDAP_SERVER_PORT	修改该参数后，若没有及时重启LdapServer服务和其关联的所有服务，会导致集群内部LdapClient的配置参数异常，影响业务运行	★ ★ ★ ★ ★	修改该参数后，请重启LdapServer服务和其关联的所有服务	无
恢复LdapServer数据	该操作会重启Manager和整个集群，影响集群的管理维护和业务	★ ★ ★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
更换LdapServer所在节点	该操作会导致部署在该节点上的服务中断，且当该节点为管理节点时，更换节点会导致重启OMS各进程，影响集群的管理维护	★ ★ ★	更换前确认操作的必要性，更换时确保同一时间无其它管理维护操作	观察是否有未恢复的告警产生，观察集群的管理维护是否正常，业务是否正常
修改LdapServer密码	修改密码需要重启LdapServer和Kerberos服务，影响集群的管理维护和业务	★ ★ ★ ★	修改前确认操作的必要性，修改时确保同一时间无其它管理维护操作	无
节点重启导致LdapServer数据损坏	如果未停止LdapServer服务，直接重启LdapServer所在节点，可能导致LdapServer数据损坏	★ ★ ★ ★ ★	使用LdapServer备份数据进行恢复	无

Loader 高危操作

表 14-14 Loader 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Loader实例的浮动IP地址loader.float.ip	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察Loader UI是否可以正常连接
修改Loader实例的启动参数LOADER_GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
往HBase导入数据时，选择清空表数据	目标表的原数据被清空	★ ★	选择时，确保目标表的数据可以清空	选择前，需确认目标表数据是否可以清空

Spark2x 高危操作

说明

MRS 3.x之前版本，服务名称为Spark。

表 14-15 Spark2x 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
配置项的修改（spark.yarn.queue）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
配置项的修改（spark.driver.extraJavaOptions）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
配置项的修改（spark.yarn.cluster.driver.extraJavaOptions）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
配置项的修改（spark.eventLog.dir）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
配置项的修改（SPARK_DAEMON_JAVA_OPTS）	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
删除所有JobHistory2x实例	导致历史应用的event log丢失	★ ★	至少保留一个JobHistory2x实例	观察JobHistory2x中是否可以查看历史应用信息
删除或修改HDFS上的/user/spark2x/jars/8.1.0.1/spark-archive-2x.zip	导致JDBCServer2x启动异常及业务功能异常	★ ★ ★	删除/user/spark2x/jars/8.1.0.1/spark-archive-2x.zip，等待10-15分钟，zip包自动恢复	观察服务能否正常启动

Storm 高危操作

表 14-16 Storm 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改插件相关的配置项： ● storm.scheduler ● nimbus.authorizer ● storm.thrift.transport ● nimbus.blobstore.class ● nimbus.topology.validator ● storm.principal.tolocal	导致服务启动异常	★ ★ ★ ★	修改相关配置项时请严格按照提示描述，确保修改后的类名是存在并有效的	观察服务能否正常启动

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改Storm实例的启动参数 GC_OPTS NIMBUS_GC_OPTS SUPERVISOR_GC_OPTS UI_GC_OPTS LOGVIEWER_GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
修改用户资源池配置参数 resource.aware.scheduler.user.pools	导致业务提交后无法正常运行	★ ★ ★	修改相关配置项时请严格按照提示描述，确保给每个用户分配的资源合理有效	观察服务能否正常启动并且业务能否正常运行
修改数据目录	该操作不当会导致服务异常，无法提供服务	★ ★ ★ ★	请勿手动操作数据目录	观察数据目录是否正常
重启服务/实例	该操作会导致服务有短暂中断，如果有业务运行也会引起业务短暂中断	★ ★ ★	确保在必要时重启服务	观察服务是否运行正常，业务是否恢复
同步配置（重启服务）	该操作会引起服务重启，导致服务短暂中断，若引起Supervisor重启会导致所运行业务短暂中断	★ ★ ★	确保在必要时修改配置	观察服务是否运行正常，业务是否恢复
停止服务/实例	该操作会导致服务停止，业务中断	★ ★ ★	确保在必要时停止服务	观察服务是否正常停止
删除/修改元数据	删除Nimbus元数据会导致服务异常，并且已运行业务丢失	★ ★ ★ ★ ★	请勿手动删除Nimbus元数据文件	观察Nimbus元数据文件是否正常

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
修改文件权限	修改元数据目录和日志目录权限不当会引起服务异常	★ ★ ★ ★	请勿手动修改文件权限	观察数据目录和日志目录权限是否正常
删除拓扑	该操作会删除正在运行中的拓扑	★ ★ ★ ★	确保在必要时删除拓扑	观察拓扑是否删除成功

Yarn 高危操作

表 14-17 Yarn 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除或者修改数据目录 yarn.nodemanager.local-dirs 和 yarn.nodemanager.log-dirs	该操作将会导致业务信息丢失	★ ★ ★	请勿手动删除数据目录	观察数据目录是否正常

ZooKeeper 高危操作

表 14-18 ZooKeeper 高危操作

操作名称	操作风险	风险等级	规避措施	重大操作观察项目
删除或者修改ZooKeeper的数据目录	该操作将会导致业务信息丢失	★ ★ ★	修改ZooKeeper目录时候，严格按照扩容指导操作	观察服务能否正常启动，关联组件能否正常启动
修改ZooKeeper实例的启动参数GC_OPTS	导致服务启动异常	★ ★	修改相关配置项时请严格按照提示描述，确保修改后的值有效	观察服务能否正常启动
设置ZooKeeper中znode的ACL信息	修改ZooKeeper中znode的权限，可能会导致其他用户无权限访问该znode，导致系统功能异常	★ ★ ★ ★	修改相关配置项时请严格按照“ZooKeeper权限设置指南”章节操作，确保修改ACL信息，不会影响其他组件正常使用ZooKeeper	观察项目观察其他依赖ZooKeeper的组件能否正常启动，并提供服务

15 故障排除

15.1 Web 页面访问类

15.1.1 无法访问 MRS 集群管理页面（MRS Manager 界面）

问题现象

集群创建完成后，无法访问集群管理页面，即无法访问MRS Manager界面。

原因分析

- 需要对MRS节点绑定弹性IP才可访问
- 需要添加安全组规则，放开9022端口

处理步骤

步骤1 登录MRS的Console页面，在现有集群列表中找到需要访问的集群，单击集群名称。

步骤2 在节点信息中单击需要访问的节点名称，选择“弹性公网IP” > “绑定弹性公网IP”。

步骤3 在“绑定弹性公网”IP页面，“选择网卡”下拉框中选择需要绑定的网卡，“选择弹性公网IP”中选择需要绑定的弹性公网IP，单击“确定”。

步骤4 弹性IP绑定成功后，需要将安全组规则中9022端口放开。

选择“安全组”页签，单击“更改安全组”。

可以选择添加已有的安全组，或者单击“新建安全组”，进入安全组管理界面进行创建，添加用户访问公网IP地址9022端口的安全组规则。

步骤5 添加成功后，可通过*https://弹性ip:9022/mrsmanager/*访问MRS。如果还未能访问，请联系技术支持。

----结束

15.1.2 升级 Python 后，无法登录 MRS Manager 页面

用户问题

升级Python后，登录不进去MRS Manager页面。

问题现象

自行升级Python后，使用admin账号且密码正确的情况下登录不进去MRS Manager页面。

原因分析

用户升级Python版本到Python3.x的过程中，修改了openssl的文件目录权限，导致LdapServer服务无法正常启动，从而引起登录认证失败。

处理步骤

步骤1 以root用户登录集群的Master节点。

步骤2 执行**chmod 755 /usr/bin/openssl**命令，修改**/usr/bin/openssl**的文件目录权限为755。

步骤3 执行**su omm**命令，切换到omm用户。

步骤4 执行**openssl**命令，查看是否能够进入openssl模式。

如果能够成功进入，则表示权限修改成功，如果不能进入，则表示权限未修改成功。

如果权限未修改成功，请检查执行的命令是否正确，或者联系运维人员。

步骤5 权限修改成功后会重启LdapServer服务，请等待LdapServer服务重启成功后，重新登录MRS Manager。

----结束

建议与总结

自行安装的软件建议和系统的分开，系统软件升级可能造成兼容性问题。

15.1.3 用户修改域名后无法登录 MRS Manager 页面

问题现象

用户修改域名后，通过console页面无法登录MRS Manager页面，或者登录MRS Manager页面异常。

问题原因

用户修改域名后，没有刷新executor用户的keytab文件，导致executor进程认证失败后不断循环认证，导致了acs进程内存溢出。

解决方案

步骤1 重启acs进程。

1. 使用root用户登录主管理节点（即MRS集群详情页面“节点管理”页签下实心五角星所在的Master节点）。
2. 执行如下命令重启进程：

   su - omm

   ps -ef|grep =acs（查找acs进程PID）

   kill -9 *PID*（PID替换为实际的ID，结束acs进程）
3. 等待几分钟后执行命令**ps -ef|grep =acs**查询进程是否已经自动启动。

步骤2 替换executor用户的keytab文件。

1. 登录MRS Manager页面，选择“系统 > 用户”，在executor用户所在的“操作”列，单击“下载认证凭据”，解压后获取keytab文件。
2. 使用root用户登录主管理节点，将获取的keytab替换“/opt/executor/webapps/executor/WEB-INF/classes/user.keytab”文件。

步骤3 替换knox用户的keytab和conf文件。

1. 登录MRS Manager页面，选择“系统 > 用户”，在knox用户所在的“操作”列，单击“下载认证凭据”，解压后获取keytab和conf文件。
2. 使用root用户登录主管理节点，将获取的keytab替换“/opt/knox/conf/user.keytab”文件。
3. 修改/opt/knox/conf/krb5JAASLogin.conf中的principal的值，把域名修改为更改后的域名。
4. 将获取的krb5.conf 替换“/opt/knox/conf/krb5.conf”文件。

步骤4 备份原有客户端目录

mv *{客户端目录}* /opt/client_init

步骤5 重新安装客户端。

步骤6 使用root用户登录主备管理节点，执行如下命令，重启knox进程。

su - omm

ps -ef | grep gateway | grep -v grep（查找knox进程PID）

kill -9 *PID*（PID替换为实际的ID，结束knox进程）

/opt/knox/bin/restart-knox.sh（启动knox进程）

步骤7 使用root用户登录主备管理节点，执行如下命令，重启executor进程。

su - omm

netstat -anp |grep 8181 |grep LISTEN（查找executor进程PID）

kill -9 *PID*（PID替换为实际的ID，结束executor进程）

/opt/executor/bin/startup.sh（启动executor进程）

----结束

15.1.4 登录 Manager，页面空白不显示

用户问题

登录到FusionInsight Manager界面后，页面空白不显示。

问题现象

登录到FusionInsight Manager界面后，页面空白不显示。

原因分析

Manager无法登录，需要清除浏览器缓存。

处理步骤

步骤1 切换至浏览器窗口（以Chrome为例），通过键盘按下“Ctrl+Shift+Delete”弹出“清除浏览数据”对话框。

步骤2 勾选待清除的浏览记录，单击“清除数据”，完成浏览器缓存清理。

----结束

15.1.5 用户名过长时下载认证凭据失败

用户问题

MRS 3.0.2~MRS 3.1.0版本集群，当用户名超过20位时（添加用户时最长限制为32位），下载Keytab文件会下载失败，状态代码：400 Bad Request。

问题现象

MRS 3.0.2~MRS 3.1.0版本集群，当用户名超过20位时（添加用户时最长限制为32位），下载Keytab文件会下载失败，状态代码：400 Bad Request。

原因分析

需要在主Master节点的“/opt/Bigdata/om-server_*/apache-tomcat-*/webapps/web/WEB-INF/validate”路径下，修改validate-common-config.xml、validate-rule-session.xml、 validate-rule-user.xml三个配置文件。

处理步骤

步骤1 以**omm**用户登录主Master节点的“/opt/Bigdata/om-server_*/apache-tomcat-*/webapps/web/WEB-INF/validate”路径。

cd /opt/Bigdata/om-server_*/apache-tomcat-*/webapps/web/WEB-INF/validate

步骤2 修改validate-common-config.xml文件。

vi validate-common-config.xml

将用户名的“maxLength”参数的值从“32”修改为“64”：

```
<!-- 用户名 -->
<validators alias="USER_NAME">
      <validator name="RANGE_LENGTH_VALIDATOR" minLength="3"
         maxLength="64" />
      <validator name="REGEXP_VALIDATOR" rule="^[_a-zA-Z0-9\- ]+$"
</validators>
```

步骤3 修改validate-rule-session.xml文件。

vi validate-rule-session.xml

将“下载当前用户凭据”的参数“rule”的值从“20”改为“64”：

```
 <!-- 下载当前用户凭据 -->
 <param_validator url="/api/v2/session/user/keytab/download" method="get"
errorHandler="com.xxx.bigdata.om.web.api.validate.SpecialValidatorErrorHandler" dataPattern="form">
            <!-- 参数名：文件名 -->
            <!--校验规则：userName_13位数字_keytab.tar；区分大小写-->
            <parameter name="file_name" required="true" errorKey="13-4000005"
errorMessage="RESID_OM_API_SESSION_0013">
                  <validator name="REGEXP_VALIDATOR" rule="[\-\w ]{3,64}_\d{13}_keytab\.tar"
caseSensitive="true" />
 </parameter>
```

步骤4 修改validate-rule-user.xml文件。

vi validate-rule-user.xml

将“下载当前用户凭据”的参数“rule”的值从“20”改为“64”：

```
<!--下载用户凭据  -->
   <param_validator url="/api/v2/permission/users/keytab/download" method="get"
errorHandler="com.xxx.bigdata.om.web.api.validate.SpecialValidatorErrorHandler" dataPattern="form">
          <!--必输；userName_13位数字_keytab.tar；区分大小写-->
          <parameter name="file_name" required="true" errorKey="12-4000005"
errorMessage="RESID_OM_API_AUTHORITY_0005">
                <validator name="REGEXP_VALIDATOR" rule="[\-\w ]{3,64}_\d{13}_keytab\.tar"
caseSensitive="true" />
          </parameter>
    </param_validator>
```

步骤5 重启Tomcat，并等待启动成功。

1. 以**omm**用户执行以下命令，查询出Tomcat进程的PID号。
   ps -ef|grep apache-tomcat
2. 使用**kill -9** *PID*命令强制停止查询出来的Tomcat进程，例如**:**
   kill -9 1203
3. 执行以下命令进行重启。
   sh ${BIGDATA_HOME}/om-server/tomcat/bin/startup.sh

步骤6 重新下载认证凭据。

----结束

15.2 集群管理类

15.2.1 缩容 Task 节点失败

用户问题

客户在MRS 2.x集群详情界面执行调整集群，将Task节点调整成0个，最终缩容失败。

问题现象

客户在MRS集群详情页面调整集群Task节点，最终缩容失败，提示“This operation is not allowed because the number of instances of NodeManager will be less than the minimum configuration after scale-in, which may cause data loss.”

原因分析

客户将Core节点的NodeManager服务停止了，导致在检查Task节点退服过程中发现Task如果全部退订，则将没有NodeManager，则Yarn服务就不可用，而MRS判断剩余的NodeManger必须大于等于1才能退服Task节点。

处理步骤

步骤1 选择“服务管理 > Yarn > 实例”。

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Yarn > 实例”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Yarn > 实例”。
- MRS 3.x及后续版本，登录FusionInsight Manager。然后选择“集群 > *待操作的集群名*称 > 服务 > Yarn > 实例”。

步骤2 勾选Core节点的NodeManager实例，选择“更多 > 启动实例”。

步骤3 在集群列表页面缩容Task节点。

1. 单击集群名称进入集群详情页面，选择“节点管理”。
2. 在Task节点组所在行的“操作”列单击“缩容”。
3. 单击“确定”并在弹出框选择“是”。

步骤4 等缩容成功后，若不想用Core节点的NodeManager再将其停止。

----结束

建议与总结

Core节点的NodeManager通常不会将其停止，客户不要随意变更集群部署结构。

15.2.2 MRS 集群添加新磁盘

用户问题

MRS HBase服务不可用。

问题现象

用户主机的磁盘占用率过高导致服务故障。

原因分析

Core节点的磁盘容量不足导致无法提供正常服务。

处理步骤

步骤1 购买云硬盘。

步骤2 挂载云硬盘。

- 若挂载云硬盘完成，请执行**步骤6**。
- 若在云硬盘控制台执行 "挂载" 操作时无法选定云服务器，请执行**步骤3**。

步骤3 登录弹性云服务器控制台，单击待扩容（挂载新磁盘）的弹性云服务器名称。

步骤4 在“云硬盘”页签，单击“挂载磁盘”。

步骤5 选择待挂载的新磁盘并单击“确定”完成磁盘挂载。

步骤6 初始化Linux数据盘。

说明

- 挂载点目录根据节点DataNode已有的实例编号递增，例如：使用df –h命令查到当前已有的编号为/srv/BigData/hadoop/data1，则新增挂载点为/srv/BigData/hadoop/data2。初始化Linux数据盘新建挂载点时，将新建挂载点命名为/srv/BigData/hadoop/data2，并将新建分区挂载到该挂载点下。例如

```
mkdir /srv/BigData/hadoop/data2
mount /dev/xvdb1 /srv/BigData/hadoop/data2
```

/srv/BigData/hadoop/data2路径说明：本章节后续提到/srv/BigData/hadoop/data2路径均请按照以下场景自行修改。

- 3.X版本目录为：/srv/BigData/data2
- 3.X之前版本目录为：/srv/BigData/hadoop/data2

步骤7 执行以下命令为新磁盘增加omm用户权限。

chown omm:wheel *新增挂载点*

例如：**chown omm:wheel /srv/BigData/hadoop/data2**

步骤8 执行chmod 701命令为新增的挂载点目录添加执行权限。

chmod 701 *新增挂载点*

例如：**chmod 701 /srv/BigData/hadoop/data2**

说明

chmod 701命令中701仅为示例，请以已有数据盘data1的数值为准。

步骤9 登录Manager，扩容DataNode实例和NodeManager实例的数据磁盘。

步骤10 修改当前节点DataNode实例配置。

MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > HDFS > 实例 > 扩容的DataNode节点 > 实例配置”，“参数类别”选择 "全部配置"。

FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例 > 扩容的DataNode节点 > 实例配置”，选择 "全部配置"。

- 方式一：手动修改当前节点DataNode实例配置。
  - 在 "搜索" 中输入"dfs.datanode.fsdataset.volume.choosing.policy"，将参数值改为

"org.apache.hadoop.hdfs.server.datanode.fsdataset.AvailableSpaceVolumeChoosingPolicy"。

- 在 "搜索" 中输入 "dfs.datanode.data.dir"，将参数值改为"/srv/BigData/hadoop/data1/dn,/srv/BigData/hadoop/data2/dn"

若此两个参数有修改，则单击 "保存配置"，并勾选 "重启角色实例" ，重启DataNode实例。

- 方式二：自动同步当前节点DataNode实例配置。
  a. 单击“同步配置”为HDFS服务启用新的配置参数。
  b. 完成同步配置后，请重启实例以使配置生效。

说明

- 如果确认当前未使用HDFS，并且希望较快完成重启，可以选择直接"重启角色实例" 。
- 如果有任务在使用HDFS，为了防止数据异常或者任务失败，必须选择滚动重启。

步骤11 修改当前节点Yarn NodeManager的实例配置。

MRS Manager界面操作入口：登录MRS Manager，依次选择 "服务管理 > Yarn > 实例 > 扩容节点的NodeManager > 实例配置"，"参数类别" 选择 "全部配置"。

FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Yarn > 实例”单击扩容节点的NodeManager，选择“实例配置 > 全部配置”。

- 方式一：手动修改当前节点Yarn NodeManager的实例配置。
  - 在“搜索”中输入 "yarn.nodemanager.local-dirs"，将参数值修改为："/srv/BigData/hadoop/data1/nm/localdir,/srv/BigData/hadoop/data2/nm/localdir" 。
  - 在“搜索”中输入 "yarn.nodemanager.log-dirs"，将参数值修改为："/srv/BigData/hadoop/data1/nm/containerlogs,/srv/BigData/hadoop/data2/nm/containerlogs"。

    若此两个参数有修改，则保存配置，并勾选 "重启角色实例" ，重启NodeManager实例。
- 方式二：自动同步当前节点Yarn NodeManager的实例配置。
  a. 单击“同步配置”为Yarn服务启用新的配置参数。
  b. 完成同步配置后，请重启实例以使配置生效。

说明

- 如果确认当前未使用Yarn，并且希望较快完成重启，可以选择直接"重启角色实例" 。
- 如果有任务在使用Yarn，为了防止数据异常或者任务失败，必须选择滚动重启。

步骤12 查看扩容是否成功。

MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > HDFS > 实例 > 扩容的DataNode节点 "。

FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > HDFS > 实例”，单击扩容的DataNode节点。

在图表区域，查看实时监控项 "DataNode存储" 中 配置的总磁盘容量是否提升，若图表区域没有监控项 "DataNode存储" ，请单击“定制”增加该监控项。

- 若配置的总磁盘容量已提升，则扩容完成。
- 若配置的总磁盘容量未提升，请联系技术支持处理。

步骤13 （可选）扩容Kafka实例的数据盘。

修改当前节点Kafka实例配置。

1. 进入Kafka扩容的Broker节点参数配置界面。

   MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka > 实例 > 扩容的Broker节点 > 实例配置"，"参数类别" 选择 "全部配置"。

   FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例> 扩容的Broker节点 > 实例配置”， 选择 “全部配置”。
2. 在 "搜索" 中输入"log.dirs"，加入新增磁盘信息，中间用英文 “,” 分割。

   例如原始只有一块Kafka数据盘，新增一块，则将"/srv/BigData/kafka/data1/kafka-logs" 改为 "/srv/BigData/kafka/data1/kafka-logs,/srv/BigData/kafka/data2/kafka-logs"。
3. 保存配置，并勾选 "重启角色实例" 后根据提示重启实例。
4. 查看扩容是否成功。

   MRS Manager界面操作入口：登录MRS Manager，依次选择 "服务管理 > Kafka > 实例 > 扩容的Broker节点 "。

   FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 实例> 扩容的Broker节点 ”。

   查看实时监控项 "Broker磁盘容量大小" 中配置的总磁盘容量是否提升。

----结束

须知

集群的节点扩容磁盘之后，若再扩容集群节点时需要在新扩容的节点上参考该页面处理步骤执行添加磁盘的操作，否则会有数据丢失的风险 。

建议与总结

- 当磁盘的使用率超过85%时，建议用户进行磁盘扩容，并将新购买的磁盘挂载到弹性云服务器上与集群进行关联。
- 具体挂载步骤、配置参数请根据实际情况进行。

15.2.3 MRS 集群更换磁盘（适用于 2.x 及之前）

用户问题

磁盘无法访问。

问题现象

客户创建本地盘系列MRS集群，其中1个Core节点的磁盘存在硬件损坏，导致读取文件失败。

原因分析

磁盘硬件故障。

处理步骤

> **说明**
>
> 该指导适用于MRS 3.x之前版本分析集群，如需为流式集群或混合集群更换磁盘，请联系技术支持处理。

步骤1 登录MRS Manager。

步骤2 选择“主机管理”并单击需要退服主机的“主机名称”，在“角色”列表中单击RegionServer，选择“更多 > 退服”。

步骤3 选择“主机管理”并单击需要退服主机的“主机名称”，在“角色”列表中单击DataNode，选择“更多 > 退服”。

步骤4 选择“主机管理”并单击需要退服主机的“主机名称”，在“角色”列表中单击NodeManager，选择“更多 > 退服”。

> **说明**
>
> 该主机下若还有其他实例，请参考该步骤方式进行退服。

步骤5 执行**vim /etc/fstab**命令编辑注释旧磁盘的挂载点。

图 15-1 注释旧磁盘的挂载点

```
[root@node-ana-coregexX0001 ~]# vim /etc/fstab

devpts   /dev/pts          devpts   mode=0620,gid=5 0 0
proc     /proc             proc     defaults        0 0
sysfs    /sys              sysfs    noauto          0 0
debugfs  /sys/kernel/debug debugfs  noauto          0 0
tmpfs    /run              tmpfs    noauto          0 0
/dev/disk/by-label/ROOT / ext4 defaults,noatime 1 1
#UUID=0f871b41-61e0-4f7f-af54-a03a1bfb3753 /srv/BigData/hadoop/data1 ext4 defaults,noatime,nodiratime 1 0
~
~
```

步骤6 迁移旧磁盘上（例如：/srv/BigData/hadoop/data1/）的用户自有数据。

步骤7 登录MRS管理控制台。

步骤8 在集群详情页面，选择“节点管理”。

步骤9 单击待更换磁盘的“节点名称”进入弹性云服务器管理控制台，单击“关机”。

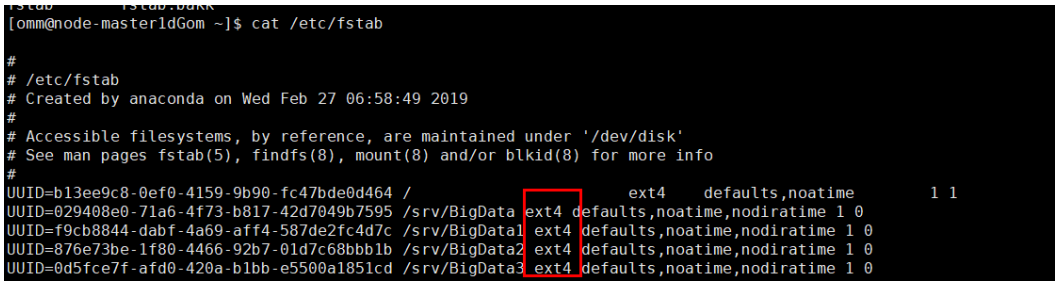
步骤10 联系支持人员在后台更换磁盘。

步骤11 在弹性云服务器管理控制台，单击“开机”，将已更换磁盘的节点开机。

步骤12 执行**fdisk -l**命令，查看新增磁盘。

步骤13 使用**cat /etc/fstab**获取盘符。

图 15-2 获取盘符



步骤14 使用对应的盘符对新磁盘进行格式化。

例如：**mkfs.ext4 /dev/sdh**

步骤15 执行如下命令挂载新磁盘。

mount *新磁盘 挂载点*

例如：**mount /dev/sdh /srv/BigData/hadoop/data1**

步骤16 执行如下命令为新磁盘增加omm用户权限。

chown omm:wheel *挂载点*

例如：**chown -R omm:wheel /srv/BigData/hadoop/data1**

步骤17 在fstab文件中新增新磁盘UUID信息。

1. 使用**blkid**命令查看新磁盘的UUID。

2. 打开“/etc/fstab”文件，新增如下信息：

   UUID=*新盘UUID* /srv/BigData/hadoop/data1 ext4 defaults,noatime,nodiratime 1 0

步骤18 （可选）执行如下命令新建日志目录。

mkdir -p /srv/BigData/Bigdata

chown omm:ficommon /srv/BigData/Bigdata

chmod 770 /srv/BigData/Bigdata

说明

执行如下命令确认Bigdata日志软链接目录是否已存在，若存在则忽略该步骤。

ll /var/log

步骤19 登录MRS Manager。

步骤20 选择“主机管理”并单击需要入服主机的“主机名称”，在“角色”列表中单击RegionServer，选择“更多 > 入服”。

步骤21 选择“主机管理”并单击需要入服主机的“主机名称”，在“角色”列表中单击DataNode，选择“更多 > 入服”。

步骤22 选择“主机管理”并单击需要入服主机的“主机名称”，在“角色”列表中单击NodeManager，选择“更多 > 入服”。

说明

该主机下若还有其他实例，请参考该步骤方式进行入服。

步骤23 选择“服务管理 > HDFS”，在“服务状态”页签的“HDFS概述”模块查看“丢失块数”是否为“0”。

- “丢失块数”是为“0”，则操作完成。
- “丢失块数”不为“0”，请联系支持人员进行处理。

----结束

15.2.4 MRS 集群更换磁盘（适用于 3.x）

用户问题

磁盘无法访问。

问题现象

客户创建本地盘系列MRS集群，其中1个Core节点的磁盘存在硬件损坏，导致读取文件失败。

原因分析

磁盘硬件故障。

处理步骤

说明

该指导适用于本地盘系列（d/i/ir/ki系列）MRS集群，针对Core、Task类型节点的磁盘存在硬件故障。

Kafka组件不支持更换磁盘，如果存储Kafka数据的节点故障，请联系技术支持处理。

步骤1 登录。

步骤2 选择“主机”并单击故障主机的“主机名称”，在“实例”列表中单击DataNode，选择“更多 > 退服”。

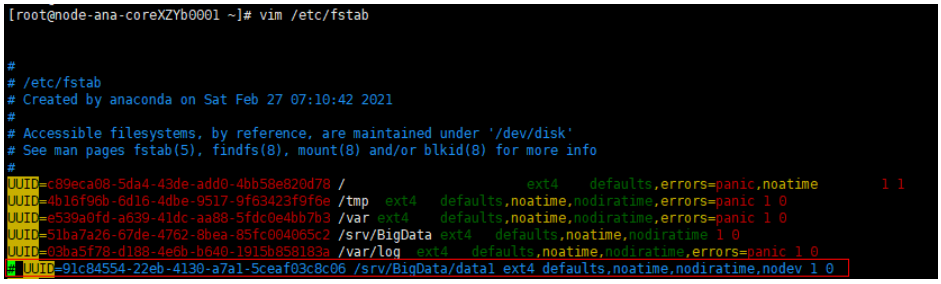
说明

- 该主机下若存在DataNode、NodeManager、RegionServer和ClickHouseServer实例，请参考该步骤进行退服操作；
- MRS 3.1.2版本之后支持退服ClickHouseServer角色实例。

步骤3 选择“主机”并勾选故障主机“主机名称”前的复选框，选择“更多 > 停止所有实例”。

步骤4 执行**vim /etc/fstab**命令编辑注释旧磁盘的挂载点。

图 15-3 注释旧磁盘的挂载点



步骤5 如果旧磁盘仍可访问，迁移旧磁盘上（例如：/srv/BigData/data1/）的用户自有数据。

cp -r *旧磁盘挂载点 临时数据保存目录*

例如：**cp -r /srv/BigData/data1 /tmp/**

步骤6 登录MRS管理控制台。

步骤7 在集群详情页面，选择“节点管理”。

步骤8 单击待更换磁盘的“节点名称”进入弹性云服务器管理控制台，单击“关机”。

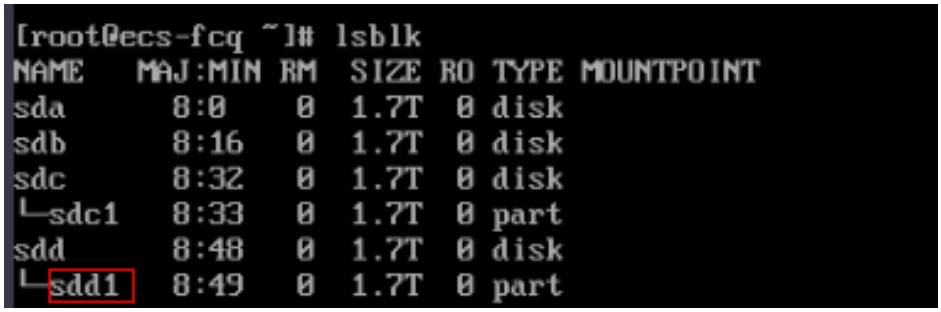
步骤9 联系支持人员在后台更换磁盘。

步骤10 在弹性云服务器管理控制台，单击“开机”，将已更换磁盘的节点开机。

步骤11 初始化Linux数据盘。

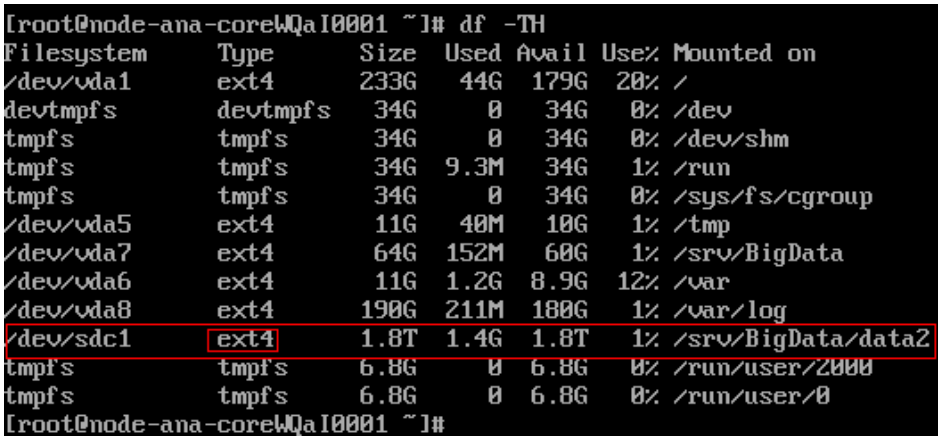
步骤12 执行**lsblk**命令，查看新增磁盘分区信息。

图 15-4 查看新增磁盘（分区）



步骤13 使用**df -TH**获取文件系统类型。

图 15-5 获取文件系统类型



步骤14 使用对应的文件系统类型对新磁盘（分区）进行格式化。

例如：**mkfs.ext4 /dev/sdd1**

步骤15 执行如下命令挂载新磁盘。

mount *新磁盘 挂载点*

例如：**mount /dev/sdd1 /srv/BigData/data1**

说明

如果挂载不上，请执行如下命令重载配置后重新挂载。

systemctl daemon-reload

步骤16 执行如下命令为新磁盘增加omm用户权限。

chown omm:wheel *挂载点*

例如：**chown -R omm:wheel /srv/BigData/data1**

步骤17 将旧磁盘上（例如：/srv/BigData/data1/）的用户自有数据迁移到新磁盘上。

cp -r*临时数据保存目录 新磁盘挂载点*

例如：**cp -r /tmp/data1/* /srv/BigData/data1/**

步骤18 在fstab文件中新增新磁盘UUID信息。

1. 使用**blkid**命令查看新磁盘的UUID。

2. 打开“/etc/fstab”文件，新增如下信息：
   UUID=*新盘UUID* /srv/BigData/data1 ext4 defaults,noatime,nodiratime,nodev 1 0

步骤19 登录。

步骤20 选择“主机”并单击需要入服主机的“主机名称”，在“实例”列表中单击DataNode，选择“更多 > 入服”。

说明

- 该主机下若存在DataNode、NodeManager、RegionServer和ClickHouseServer实例，请参考该步骤进行入服操作；
- MRS 3.1.2版本之后支持入服ClickHouseServer角色实例。

步骤21 选择“主机”，并勾选故障主机“主机名称”前的复选框，选择“更多 > 启动所有实例”。

步骤22 选择“集群 > HDFS”，在“概览”页签的“基本信息”模块查看“丢失块数”是否为“0”。

- “丢失块数”是为“0”，则操作完成。
- “丢失块数”不为“0”，请联系支持人员进行处理。

----结束

15.2.5 MRS 备份失败

用户问题

MRS备份总是失败。

问题现象

MRS备份总是失败。

原因分析

备份目录软链接到系统盘，系统盘满了之后备份便会失败。

处理步骤

步骤1 检查备份目录是否软链接到系统盘。

1. 以root用户登录集群主备Master节点。
2. 执行**df -h**命令查看磁盘情况，检查系统盘的存储情况。
3. 执行 **ll /srv/BigData/LocalBackup**命令， 查看备份目录是否软连接到**/opt/Bigdata/LocalBackup**。

   检查备份文件是否软链接到系统盘且系统盘空间是否足够。如果软链接到系统盘且系统盘空间不足，请执行**步骤2**。如果否，说明不是由于系统盘空间不足导致，请联系技术服务。

步骤2 将历史备份数据移到数据盘的新目录中。

1. 以root用户登录Master节点。
2. 执行**su - omm**命令，切换到omm用户。
3. 执行**rm -rf /srv/BigData/LocalBackup**命令，删除备份目录软连接。
4. 执行**mkdir -p /srv/BigData/LocalBackup**命令，创建备份目录。
5. 执行**mv /opt/Bigdata/LocalBackup/* /srv/BigData/LocalBackup/**命令，将历史备份数据移到新目录。

----结束

15.2.6 Core 节点出现 df 显示的容量和 du 显示的容量不一致

用户问题

Core节点出现df显示的容量和du显示的容量不一致

问题现象

Core节点出现df显示的容量和du显示的容量不一致：

分别使用命令**df -h** 和命令**du -sh /srv/BigData/hadoop/data1/**查询得到的/srv/BigData/hadoop/data1/目录磁盘占用量相差较大（大于10G）。

原因分析

使用命令**lsof |grep deleted**可以查询到此目录下有大量log文件处于deleted状态。

出现此问题的一种情况是长时间运行某些spark任务，任务中的一些container一直运行，并且持续产生日志；spark的executor在打印日志的时候使用了log4j的日志滚动功能，将日志输出到stdout文件下；而container同时也会监控这个文件，导致此文件被两个进程同时监控。当其中一个进程按照配置滚动的时候，删除了最早的日志文件，但是另一个进程依旧占用此文件句柄。从而产生了deleted状态的文件。

处理步骤

将spark的executor日志输出目录修改成其他名称

1. 打开日志配置文件，默认在<客户端地址>/Spark/spark/conf/log4j-executor.properties。
2. 将日志输出文件改名，例如：

   log4j.appender.sparklog.File = ${spark.yarn.app.container.log.dir}/stdout改为：log4j.appender.sparklog.File = ${spark.yarn.app.container.log.dir}/stdout.log
3. 保存退出
4. 重新提交任务。

15.2.7 如何解除关联子网

操作场景

您可根据自身网络需求，解除网络ACL与子网关联。

操作步骤

步骤1 登录管理控制台。

步骤2 在系统首页，单击“网络 > 虚拟私有云”。

步骤3 在左侧导航栏单击“网络ACL”。

步骤4 在右侧在“网络ACL”列表区域，选择网络ACL的名称列，单击您需要修改的“网络ACL名称”进入网络ACL详情页面。

步骤5 在详情页面，单击“关联子网”页签。

步骤6 在“关联子网”页签详情区域，选择对应子网的“操作”列，单击“取消关联”。

步骤7 单击“确认”。

----结束

15.2.8 修改 hostname，导致 MRS 状态异常

用户问题

修改hostname后，MRS状态异常怎么处理?

问题现象

修改hostname，导致MRS状态异常。

原因分析

修改hostname导致兼容性问题和故障。

处理步骤

步骤1 以root用户登录集群的任意节点。

步骤2 在集群节点中执行**cat /etc/hosts**命令，查看各个节点的hostname值，根据此值来配置newhostname变量值 。

步骤3 在hostname被修改的节点上执行**sudo hostnamectl set-hostname ${newhostname}**命令，恢复正确的hostname。

说明

${newhostname}：表示新的hostname取值。

步骤4 修改完成后，重新登录修改hostname的节点，查看修改的hostname是否生效。

----结束

15.2.9 如何定位进程被 kill

问题背景与现象

在某环境出现DataNode异常重启，且确认此时未从页面做重启DataNode的操作，需要定位是什么进程kill DataNode服务端进程。

原因分析

常见的进程被异常终止有2种原因：

- **Java进程OOM被Kill**

  一般Java进程都会配置OOM Killer，当检测到OOM会自动Kill，OOM日志通常被打印到out日志中，此时可以看运行日志（如DataNode的日志路径为 /var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-主机名.log），看是否有OutOfMemory 内存溢出的打印。

- **被其他进程kill，或者人为kill。**

  排查DataNode运行日志（/var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-主机名.log），是先收到“RECEIVED SIGNAL 15”再健康检查失败。即如下示例中DataNode先于 11:04:48被kill，然后过2分钟，于 11:06:52启动。

  ```
  2018-12-06 11:04：48,433 | ERROR | SIGTERM handler | RECEIVED SIGNAL 15: SIGTERM | LogAdapter.java:69
  2018-12-06 11:04:48,436 | INFO  | Thread-1 | SHUTDOWN_MSG:
  /************************************************************
  SHUTDOWN_MSG: Shutting down DataNode at 192-168-235-85/192.168.235.85
  ************************************************************/ | LogAdapter.java:45
  2018-12-06 11:06:52,744 | INFO  | main | STARTUP_MSG:
  ```

  以上日志说明，DataNode先被其他进程关闭，然后健康检查失败，2分钟后，被NodeAgent启动DataNode进程。

处理步骤

打开操作系统审计日志，给审计日志增加记录kill命令的规则，即可定位是何进程发送的kill命令。

操作影响

- 打印审计日志，会消耗一定操作系统性能，经过分析仅影响不到1%。
- 打印审计日志，会占用一定磁盘空间。该日志打印量不大，MB级别，且默认配置有老化机制和检测磁盘剩余空间机制，不会占满磁盘。

定位方法

在DataNode进程可能发生重启的所有节点，分别执行以下操作。

步骤1 以**root**用户登录节点，执行**service auditd status**命令，确认该服务状态。

```
Checking for service auditd  running
```

如果该服务未启动，执行**service auditd restart**命令重启该服务（无影响，耗时不到1秒）

```
Shutting down auditd done
Starting auditd done
```

步骤2 审计日志临时增加kill命令审计规则。

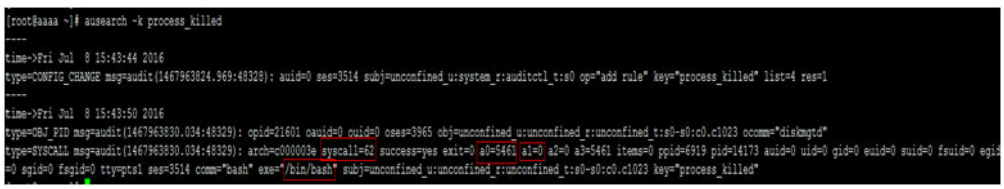
增加规则：

auditctl -a exit,always -F arch=b64 -S kill -S tkill -S tgkill -F a1!=0 -k process_killed

查看规则：

auditctl -l

步骤3 当进程有异常被kill后，使用**ausearch -k process_killed**命令，可以查询kill历史。



说明

a0是被kill进程的pid（16进制），a1是kill命令的信号量。

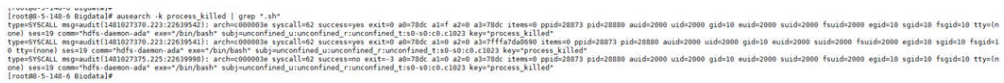
----结束

验证方法

步骤1 从MRS页面重启该节点一个实例，如DataNode。

步骤2 执行**ausearch -k process_killed**命令，确认是否有日志打印。

例如以下命令ausearch -k process_killed |grep “.sh” ，可以看到是hdfs-daemon-ada* 脚本，关闭的DataNode进程。



----结束

停止审计kill命令方法

步骤1 执行**service auditd restart**命令，即会清理临时增加的kill审计日志。

步骤2 执行**auditctl -l**命令，如果没有kill相关信息，即说明已清理该规则。

----结束

15.2.10 MRS 集群使用 pip3 安装 python 包提示网络不可达

用户问题

使用pip3安装python包报错网络不可达。

问题现象

执行pip3 install 安装python包报错网络不可达。具体如下图所示：

原因分析

客户未给Master节点绑定弹性公网IP，造成报错的发生。

处理步骤

步骤1 登录MRS服务管理控制台。

步骤2 选择“集群列表 > 现有集群”，选中当前安装出问题的集群并单击集群名称，进入集群基本信息页面。

步骤3 在“节点管理”页签单击Master节点组中某一Master节点名称，登录到弹性云服务器管理控制台。

步骤4 选择“弹性公网IP”页签，单击“绑定弹性公网IP”为弹性云服务器绑定一个弹性公网IP。

步骤5 登录Master节点执行pip3 install安装python包。

----结束

15.2.11 MRS 集群客户端无法下载

用户问题

在本地的Master主机上想给另外一台远端主机下载一个MRS集群客户端进行使用，但是一直提示网络或者参数有问题

问题现象

在本地的Master主机上想给另外一台远端主机下载一个MRS集群客户端进行使用，但是一直提示网络或者参数有问题

原因分析

- 可能是两台主机处于不同VPC网络中
- 密码填写错误
- 远端主机开启防火墙

处理步骤

- 两台主机处于不同VPC网络中

  放开远端主机的22端口
- 密码填写错误

  请检查密码是否正确，密码中不能有特殊符号。
- 远端主机开启防火墙

  使用规避方案，先将这个MRS集群客户端下载到“服务器端”主机，然后通过linux提供的scp命令远程发送到远端主机。

15.2.12 扩容失败

用户问题

Console界面正常，MRS集群扩容失败

问题现象

Console界面正常，查看MRS Manager界面也无警告无错误，但MRS集群无法扩容报“集群存在非运行状态节点，请稍后重试”的错误。

原因分析

MRS集群的扩缩容要建立在集群处于正常运行的基础上，所以首先要检查集群是否处于正常与否，现在报的是集群存在非运行状态节点，而console界面和MRS Manager

界面都是正常的，所以可能原因就是数据库中集群状态不正常或未刷新导致集群相关节点处于非正常运行状态导致的。

处理步骤

步骤1 登录MRS控制台，单击集群名称进入集群详情页面查看集群状态，确保集群状态为“运行中”。

步骤2 单击“节点管理”，查看所有节点的状态，确保所有节点的状态为“运行中”。

步骤3 登录集群的podMaster节点跳转到MRS的deployer节点，查看api-gateway.log的日志。

1. 用**kubectl get pod -n mrs**命令查看MRS 对应的**deployer**节点的**pod**。
2. 用**kubectl exec -ti ${deployer节点的pod} -n mrs /bin/bash**命令登录相应的pod，如执行**kubectl exec -ti mrsdeployer-78bc8c76cf-mn9ss -n mrs /bin/bash**命令进入MRS的deployer容器。
3. 在/opt/cloud/logs/apigateway目录下查看最新的api-gateway.log日志，检索里面的关键信息（如：ERROR，scaling，clusterScaling，HostState，state-check，集群ID等)查看报错类型。
4. 根据报错提示信息进行相应处理，然后再次执行扩容操作。
   - 扩容成功，则处理完成。
   - 扩容失败，则执行**步骤4**。

步骤4 用**/opt/cloud/mysql -u${用户名} -P${端口} -h${地址} -p${密码}**登录数据库。

步骤5 执行**select cluster_state from cluster_detail where cluster_id="集群ID";**查看cluster_state。

- cluster_state为2，则集群状态正常，执行**步骤6**。
- cluster_state不为2，说明集群状态在数据库中异常，可用**update cluster_detail set cluster_state=2 where cluster_id="集群ID";**刷新集群状态，并查看cluster_state。
  - cluster_state为2，则集群状态正常，执行**步骤6**
  - cluster_state不为2，则请联系技术工程师处理。

步骤6 执行**select host_status from host where cluster_di="clusterID";**命令查询集群主机状态。

- 如果主机状态为started，则处理完成。
- 如果主机状态不为started，则可执行**update host set host_status='started' where cluster_id="集群ID";**命令更新主机状态到数据库。
  - 如果主机状态为started，则处理完成。
  - 如果主机状态不为started，则请联系技术工程师处理。

----结束

15.2.13 MRS 通过 beeline 执行插入命令的时候出错

用户问题

MRS通过beeline执行插入命令的时出错

问题现象

在hive的beeline中执行**insert into**插入语句的时候会报以下的错误:

```
Mapping run in Tez on Hive transactional table fails when data volume is high with error: "org.apache.hadoop.hive.ql.lockmgr.LockException Reason: Transaction... already aborted, Hive SQL state [42000]."
```

原因分析

对于Join操作，由于集群配置不理想和Tez资源设置不合理导致该问题。

处理步骤

可以在beeline上设置配置参数进行解决。

步骤1 设置以下属性以优化性能（建议在集群级别进行更改）

- 设置hive.auto.convert.sortmerge.join = true
- 设置hive.optimize.bucketmapjoin = true
- 设置hive.optimize.bucketmapjoin.sortedmerge = true

步骤2 更改以下内容以调整Tez的资源 。

- 设置hive.tez.container.size = {与YARN容器相同的大小}
- 将hive.tez.container.size设置为与YARN容器大小yarn.scheduler.minimum-allocation-mb相同或更小的值（如设置为二分之一或四分之一的值），但不要超过yarn.scheduler.maximum-allocation-mb 。

----结束

15.2.14 MRS 集群如何进行 Euleros 系统漏洞升级?

用户问题

Euleros系统底层存在漏洞，MRS集群如何进行漏洞升级?

问题现象

在使用绿盟软件测试集群，发现有Euleros系统底层存在漏洞，漏洞报告如下:

绿盟科技风险评估报告
2020-06-18 06:16:23(UTC+08:00)
目录
1.风险概况
1.1风险总览
主机风险
资产总数
漏洞总数
非常危险 （10.0 分）
16
199
1.2风险分布
1.2.1 主机风险分布
总数
16
非常安全
比较安全
比较危险
非常危险
未知
1.2.2 高中低危漏洞
总数
199
高危漏洞
中危漏洞
低危漏洞

IP地址	操作系统	漏洞风险（个）				主机风险值
		高	中	低	合计	
[redacted]	Linux	7	13	21	41	10.0
	Linux	7	11	17	35	10.0
	Linux	6	11	9	26	8.9
	Linux	6	10	8	24	8.9
		6	11	9	26	8.9
	Linux	6	14	23	43	8.9
		6	13	10	29	8.9
		6	13	22	41	8.9
		6	10	9	25	8.9
	Linux	6	9	8	23	8.8
	Linux	6	9	7	22	8.8
		6	9	7	22	8.8
		6	9	8	23	8.8
	Linux	6	9	8	23	8.8
	Windows Server 2008 R2 Sta	1	138	28	167	8.3

原因分析

在使用绿盟软件测试集群，发现有Euleros系统底层存在漏洞，MRS服务部署在Euleros系统中，因此需要进行漏洞升级。

处理步骤

说明

修复漏洞前请确认是否开启了企业主机安全（Host Security Service，简称HSS）服务，如果已开启，需要先暂时关闭HSS服务对MRS集群的监测，漏洞修复完成后重新开启HSS服务。

步骤1 登录MRS服务控制台。

步骤2 单击集群名称进入集群详情页面，并选择“节点管理”。

步骤3 在Core节点组中勾选任意一个Core节点，单击“节点操作 > 停止所有角色”。

步骤4 通过远登录Core节点后台，配置yum源。

步骤5 使用**uname –r**或**rpm –qa |grep kernel**命令，查询并记录当前节点内核版本。

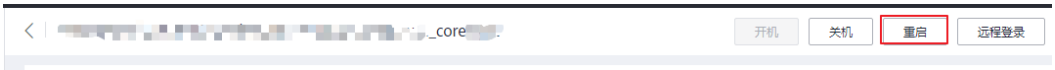
步骤6 执行**yum update –y --skip-broken --setopt=protected_multilib=false**命令更新补丁。

步骤7 完成更新后查询内核版本，并执行**rpm -e** ***旧内核版本*** 命令删除旧内核版本。

步骤8 在集群详情页，选择“节点管理”。

步骤9 在Core节点组中单击已更新补丁的Core名称，进入弹性云服务管理控制台。

步骤10 在页面右上角单击“重启”，重启Core节点。



步骤11 重启完成后，在集群详情页的“节点管理”的Core节点组中勾选Core节点，单击“节点操作 > 启动所有角色”。

步骤12 重复**步骤1**~**步骤11**的操作，升级其他Core节点。

步骤13 所有Core节点升级完成后，参考**步骤1**~**步骤11**的操作先升级备Master节点，再升级主Master节点。

----结束

15.2.15 使用 CDM 迁移数据至 HDFS

用户问题

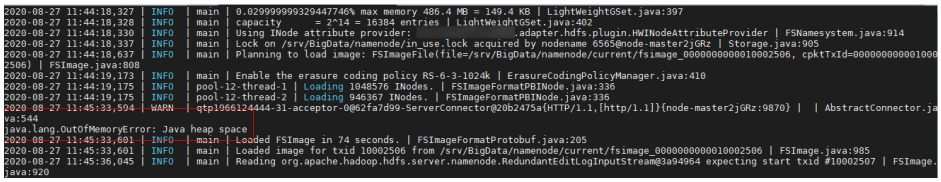
使用CDM从旧的集群迁移数据至新集群的HDFS过程失败。

问题现象

使用CDM从源HDFS导入目的端HDFS，发现目的端MRS集群故障，NameNode无法启动。

查看日志发现在启动过程中存在Java heap space报错，需要修改NN的JVM参数。

图 15-6 故障日志



原因分析

客户在使用CDM迁移数据的过程中，HDFS的数据量过大，导致在合并元数据时发生堆栈异常。

处理步骤

步骤1 进入HDFS服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > HDFS > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > HDFS > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 > HDFS > 配置 > 全部配置”。

步骤2 搜索并修改“HDFS->NameNode” 中的“GC_OPTS”参数，将其中的“-Xms512M -Xmx512M”两个参数的值根据实际情况调整为较大的值。

步骤3 保存配置并重启受影响的服务或实例。

----结束

15.2.16 MRS 集群频繁产生告警

用户问题

集群频繁发出Manager主备节点间心跳中断，DBService主备节点间心跳中断，节点故障等告警，偶尔会造成hive不可用。

问题现象

集群频繁发出Manager主备节点间心跳中断，DBService主备节点间心跳中断，节点故障等告警，偶尔会造成hive不可用，影响客户业务。

原因分析

1. 在出现告警时间点发现虚拟机发生了重启，告警发生的原因是因虚拟机重启导致的。

```
[omm@node-master1yqIY nodeagent]$ last
omm      pts/0        100.125.0.70     Thu Sep 24 10:33   still logged in
omm      pts/1        100.125.0.70     Thu Sep 24 09:26 - 09:47  (00:20)
omm      pts/0        100.125.0.70     Thu Sep 24 09:22 - 10:21  (00:59)
omm      pts/1        100.125.0.70     Wed Sep 23 17:32 - 17:37  (00:05)
root     pts/0        10.203.216.102   Wed Sep 23 17:13 - 18:35  (01:21)
omm      pts/0        100.125.0.70     Wed Sep 23 16:55 - 16:56  (00:00)
omm      pts/0        100.125.0.70     Wed Sep 23 16:20 - 16:25  (00:05)
reboot   system boot  4.19.36-vhulk190 Wed Sep 23 16:06   still running
root     pts/1        10.203.216.102   Tue Sep 22 19:13 - 19:48  (00:34)
omm      pts/0        100.125.0.70     Tue Sep 22 19:08 - 20:03  (00:54)
root     pts/0        10.203.216.102   Tue Sep 22 17:03 - 17:52  (00:48)
omm      pts/1        100.125.0.70     Tue Sep 22 15:55 - 16:00  (00:05)
```

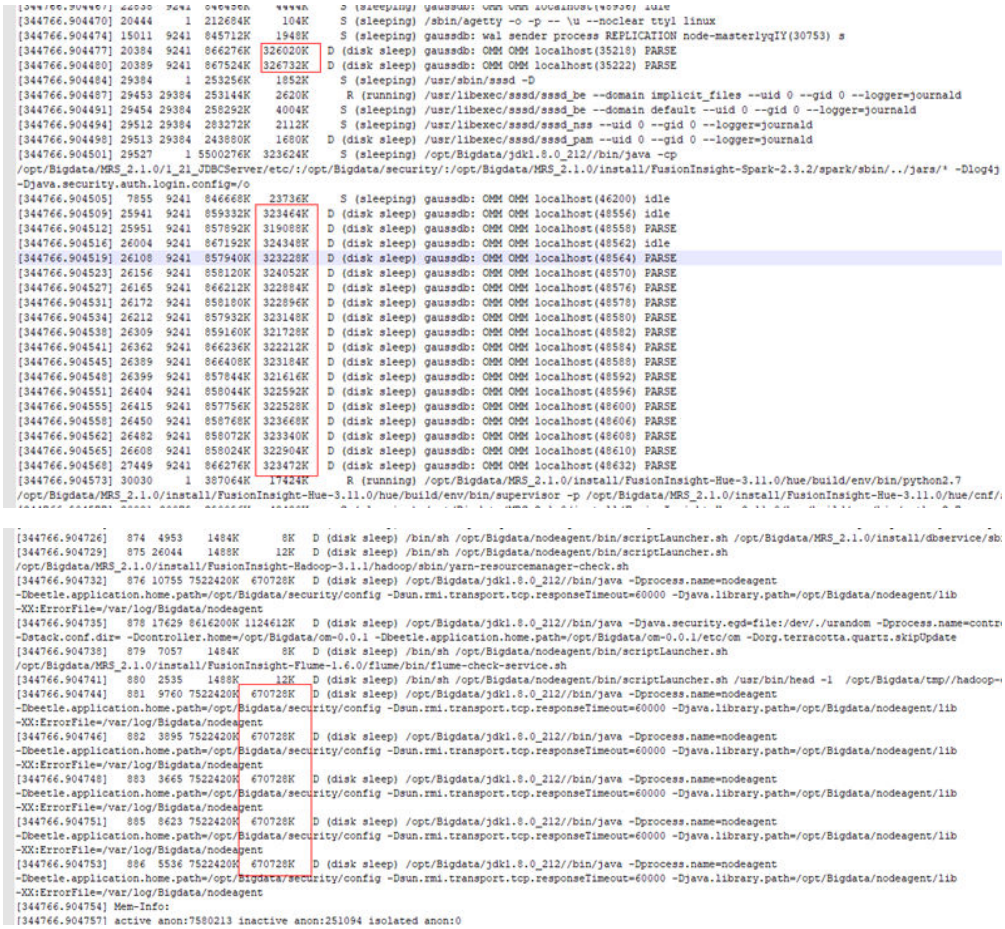
```
[omm@node-master2WbYp ~]$ last
omm      pts/0        10.80.0.56       Thu Sep 24 11:00   still logged in
omm      pts/0        10.80.0.56       Thu Sep 24 09:24 - 10:21  (00:56)
omm      pts/0        10.80.0.56       Wed Sep 23 17:32 - 17:37  (00:05)
omm      pts/0        10.80.0.56       Tue Sep 22 19:15 - 19:15  (00:00)
omm      pts/0        10.80.0.56       Tue Sep 22 15:57 - 16:21  (00:23)
omm      pts/0        10.80.0.56       Tue Sep 22 15:23 - 15:35  (00:12)
omm      pts/0        10.80.0.56       Tue Sep 22 15:07 - 15:12  (00:05)
omm      pts/0        10.80.0.56       Tue Sep 22 14:21 - 14:26  (00:05)
omm      pts/0        10.80.0.56       Mon Sep 21 10:57 - 11:06  (00:09)
omm      pts/0        10.80.0.56       Mon Sep 21 10:42 - 10:56  (00:14)
omm      pts/0        10.80.0.56       Thu Sep 17 16:05 - 16:15  (00:10)
omm      pts/0        10.80.0.56       Wed Sep 16 20:52 - 20:58  (00:06)
reboot   system boot  4.19.36-vhulk190 Wed Sep 16 18:05   still running
omm      pts/0        10.80.0.56       Wed Sep 16 15:43 - 16:10  (00:26)
omm      pts/0        10.80.0.56       Wed Sep 16 14:35 - 14:53  (00:17)
omm      pts/0        10.80.0.56       Wed Sep 16 14:33 - 14:33  (00:00)
omm      pts/0        10.80.0.56       Wed Sep 16 14:11 - 14:29  (00:17)
omm      pts/0        10.80.0.56       Wed Sep 16 14:02 - 14:09  (00:06)
omm      pts/0        10.80.0.56       Wed Sep 16 11:56 - 12:04  (00:08)
omm      pts/0        10.80.0.56       Wed Sep 16 11:26 - 11:31  (00:04)
omm      pts/0        10.80.0.56       Wed Sep 16 11:09 - 11:24  (00:15)
root     pts/0        10.203.230.193   Mon Sep 14 15:54 - 16:30  (00:35)
root     pts/0        10.203.172.29    Fri Sep 11 17:15 - 17:45  (00:30)
root     pts/0        10.203.172.29    Fri Sep 11 16:53 - 17:12  (00:19)
root     tty1                          Fri Sep 11 16:23 - 17:25  (01:01)
reboot   system boot  4.19.36-vhulk190 Fri Sep 11 10:07   still running
reboot   system boot  4.19.36-vhulk190 Thu Aug 27 16:41   still running
root     tty1                          Thu Aug 20 09:46 - 10:17  (00:30)
reboot   system boot  4.19.36-vhulk190 Wed Aug 19 17:48   still running
reboot   system boot  4.19.36-vhulk190 Wed Aug 19 17:46   still running
```

2. 经OS定位虚拟机发生重启的原因是节点没有可用的内存，系统发生内存溢出触发了oom-killer，当进程处于被调用的状态会使进程处于disk sleep状态，最终导致虚拟机发生重启。

```
mem info:
[344766.903734] MemTotal:       32397404 kB
MemFree:          160404 kB
MemAvailable:      31668 kB
Buffers:            2172 kB
Cached:          2768904 kB
SwapCached:            0 kB
Active:         30328872 kB
Inactive:        1035844 kB
Active(anon):   30320852 kB
Inactive(anon):  1004376 kB
Active(file):       8020 kB
Inactive(file):    31468 kB
Unevictable:           0 kB
Mlocked:               0 kB
[344766.903738] SwapTotal:             0 kB
SwapFree:              0 kB
```

总内存

用户态使用



3. 查看占用的内存进程，发现占用内存都是正常的业务进程。

结论：虚拟机内存不能满足服务需求。

处理步骤

- 建议扩大节点内存。
- 建议关闭不需要的服务来规避该问题。

15.2.17 PMS 进程占用内存高问题处理

用户问题

主Master节点内存使用率高如何处理?

问题现象

主Master节点内存使用率高，且用**top -c**命令查询得内存占用量高的是如下idle的进程。

```
12180 ommdba    20   0 1395492 1.180g 1.082g S   0.0  3.8  23:14.29 gaussdb: OMM OMM localhost(60598) idle
14828 ommdba    20   0 1395904 1.180g 1.081g S   0.0  3.8  23:17.08 gaussdb: OMM OMM localhost(60698) idle
15016 ommdba    20   0 1395840 1.180g 1.081g S   0.0  3.8  23:11.19 gaussdb: OMM OMM localhost(60824) idle
14943 ommdba    20   0 1395900 1.180g 1.081g S   0.0  3.8  23:14.76 gaussdb: OMM OMM localhost(60764) idle
14908 ommdba    20   0 1395840 1.180g 1.081g S   0.0  3.8  23:15.18 gaussdb: OMM OMM localhost(60738) idle
14953 ommdba    20   0 1395824 1.180g 1.081g S   0.0  3.8  23:15.96 gaussdb: OMM OMM localhost(60770) idle
14995 ommdba    20   0 1395560 1.180g 1.081g S   0.0  3.8  23:13.28 gaussdb: OMM OMM localhost(60812) idle
15062 ommdba    20   0 1395820 1.180g 1.081g S   0.0  3.8  23:16.12 gaussdb: OMM OMM localhost(60868) idle
15064 ommdba    20   0 1395512 1.180g 1.081g S   0.0  3.8  23:13.33 gaussdb: OMM OMM localhost(60870) idle
14973 ommdba    20   0 1395528 1.180g 1.081g S   0.0  3.8  23:12.74 gaussdb: OMM OMM localhost(60790) idle
14835 ommdba    20   0 1395536 1.180g 1.081g S   0.0  3.8  23:17.39 gaussdb: OMM OMM localhost(60704) idle
14822 ommdba    20   0 1395524 1.180g 1.081g S   0.0  3.8  23:13.80 gaussdb: OMM OMM localhost(60692) idle
14991 ommdba    20   0 1395808 1.180g 1.081g S   0.0  3.8  23:17.96 gaussdb: OMM OMM localhost(60808) idle
14975 ommdba    20   0 1395812 1.180g 1.081g S   0.0  3.8  23:12.57 gaussdb: OMM OMM localhost(60792) idle
15038 ommdba    20   0 1395520 1.180g 1.081g S   0.0  3.8  23:12.75 gaussdb: OMM OMM localhost(60846) idle
14919 ommdba    20   0 1395540 1.180g 1.081g S   0.0  3.8  23:11.58 gaussdb: OMM OMM localhost(60744) idle
14832 ommdba    20   0 1395476 1.180g 1.081g S   0.0  3.8  23:13.11 gaussdb: OMM OMM localhost(60702) idle
14989 ommdba    20   0 1395500 1.180g 1.081g S   0.0  3.8  23:15.63 gaussdb: OMM OMM localhost(60806) idle
14979 ommdba    20   0 1395448 1.180g 1.081g S   0.0  3.8  23:13.17 gaussdb: OMM OMM localhost(60796) idle
15047 ommdba    20   0 1395512 1.180g 1.081g S   0.0  3.8  23:12.10 gaussdb: OMM OMM localhost(60854) idle
14977 ommdba    20   0 1395496 1.180g 1.081g S   0.0  3.8  23:16.90 gaussdb: OMM OMM localhost(60794) idle
15028 ommdba    20   0 1395800 1.180g 1.081g S   0.0  3.8  23:09.35 gaussdb: OMM OMM localhost(60836) idle
```

原因分析

- PostgreSQL缓存：除了常见的执行计划缓存、数据缓存，PostgreSQL为了提高生成执行计划的效率，还提供了catalog，relation等缓存机制。长连接场景下这些缓存中的某些缓存是不会主动释放的，因此可能导致长连接占用大量的内存不释放。
- PMS是MRS的监控进程，此进程会经常创建表分区或者新表，由于PostgreSQL会缓存当前会话访问过的对象的元数据，且PMS的数据库连接池连接会长时间存在，所以连接占用的内存会逐渐上升。

处理步骤

步骤1 以**root**用户登录主Master节点。

步骤2 执行如下命令查询PMS进程号。

ps -ef | grep =pmsd |grep -v grep

步骤3 执行如下命令关闭PMS进程，其中**PID**为步骤2中获取的PMS进程号。

kill -9 PID

步骤4 等待PMS进程自动启动。

PMS启动需要2-3分钟。PMS是监控进程，重启不影响大数据业务。

----结束

15.2.18 Knox 进程占用内存高

用户问题

knox进程占用内存高

问题现象

主Master节点内存使用率高，用**top -c**命令查看到占用内存较高的进程中有knox进程，且此进程占用内存超过4G。

原因分析

knox进程没有单独配置内存，进程会自动根据系统内存大小按照比例划分可用内存，导致knox占用内存大。

处理步骤

步骤1 以root用户分别登录Master节点。

步骤2 打开文件“/opt/knox/bin/gateway.sh”，查找APP_MEM_OPTS，并设置该参数的值为：“-Xms3072m -Xmx4096m”。

步骤3 登录Manager页面，单击“主机管理”，找到主Master节点的IP（即主机名称前带有实心五角星的节点），并登录该节点后台。

步骤4 执行如下命令重启进程。

su - omm

sh /opt/knox/bin/restart-knox.sh

----结束

15.2.19 安全集群外节点安装客户端访问 HBase 很慢

用户问题

安全集群外节点安装了集群的客户端，并使用客户端命令hbase shell访问hbase，发现访问hbase非常慢。

问题现象

客户创建了安全集群，在集群外节点安装了集群的客户端，并使用客户端命令hbase shell访问hbase，发现访问hbase非常慢。

原因分析

安全集群需要进行Kerberos认证，需要在客户端节点的hosts中配置信息，访问速度才不会受到影响。例如，hosts配置信息为：

```
1.1.1.1 hadoop.782670e3_1364_47e2_8c70_1b61bb80479c.com
1.1.1.1 hadoop.hadoop.com
1.1.1.1 hacluster
1.1.1.1 haclusterX
1.1.1.1 haclusterX1
1.1.1.1 haclusterX2
1.1.1.1 haclusterX3
1.1.1.1 haclusterX4
1.1.1.1 ClusterX
1.1.1.1 manager
ip1 hostname1
ip2 hostname2
ip3 hostname3
ip4 hostname4
```

处理步骤

将集群节点上的hosts文件内容复制到安装客户端节点的hosts文件中。

15.2.20 作业无法提交如何定位?

问题背景与现象

客户通过DGC或者在MRS Console无法提交作业。

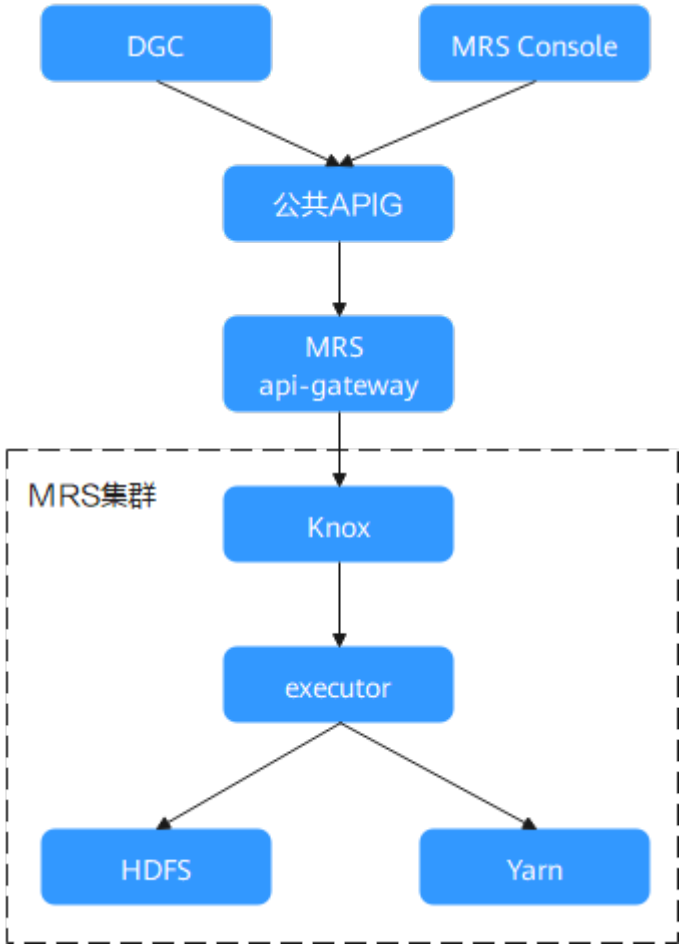
问题影响

作业无法提交，业务中断。

作业流程简介

1. 所有请求会先经过APIG网关，受到APIG配置的流控限制。
2. APIG将请求转发到MRS管控面的api-gateway中。
3. MRS管控面API节点轮询主备oms的Knox，确认主oms的Knox。
4. MRS管控面API提交任务到主oms的Knox。
5. Knox转发请求到本节点的executor进程。
6. executor进程提交任务到Yarn。

图 15-7 作业流程



处理步骤

前期准备：

- 确定作业是通过DGC或在MRS Console提交。
- 准备如表15-1信息。

表 15-1 修复前准备事项

序号	项目	操作方式
1	集群账号信息	申请集群**admin**账户的密码。
2	节点账号信息	申请集群内节点的**omm**、**root**用户密码。
3	SSH远程登录工具	准备PuTTY或SecureCRT等工具。
4	客户端	已提前安装好客户端。

步骤1 确认异常来源。

查看作业日志中收到的错误码，确认错误码是属于APIG还是MRS。

- 若是公共APIG的错误码（APIG的错误码是APIGW开头），联系公共APIG维护人员。
- 若是MRS侧错误，继续下一步。

步骤2 排查服务和进程运行状态等基本情况。

1. 登录Manager界面确认是否有服务故障，如果有作业相关服务故障或者底层基础服务故障，需要解决故障。
2. 查看是否有严重告警。
3. 登录主Master节点。
4. 执行如下命令查看oms状态是否正常，主oms节点executor和knox进程是否正常。knox是双主模式，executor是单主模式。

   /opt/Bigdata/om-0.0.1/sbin/status-oms.sh
5. 以omm用户执行**jmap -heap PID**检查knox和executor进程内存使用情况，如果多次执行查看到老生代内存使用率为99.9%说明有内存溢出。

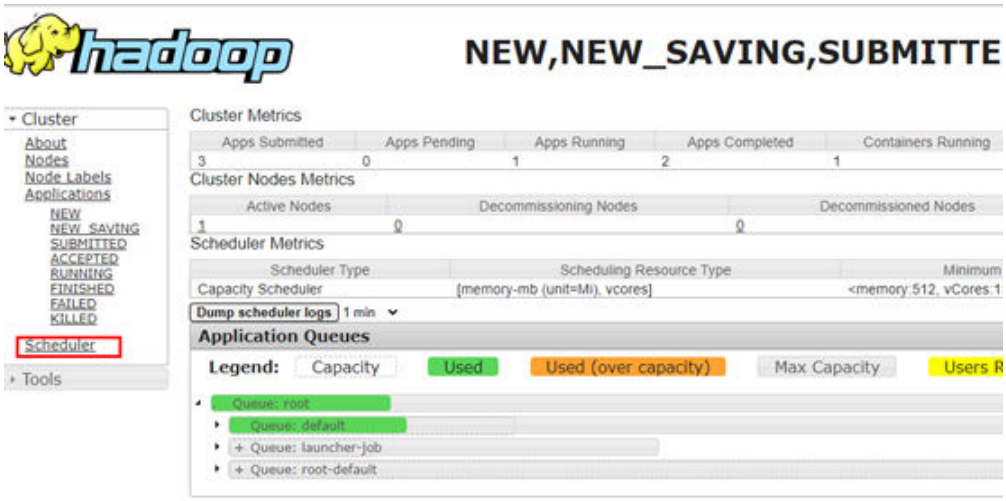
   查询executor进程PID：netstat -anp | grep 8181 | grep LISTEN

   查询knox进程PID：ps -ef|grep knox | grep -v grep

   如果内存溢出，需要现在执行**jmap -dump:format=b,file=/home/omm/temp.bin PID**，导出内存信息后重启进程进行恢复。
6. 查看Yarn的原生界面，确认队列资源情况，以及任务是否提交到了yarn上。

   Yarn的原生界面：在集群详情页选择“组件管理 > Yarn > ResourceManager WebUI > ResourceManager (主)”

图 15-8 Yarn 界面队列资源情况



步骤3 排查任务提交失败点。

1. 登录MRS控制台，单击集群名称进入集群详情页面。
2. 选择“作业管理”页签，单击作业所在行“操作”列的“查看日志”。
3. 若没有日志或者日志信息不详细，则在“作业名称/ID”列复制作业ID。
4. 在主oms节点执行如下命令确认任务请求是否下发到了knox，如果请求没有到knox则可能是knox出了问题，需要尝试重启knox进行恢复。

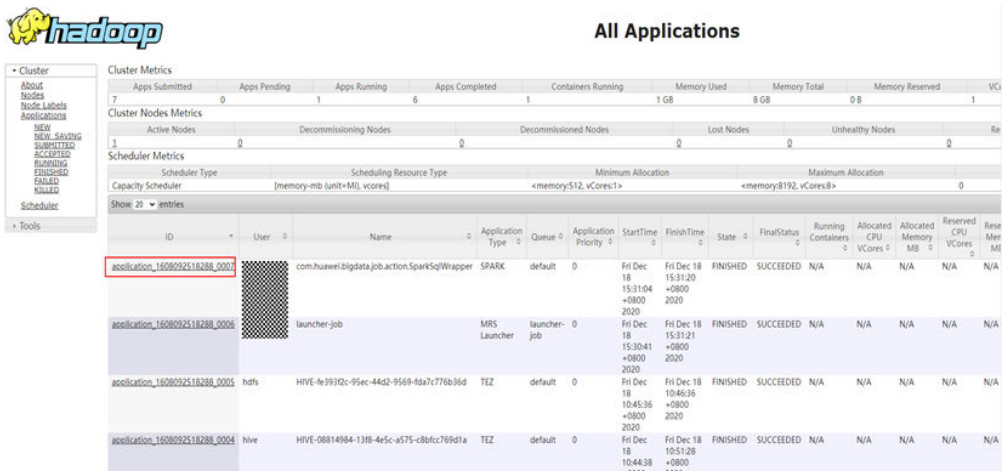
   grep "mrsjob" /var/log/Bigdata/knox/logs/gateway-audit.log | tail -10
5. 进入executor的日志中搜索作业ID，查看报错信息。

   日志路径：/var/log/Bigdata/executor/logs/exe.log
6. 修改“/opt/executor/webapps/executor/WEB-INF/classes/log4j.properties”文件开启executor的debug日志，提交测试任务，查看executor的日志并确认作业提交过程中的报错。

   日志路径：/var/log/Bigdata/executor/logs/exe.log
7. 如果当前任务在exeutor中出错，执行如下命令打印executor的jstack信息，确认线程当前执行状态。

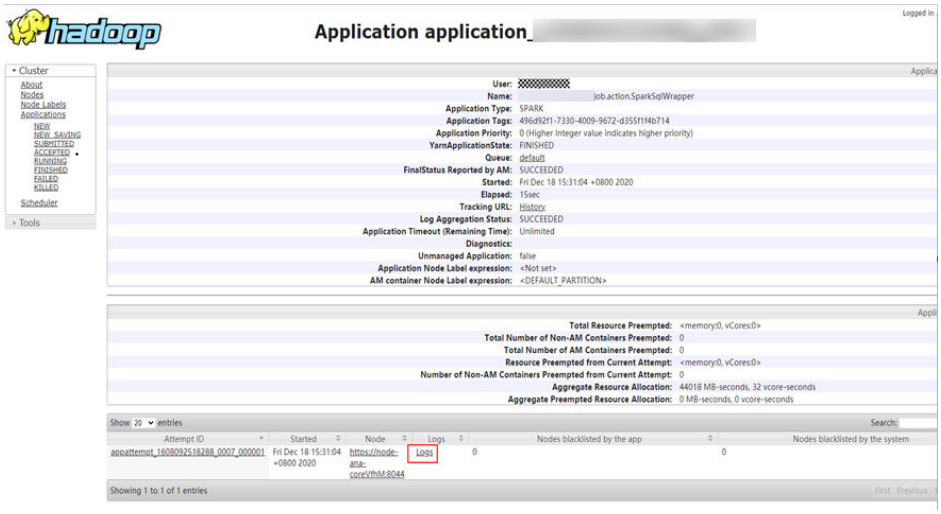
   jstack PID > xxx.log
8. 在集群详情页面选择“作业管理”页签，单击作业所在行“操作”列的“查看详情”，获取“实际作业编号”applicationID。
9. 在集群详情页选择“组件管理 > Yarn > ResourceManager WebUI > ResourceManager (主)”进入Yarn的原生界面，单击applicationID。

图 15-9 Yarn 的 Applications



10. 在任务详情页面查看日志。

图 15-10 任务日志



----结束

15.2.21 HBase 日志文件过大导致 OS 盘空间不足

用户问题

OS盘/var/log分区空间不足。

问题现象

“/var/log/Bigdata/hbase/*/hbase-omm-*.out”日志文件过大，造成OS盘/var/log分区空间不足。

原因分析

在HBase长时间运行场景下，操作系统会把JVM创建的“/tmp/.java_pid*”文件定期清理。因为HBase的内存监控使用了JVM的jinfo命令，而jinfo依赖“/tmp/.java_pid*”文

件，当该文件不存在时，jinfo会执行**kill -3**将jstack信息打印到.out日志文件里，从而导致.out日志文件过大。

处理步骤

在每个HBase实例的节点上部署定期清理.out日志文件的定时任务。后台登录HBase的实例节点，在crontab -e中添加每天0点清理.out日志的定时任务。

crontab -e

00 00 * * * for file in `ls /var/log/Bigdata/hbase/*/hbase-omm-*.out`; do echo "" > $file; done

说明

如果.out大文件出现比较频繁，可以每天清理多次或者调整操作系统的自动清理策略。

15.2.22 Manager 页面新建的租户删除失败

问题现象

在FusionInsight Manager的“租户资源”页面添加租户后，删除租户时，报“删除租户角色失败”。

原因分析

在创建租户时会生成对应的角色，执行删除租户操作时会首先删除对应的角色。此时如果支持权限配置的组件状态异常，则会导致删除这个角色对应的资源权限失败。

处理步骤

步骤1 登录FusionInsight Manager，选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，在“配置资源权限”中单击集群名称，确认可配置资源权限的组件。

步骤3 选择“集群 > 服务”，查看可配置资源权限的组件的运行状态是否都为“良好”。

步骤4 如果不为“良好”，请启动或者修复组件，直至状态为“良好”。

步骤5 再次执行删除租户操作。

----结束

15.3 使用 Alluxo

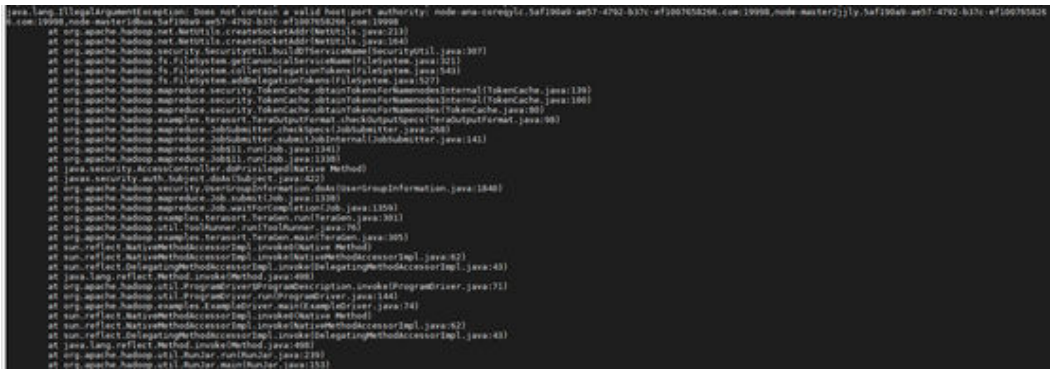
15.3.1 Alluxo 在 HA 模式下出现 Does not contain a valid host:port authority 报错

用户问题

安全集群Alluxo在HA模式下出现Does not contain a valid host:port authority的报错，如何处理?

问题现象

安全集群中，Alluixo在HA模式下出现Does not contain a valid host:port authority的报错。



原因分析

org.apache.hadoop.security.SecurityUtil.buildDTServiceName不支持在uri中填写多个alluxiomaster的地址。

处理步骤

使用alluxio:///或者alluxio://<主AlluxioMaster的ip或hostname>:19998/进行访问。

15.4 使用 ClickHouse

15.4.1 ZooKeeper 上数据错乱导致 ClickHouse 启动失败问题

问题现象

ClickHouse集群中某实例节点启动失败，该实例节点启动日志中有如下类似报错信息：

```
2021.03.15 21:01:19.816593 [ 11111 ] {} <Error> Application: DB::Exception:
The local set of parts of table DEFAULT.lineorder doesn't look like the set ofdoesn't look like the set of
parts in ZooKeeper: 59.99 million rows of 59.99 million total rows in
filesystem are suspicious. There are 30 unexpected parts with 59986052 rows
(14 of them is not just-written with 59986052 rows), 0 missing parts (with 0
blocks).: Cannot attach table `DEFAULT`.`lineorder` from metadata file
...
: while loading database
```

原因分析

使用ClickHouse过程中，ClickHouse实例异常场景下，重复创建集群ReplicatedMergeTree引擎表，后续又进行删除表等操作导致ZooKeeper上的数据异常，致使ClickHouse启动失败。

解决办法

步骤1 备份问题节点数据库下所有表数据到其他目录。

- 备份表数据：
  cd /srv/BigData/data*1*/**clickhouse/data/***数据库名*
  mv *表名 待备份的目录*/**data***1*

  说明

  如果存在多磁盘的场景，需要对**data1**到**dataN**的磁盘数据都执行相同的备份操作。
- 备份元数据信息：
  cd /srv/BigData/data1/**clickhouse_path/metadata**
  mv *表名*.**sql** *待备份的目录*

例如，下面是备份default数据库下的表lineorder数据到**/home/backup**目录下。

cd /srv/BigData/data1/clickhouse/data/default

mv lineorder /home/backup/data1

cd /srv/BigData/data1/clickhouse_path/metadata

mv lineorder.sql /home/backup

步骤2 登录MRS Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，选择对应的实例节点，单击“启动实例”，完成实例启动。

步骤3 实例启动成功后，使用ClickHouse客户端登录问题节点。

clickhouse client --host *clickhouse实例IP* **--user** *用户名* **--password** *密码*

步骤4 执行以下命令获取当前表所在的ZooKeeper路径**zookeeper_path**和对应节点所在的副本编号**replica_num**。

SELECT zookeeper_path FROM system.replicas WHERE database = '*数据库名***' AND table = '***表名***';**

SELECT replica_num,host_name FROM system.clusters;

步骤5 执行以下命令连接ZooKeeper命令行界面。

zkCli.sh -server *ZooKeeper所在节点的IP*:**2181**

步骤6 找到对应故障节点表数据对应的ZooKeeper路径。

ls *zookeeper_path*/**replicas/***replica_num*

说明

zookeeper_path为**步骤4**中查询到的**zookeeper_path**值。
replica_num为**步骤4**中节点主机对应的副本编号**replica_num**的值。

步骤7 执行以下命令，删除ZooKeeper上的副本数据。

deleteall *zookeeper_path*/**replicas/***replica_num*

步骤8 使用ClickHouse客户端登录问题节点，重新执行create创建集群ReplicatedMergeTree引擎表。

clickhouse client --host *clickhouse实例IP* **--multiline --user** *用户名* **--password** *密码*

CREATE TABLE *数据库名.表名* **ON CLUSTER** *集群名*

...

ENGINE = ReplicatedMergeTree ...

其他副本节点有如下提示表已经存在的报错信息，属于正常现象，可以忽略。

```
Received exception from server (version 20.8.7):
Code: 57. DB::Exception: Received from x.x.x.x:9000. DB::Exception:
There was an error on [x.x.x.x:9000]: Code: 57, e.displayText() =
DB::Exception: Table DEFAULT.lineorder already exists. (version 20.8.11.17
(official build)).
```

建表成功后问题节点上表数据会自动进行同步，数据恢复完成。

----结束

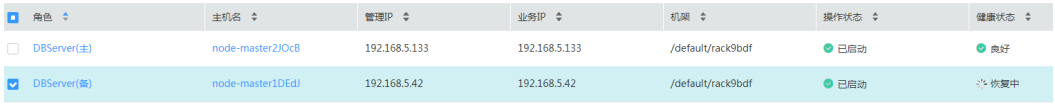
15.5 使用 DBservice

15.5.1 DBServer 实例状态异常

问题背景与现象

DBServer实例状态一直是concerning。

图 15-11 DBServer 实例状态



原因分析

数据目录下文件或目录的权限不对，GaussDB要求文件权限至少是600，目录权限至少为700。

图 15-12 目录权限列表

```
omm@ 192-168-234-176:/srv/BigData/dbdata_service> ll
total 4
drwx------ 19 omm wheel 4096 Dec 14 10:15 data
```

图 15-13 文件权限列表

```
omm@192-168-234-176:/srv/BigData/dbdata_service/data> ll
total 128
drwx------ 6 omm wheel  4096 Dec  9 15:47 base
-rw------- 1 omm wheel   922 Dec  9 15:34 dblink.conf
-rw------- 1 omm wheel    16 Dec 14 10:15 gaussdb.state
drwx------ 2 omm wheel  4096 Dec 14 10:17 global
drwx------ 2 omm wheel  4096 Dec 11 00:00 pg_audit
drwx------ 2 omm wheel  4096 Dec 14 10:15 pg_blackbox
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_clog
drwx------ 2 omm wheel  4096 Dec 14 10:15 pg_confile_backup
-rw------- 1 omm wheel  1024 Dec  9 15:34 pg_ctl.lock
-rw------- 1 omm wheel  4245 Dec  9 15:47 pg_hba.conf
-rw------- 1 omm wheel  1024 Dec  9 15:47 pg_hba.conf.lock
-rw------- 1 omm wheel  1636 Dec  9 15:34 pg_ident.conf
drwx------ 2 omm wheel  4096 Dec  9 15:38 pg_log
drwx------ 4 omm wheel  4096 Dec  9 15:34 pg_multixact
drwx------ 2 omm wheel  4096 Dec 14 10:15 pg_notify
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_serial
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_snapshots
drwx------ 2 omm wheel  4096 Dec 14 11:56 pg_stat_tmp
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_subtrans
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_tblspc
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_twophase
-rw------- 1 omm wheel     4 Dec  9 15:34 PG_VERSION
drwx------ 2 omm wheel  4096 Dec  9 15:34 pg_wallet
drwx------ 3 omm wheel  4096 Dec  9 15:39 pg_xlog
-rw------- 1 omm wheel 13309 Dec 14 10:15 postgresql.conf
-rw------- 1 omm wheel  1024 Dec  9 15:34 postgresql.conf.lock
-rw------- 1 omm wheel   105 Dec 14 10:15 postmaster.opts
-rw------- 1 omm wheel    96 Dec 14 10:15 postmaster.pid
```

解决办法

步骤1 按照图15-12和图15-13的权限列表，修改相应文件和目录的权限。

步骤2 重启相应的DBServer实例。

----结束

15.5.2 DBServer 实例一直处于 Restoring 状态

问题背景与现象

DBServer实例状态一直是Restoring状态，重启之后仍然不恢复。

原因分析

1. DBService组件会对“${BIGDATA_HOME}/MRS_*XXX*/install/dbservice/ha/module/harm/plugin/script/gsDB/.startGS.fail”这个文件监控。其中*XXX*是产品版本号。
2. 如果这个文件中的值大于3就会启动失败，NodeAgent会一直尝试重启该实例，此时仍会失败而且这个值每启动失败一次就会加1。

解决办法

步骤1 登录Manager管理界面。

步骤2 停止该DBServer实例。

步骤3 使用**omm**用户登录到DBServer实例异常的节点。

步骤4 修改“${BIGDATA_HOME}/MRS_*XXX*/install/dbservice/ha/module/harm/plugin/script/gsDB/.startGS.fail”配置文件中的值为0。其中*XXX*是产品版本号。

步骤5 启动该DBServer实例。

----结束

15.5.3 默认端口 20050 或 20051 被占用

问题背景与现象

执行DBService服务重启操作时，DBService服务启动失败，打印的错误日志中出现20050或20051端口被占用等信息。

原因分析

1. 由于DBService使用的默认端口20050或20051被其他进程占用。
2. DBService进程没有停止成功，使用的端口未释放。

解决办法

该解决办法以20051端口被占用为例，20050端口被占用的解决办法与该办法类似。

步骤1 以**root**用户登录DBService安装报错的节点主机，执行命令：**netstat -nap | grep 20051**查看占用20051端口的进程。

步骤2 使用**kill**命令强制终止使用20051端口的进程。

步骤3 约2分钟后，再次执行命令：**netstat -nap | grep 20051**，查看是否还有进程占用该端口。

步骤4 确认占用该端口进程所属的服务，并修改为其他端口。

步骤5 分别在“/tmp”和“/var/run/MRS-DBService/”目录下执行**find . -name "*20051*"**命令，将搜索到的文件全部删除。

步骤6 登录Manager，重启DBService服务。

----结束

15.5.4 /tmp 目录权限不对导致 DBserver 实例状态一直处于 Restoring

问题背景与现象

DBServer实例状态一直是Restoring状态，重启之后仍然不恢复。

原因分析

1. 查看“/var/log/Bigdata/dbservice/healthCheck/dbservice_processCheck.log”，可以看到gaussdb异常。

图 15-14 gaussdb 异常

```
[2019-07-22 10:57:00] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:00] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:10] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:10] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
Normal.
[2019-07-22 10:57:10] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:10] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:20] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:20] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
Normal.
[2019-07-22 10:57:20] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:20] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:30] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:31] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
Normal.
[2019-07-22 10:57:31] ERROR: [:108]: Host 192.168.5.42 gaussdb status is Exception.
[2019-07-22 10:57:31] ERROR: [:154]: Check DBService health failed.
[2019-07-22 10:57:41] INFO: [:84]: check host:192.168.5.42 DBService health.
[2019-07-22 10:57:41] INFO: [:99]: Host 192.168.5.42 floatip status is Normal
```

2. 检查发现“/tmp”权限不对。

图 15-15 /tmp 权限

```
[root@node-master1DEdJ DB]# ll / -rlth
total 76K
drwxr-xr-x.   2 root root 4.0K Dec 12  2016 mnt
drwxr-xr-x.   2 root root 4.0K Dec 12  2016 media
drwxr-xr-x.  13 root root 4.0K Jul 15 16:25 usr
-rwxr-xr-x.   1 root root 3.8K Jul 15 16:25 README
-rwxr-xr-x.   1 root root    0 Jul 15 16:25 OTC_EulerOS_2.x86_64-0.9.1-20170904-0513
lrwxrwxrwx.   1 root root    8 Jul 15 16:26 sbin -> usr/sbin
lrwxrwxrwx.   1 root root    9 Jul 15 16:26 lib64 -> usr/lib64
lrwxrwxrwx.   1 root root    7 Jul 15 16:26 lib -> usr/lib
lrwxrwxrwx.   1 root root    7 Jul 15 16:26 bin -> usr/bin
drwxr-xr-x.   3 root root 4.0K Jul 15 16:29 srv
drwxr-xr-x.   7 root root 4.0K Jul 15 16:39 CloudResetPwdUpdateAgent
drwxr-xr-x.   7 root root 4.0K Jul 15 16:39 CloudrResetPwdAgent
drwx------.   2 root root  16K Jul 15 16:46 lost+found
dr-xr-xr-x. 236 root root    0 Jul 19 17:36 proc
dr-xr-xr-x.   4 root root 4.0K Jul 19 17:37 boot
dr-xr-xr-x.  13 root root    0 Jul 19 17:37 sys
drwxr-xr-x.  19 root root 4.0K Jul 19 17:37 var
drwxr-xr-x.  19 root root 3.0K Jul 19 17:37 dev
drwxr-xr-x.   2 root root 4.0K Jul 19 17:38 tmpdir
drwxr-xr-x.   7 root root 4.0K Jul 19 17:38 opt
-rw-------.   1 root root    0 Jul 19 17:39 install_os_optimization.log
drwxr-xr-x.   6 root root 4.0K Jul 19 17:54 home
drwxr-xr-x.  86 root root 4.0K Jul 19 17:54 etc
drwxr-xr-x.  30 root root  960 Jul 22 10:49 run
drwx------.  23 root root 4.0K Jul 22 11:42 tmp
drwx------.   5 root root 4.0K Jul 22 11:50 root
```

解决办法

步骤1 修改/tmp的权限。

chmod 1777 /tmp

步骤2 等待实例状态恢复。

----结束

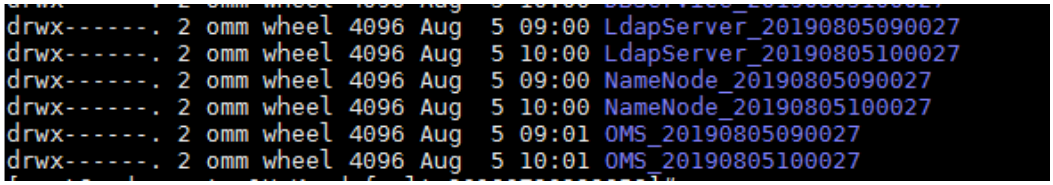
15.5.5 DBService 备份失败

问题背景与现象

ls /srv/BigData/LocalBackup/default_20190720222358/ -rlth

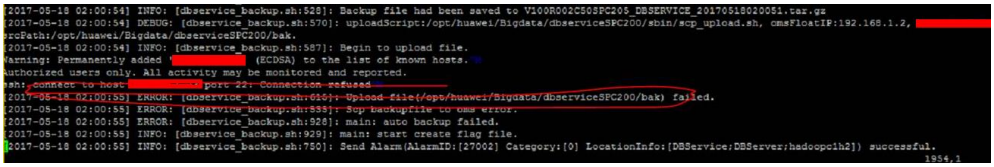
查看备份文件路径中没有DBService的备份文件。

图 15-16 查看备份文件



原因分析

- 查看DBService的备份日志/var/log/Bigdata/dbservice/scriptlog/backup.log，其实备份已经成功，只是上传至OMS节点时失败。



- 失败原因是由于ssh不通。

解决办法

步骤1 网络问题，联系网络工程师处理。

步骤2 网络问题解决之后重新备份即可。

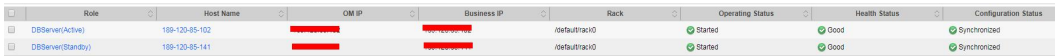
----结束

15.5.6 DBService 状态正常，组件无法连接 DBService

问题背景与现象

上层组件连接DBService失败，检查DBService组件状态正常，两个实例状态也正常。

图 15-17 DBService 状态



原因分析

1. 上层组件是通过dbservice.floatip连接的DBService。
2. 在DBServer所在节点执行命令netstat -anp | grep 20051发现，DBService的Gauss进程在启动时并未绑定floatip，只监听了127.0.0.1的本地ip。

解决办法

步骤1 重新启动DBService服务。

步骤2 启动完成之后在主DBServer节点执行**netstat -anp | grep 20051**命令检查是否绑定了dbservice.floatip。

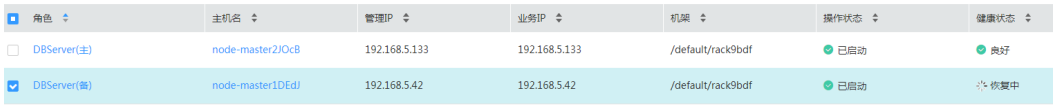
----结束

15.5.7 DBServer 启动失败

问题背景与现象

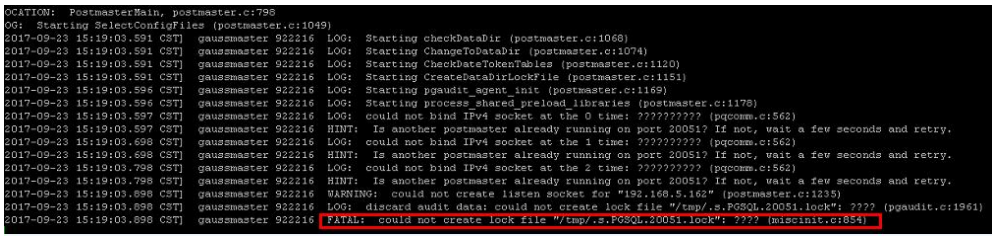
DBService组件启动失败，重启还是失败，实例状态一直为正在恢复状态。

图 15-18 DBService 的状态

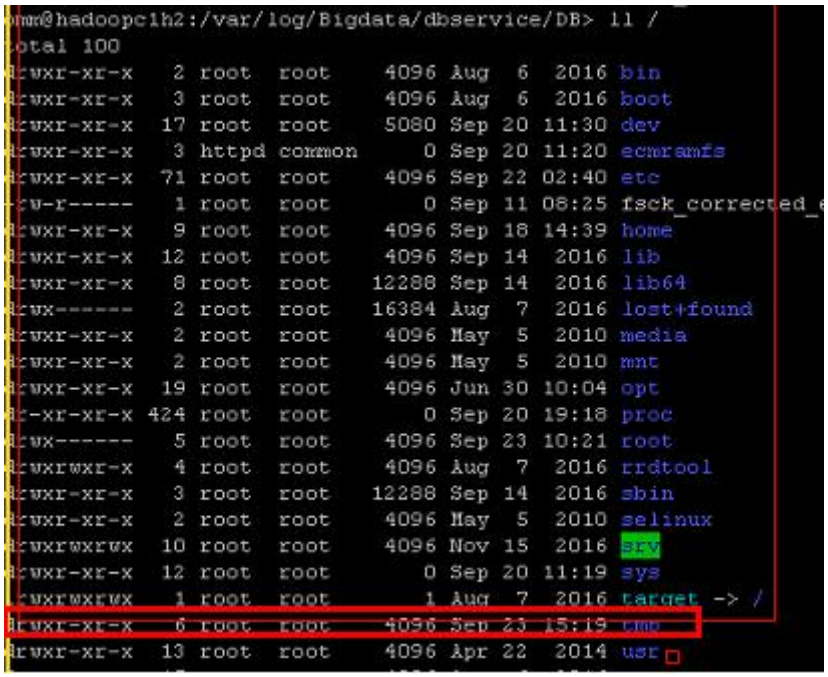


原因分析

1. 查看DBService的日志/var/log/Bigdata/dbservice/DB/gs_ctl-current.log，报如下错误。



2. 检查发现/tmp权限不正确，正确的权限应该为777。



解决办法

步骤1 修改/tmp权限为777。

步骤2 重新启动DBService组件。

----结束

15.5.8 浮动 IP 不通导致 DBService 备份失败

问题背景与现象

在默认备份default中DBService备份失败，其他备份（NameNode、LdapServer、OMS备份）成功。

原因分析

1. 查看DBService的备份页面错误信息，有如下错误信息提示：

```
Clear temporary files at backup checkpoint DBService_test_DBService_DBService_20180326155921 that fialed last time.
Temporyary files at backup checkpoint DBService_test_DBService_DBService20180326155921 that failed last time are cleared successfully.
Start executing the backup task.
The backup of configuration DBService is started.
Check the backup available disk space.
Backup initialization succeeded for configuration DBService.
Clear temporary files at backup checkpoint DBService_test_DBService_DBService_20180326155921 that failed last time.
Temporary files at backup checkpoint DBService_test_DBService_DBService_20180326155921 that failed last time are cleared successfully.
Checkpoint DBService_test_DBService_DBService_20180326162235 is verified successfully before backup.
Temporary files are cleared successfully before backup checkpoint DBService_test_DBService_DBService_20180326162235.
Prestart backup succeeded for checkpoint DBService_test_DBService_DBService_20180326162235.
The snapshot is created successfully for checkpoint DBService_test_DBService_DBService_20180326162235 before backup.
Backup is being performed for checkpoint DBService_test_DBService_DBService_20180326162235.
Backup execution failed. Task ID: 2
Detail: DBService backup task failed, please view details in logs.
Temporary files are cleared successfully after backup checkpoint DBService_test_DBService_DBService_20180326162235.
checkpoint DBService_test_DBService_DBService_20180326162235 is deleted successfully after backup failure.
Failed to backup configuration DBService.
```

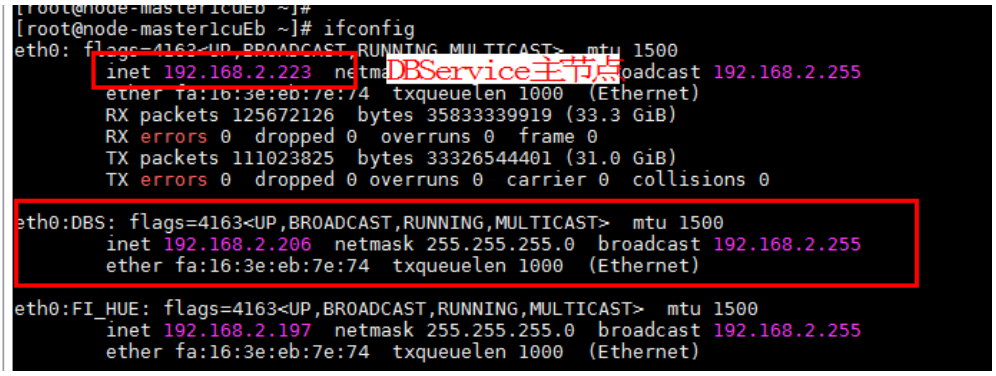
2. 查看/var/log/Bigdata/dbservice/scriptlog/backup.log文件，发现日志停止打印，并没有备份相关信息。
3. 查看主OMS节点 /var/log/Bigdata/controller/backupplugin.log日志发现如下错误信息：

```
result error is ssh:connect to host 172.16.4.200 port 22 : Connection refused (172.16.4.200是DBService的浮动IP)
DBService backup failed.
```

```
2018-03-27 07:00:35,758 INFO  [pool-1-thread-5] Create adapter from com.huawei.bigdata.om.backup.MetadataPluginAdapter success.
com.huawei.bigdata.om.backup.plugin.AbstractBackupRecoveryPlugin.initializePluginAdapter(AbstractBackupRecoveryPlugin.java:92)
2018-03-27 07:00:35,759 INFO  [pool-1-thread-5] floatIp is 172.16.4.200. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.getFloatIp(BackupRecoveryPlugin.java:233)
2018-03-27 07:00:35,759 INFO  [pool-1-thread-5] cmd is ssh 172.16.4.200 /opt/huawei/Bigdata/FusionInsight_V100R002C60U20/dbservice/sbin/dbservice_backup.sh -b -d
/srv/BigData/LocalBackup/default_20180326213206/DBService_20180327070010. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.startBackup(BackupRecoveryPlugin.java:166)
2018-03-27 07:00:35,759 INFO  [pool-1-thread-5] create task taskId is 6. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.startBackup(BackupRecoveryPlugin.java:169)
2018-03-27 07:00:35,760 INFO  [pool-1-thread-5] startBackup result OperateResult(errorCode:RUNNING, result:6, detailInfo:, packageName:null).
com.huawei.bigdata.om.backup.BackupPluginContainerHandler.startBackup(BackupPluginContainerHandler.java:246)
2018-03-27 07:00:35,760 INFO  [Thread-132] Executing the command with arguments and env, timeout: 900000
com.huawei.bigdata.om.controller.api.extern.monitor.script.LinuxScriptExecutionHandler.logMessage(LinuxScriptExecutionHandler.java:64)
2018-03-27 07:00:35,863 INFO  [Thread-132] Execute command : /opt/huawei/Bigdata/om-0.0.1/sbin/scriptLauncher.sh ssh 172.16.4.200
/opt/huawei/Bigdata/FusionInsight_V100R002C60U20/dbservice/sbin/dbservice_backup.sh -b -d /srv/BigData/LocalBackup/default_20180326213206/DBService_20180327070010.
com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:48)
2018-03-27 07:00:35,863 INFO  [Thread-132] result status is 255. com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:49)
2018-03-27 07:00:35,863 INFO  [Thread-132] result output is . com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:50)
2018-03-27 07:00:35,863 INFO  [Thread-132] result erro is ssh: connect to host 172.16.4.200 port 22: Connection refused
. com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:51)
2018-03-27 07:00:35,863 ERROR [Thread-132] DBService backup failed. com.huawei.bigdata.om.dbservice.backup.BackupTask.run(BackupTask.java:64)
2018-03-27 07:00:40,868 INFO  [pool-1-thread-5] query backup taskId is 6. com.huawei.bigdata.om.dbservice.backup.BackupRecoveryPlugin.getBackupProgress(BackupRecoveryPlugin.java:247)
```

解决办法

步骤1 登录DBService主节点（绑定有DBService浮动IP的master节点）。



步骤2 检查 /etc/ssh/sshd_config文件中ListenAddress配置项，添加DBService浮动IP到ListenAdress或者注释掉ListenAdress配置项。

步骤3 执行如下命令重启sshd服务。

service sshd restart

步骤4 观察下次备份DBServie是否备份成功。

----结束

15.5.9 DBService 配置文件丢失导致启动失败

问题背景与现象

节点异常下电，重启备DBService失败。

原因分析

1. 查看/var/log/Bigdata/dbservice/DB/gaussdb.log日志没有内容。
2. 查看/var/log/Bigdata/dbservice/scriptlog/preStartDBService.log日志，发现如下信息，判断为配置信息丢失。

```
The program "gaussdb" was found by "
/opt/Bigdata/MRS_xxx/install/dbservice/gaussdb/bin/gs_guc)
But not was not the same version as gs_guc.
Check your installation.
```

3. 比对主备DBServer节点/srv/BigData/dbdata_service/data目录下的配置文件发现差距比较大。

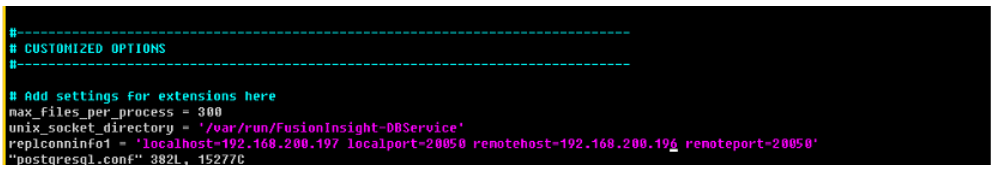
```
omm@hadoopc1h3:/srv/BigData/dbdata_service/data> ll
total 128
-rw------- 1 omm wheel     4 May  8 09:54 PG_VERSION
drwx------ 2 omm wheel  4096 May  8 09:54 bak
drwx------ 7 omm wheel  4096 May  8 09:54 base
-rw------- 1 omm wheel   922 May  8 09:54 dblink.conf
-rw------- 1 omm wheel    16 May  8 09:59 gaussdb.state
drwx------ 2 omm wheel  4096 May  8 09:58 global
drwx------ 2 omm wheel  4096 May  8 09:54 pg_audit
drwx------ 2 omm wheel  4096 May  8 09:58 pg_blackbox
drwx------ 2 omm wheel  4096 May  8 09:54 pg_clog
drwx------ 2 omm wheel  4096 May  8 09:58 pg_confile_backup
-rw------- 1 omm wheel     0 May  8 09:54 pg_ctl.lock
-rw------- 1 omm wheel  4287 May 18  2017 pg_hba.conf
-rw------- 1 omm wheel  1024 May  8 09:54 pg_hba.conf.lock
-rw------- 1 omm wheel  1636 May  8 09:54 pg_ident.conf
drwx------ 2 omm wheel  4096 May  8 09:54 pg_log
drwx------ 4 omm wheel  4096 May  8 09:54 pg_multixact
drwx------ 2 omm wheel  4096 May  8 09:58 pg_notify
drwx------ 2 omm wheel  4096 May  8 09:54 pg_serial
drwx------ 2 omm wheel  4096 May  8 09:54 pg_snapshots
drwx------ 2 omm wheel  4096 May  8 09:58 pg_stat_tmp
drwx------ 2 omm wheel  4096 May  8 09:54 pg_subtrans
drwx------ 2 omm wheel  4096 May  8 09:54 pg_tblspc
drwx------ 2 omm wheel  4096 May  8 09:54 pg_twophase
drwx------ 2 omm wheel  4096 May  8 09:54 pg_wallet
drwx------ 3 omm wheel  4096 May  8 09:54 pg_xlog
-rw------- 1 omm wheel 15277 May  8 09:59 postgresql.conf
-rw------- 1 omm wheel  1024 May  8 09:54 postgresql.conf.lock
-rw------- 1 omm wheel   134 May  8 09:59 postmaster.opts
-rw------- 1 omm wheel   127 May  8 09:58 postmaster.pid
```

```
omm@hadoopc1h3:/srv/BigData/dbdata_service> cd data_bak/
omm@hadoopc1h3:/srv/BigData/dbdata_service/data_bak> ll
total 64
-rw------- 1 omm wheel   202 Feb 11 10:43 backup_label
-rw------- 1 omm wheel     0 Feb 11 10:42 build_completed.start
-rw------- 1 omm wheel    16 Apr 28 17:32 gaussdb.state
-rw------- 1 omm wheel     7 Apr 28 17:32 gs_build.pid
drwx------ 2 omm wheel  4096 Feb 11 10:44 pg_audit
drwx------ 2 omm wheel  4096 Feb 11 10:41 pg_blackbox
drwx------ 2 omm wheel  4096 Feb 11 10:09 pg_confile_backup
-rw------- 1 omm wheel     0 Apr 28 17:32 pg_ctl.lock
-rw------- 1 omm wheel  4287 May 18  2017 pg_hba.conf
drwx------ 2 omm wheel  4096 Feb 11 10:43 pg_notify
drwx------ 2 omm wheel  4096 Feb 11 10:43 pg_xlog
-rw------- 1 omm wheel 15155 May  7 15:33 postgresql.conf
-rw------- 1 omm wheel  1024 May  7 15:33 postgresql.conf.lock
-rw------- 1 omm wheel   134 Feb 11 10:42 postmaster.opts
```

解决办法

步骤1 把主节点**/srv/BigData/dbdata_service/data**的内容拷贝到备节点，保持文件权限和属组与主节点一样。

步骤2 修改postgresql.conf配置信息，localhost修改成本节点IP，remotehost修改成对端节点IP。



步骤3 登录Manager页面重启备DBServer节点。

----结束

15.6 使用 Flink

15.6.1 安装客户端执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file :"security.kerberos.login.keytab"

问题背景与现象

客户端安装成功，执行客户端命令例如yarn-session.sh命令报错，提示IllegalConfigurationException: Error while parsing YAML configuration file : "security.kerberos.login.keytab: "

```
[root@8-5-131-10 bin]# yarn-session.sh
2018-10-25 01:22:06,454 | ERROR | [main] | Error while trying to split key and value in configuration
file /opt/flinkclient/Flink/flink/conf/flink-conf.yaml:80: "security.kerberos.login.keytab: " |
org.apache.flink.configuration.GlobalConfiguration (GlobalConfiguration.java:160)
Exception in thread "main" org.apache.flink.configuration.IllegalConfigurationException: Error while parsing
YAML configuration file :80: "security.kerberos.login.keytab: "
        at org.apache.flink.configuration.GlobalConfiguration.loadYAMLResource(GlobalConfiguration.java:161)
        at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:112)
        at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:79)
        at org.apache.flink.yarn.cli.FlinkYarnSessionCli.main(FlinkYarnSessionCli.java:482)
[root@8-5-131-10 bin]#
```

原因分析

在安全集群环境下，Flink需要进行安全认证。当前客户端未进行相关安全认证设置。

1. Flink整个系统有两种认证方式：
   - 使用kerberos认证：Flink yarn client、Yarn Resource Manager、JobManager、HDFS、TaskManager、Kafka和Zookeeper。
   - 使用YARN内部的认证机制：Yarn Resource Manager与Application Master（简称AM）。
2. 如果用户安装安全集群需要使用kerberos认证和security cookie认证。根据日志提示，发现配置文件中“security.kerberos.login.keytab :”配置项错误，未进行安全配置。

解决办法

步骤1 从MRS上下载用户keytab，并将keytab放到Flink客户端所在主机的某个文件夹下。

步骤2 在“flink-conf.yaml”上配置：

1. keytab路径。
   ```
   security.kerberos.login.keytab: /home/flinkuser/keytab/abc222.keytab
   ```
   说明
   - “/home/flinkuser/keytab/abc222.keytab”表示的是用户目录，为步骤1中放置目录。
   - 请确保客户端用户具备对应目录权限。
2. principal名。
   ```
   security.kerberos.login.principal: abc222
   ```
3. 对于HA模式，如果配置了ZooKeeper，还需要设置ZK kerberos认证相关的配置。配置如下：
   ```
   zookeeper.sasl.disable: false
   security.kerberos.login.contexts: Client
   ```
4. 如果用户对于Kafka client和Kafka broker之间也需要做kerberos认证，配置如下：
   ```
   security.kerberos.login.contexts: Client,KafkaClient
   ```

----结束

15.6.2 安装客户端修改配置后执行命令错误，提示 IllegalConfigurationException: Error while parsing YAML configuration file

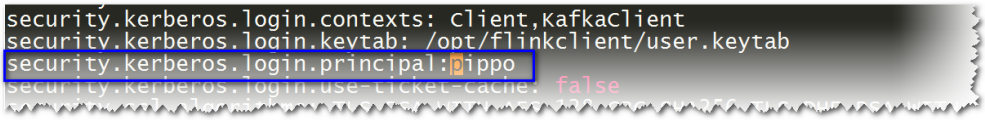
问题背景与现象

客户端安装成功，执行客户端命令例如yarn-session.sh命令报错，提示IllegalConfigurationException: Error while parsing YAML configuration file :81: "security.kerberos.login.principal:pippo "

```
[root@8-5-131-10 bin]# yarn-session.sh
2018-10-25 19:27:01,397 | ERROR | [main] | Error while trying to split key and value in configuration file /opt/flinkclient/Flink/flink/conf/flink-conf.yaml:81: "security.kerberos.login.principal:pippo " | org.apache.flink.configuration.GlobalConfiguration (GlobalConfiguration.java:160)
Exception in thread "main" org.apache.flink.configuration.IllegalConfigurationException: Error while parsing YAML configuration file :81: "security.kerberos.login.principal:pippo "
        at org.apache.flink.configuration.GlobalConfiguration.loadYAMLResource(GlobalConfiguration.java:161)
        at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:112)
        at org.apache.flink.configuration.GlobalConfiguration.loadConfiguration(GlobalConfiguration.java:79)
        at org.apache.flink.yarn.cli.FlinkYarnSessionCli.main(FlinkYarnSessionCli.java:482)
```

原因分析

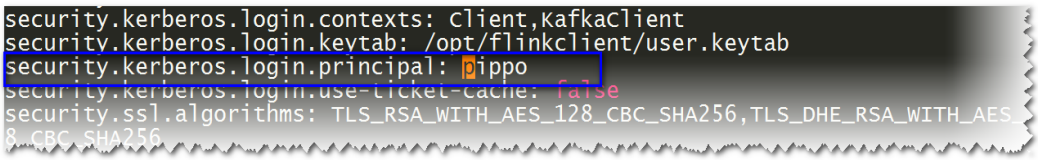
配置文件flink-conf.yaml中配置项"security.kerberos.login.principal:pippo”错误。



解决办法

修改flink-conf.yaml中配置。

注意：配置项名称和值之间存在空格。



15.6.3 创建 Flink 集群时执行 yarn-session.sh 命令失败

问题背景与现象

创建Fllink集群时，执行yarn-session.sh命令卡住一段时间后报错：

```
2018-09-20 22:51:16,842 | WARN  | [main] | Unable to get ClusterClient status from Application Client | org.apache.flink.yarn.YarnClusterClient (YarnClusterClient.java:253)
org.apache.flink.util.FlinkException: Could not connect to the leading JobManager. Please check that the JobManager is running.
    at org.apache.flink.client.program.ClusterClient.getJobManagerGateway(ClusterClient.java:861)
    at org.apache.flink.yarn.YarnClusterClient.getClusterStatus(YarnClusterClient.java:248)
    at org.apache.flink.yarn.YarnClusterClient.waitForClusterToBeReady(YarnClusterClient.java:516)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli.run(FlinkYarnSessionCli.java:717)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli$1.call(FlinkYarnSessionCli.java:514)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli$1.call(FlinkYarnSessionCli.java:511)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1729)
    at org.apache.flink.runtime.security.HadoopSecurityContext.runSecured(HadoopSecurityContext.java:41)
    at org.apache.flink.yarn.cli.FlinkYarnSessionCli.main(FlinkYarnSessionCli.java:511)
Caused by: org.apache.flink.runtime.leaderretrieval.LeaderRetrievalException: Could not retrieve the leader gateway.
    at org.apache.flink.runtime.util.LeaderRetrievalUtils.retrieveLeaderGateway(LeaderRetrievalUtils.java:79)
    at org.apache.flink.client.program.ClusterClient.getJobManagerGateway(ClusterClient.java:856)
    ... 10 common frames omitted
Caused by: java.util.concurrent.TimeoutException: Futures timed out after [10000 milliseconds]
```

可能原因

Flink开启了SSL通信加密，却没有正确的配置SSL证书。

解决办法

针对MRS 2.x及之前版本，操作如下：

方法1：

关闭Flink SSL通信加密，修改客户端配置文件**conf/flink-conf.yaml**。

```
security.ssl.internal.enabled: false
```

方法2：

开启Flink SSL通信加密，security.ssl.internal.enabled 保持默认。正确配置SSL：

- 配置keystore或truststore文件路径为**相对路径**时，Flink Client执行命令的目录需要可以直接访问该相对路径

  ```
  security.ssl.internal.keystore: ssl/flink.keystore
  security.ssl.internal.truststore: ssl/flink.truststore
  ```

在Flink的CLI yarn-session.sh命令中增加“-t”选项来传输keystore和truststore文件到各个执行节点。如：

```
yarn-session.sh -t ssl/ 2
```

- 配置keystore或truststore文件路径为**绝对路径**时，需要在Flink Client以及各个节点的该绝对路径上放置keystore或truststore文件。

```
security.ssl.internal.keystore: /opt/client/Flink/flink/conf/flink.keystore
security.ssl.internal.truststore: /opt/client/Flink/flink/conf/flink.truststore
```

针对MRS3.x及之后版本，操作如下：

方法1：

关闭Flink SSL通信加密，修改客户端配置文件**conf/flink-conf.yaml**。

```
security.ssl.enabled: false
```

方法2：

开启Flink SSL通信加密，security.ssl.enabled 保持默认。正确配置SSL：

- 配置keystore或truststore文件路径为**相对路径**时，Flink Client执行命令的目录需要可以直接访问该相对路径

```
security.ssl.keystore: ssl/flink.keystore
security.ssl.truststore: ssl/flink.truststore
```

在Flink的CLI yarn-session.sh命令中增加“-t”选项来传输keystore和truststore文件到各个执行节点。如：

```
yarn-session.sh -t ssl/ 2
```

- 配置keystore或truststore文件路径为**绝对路径**时，需要在Flink Client以及各个节点的该绝对路径上放置keystore或truststore文件。

```
security.ssl.keystore: /opt/Bigdata/client/Flink/flink/conf/flink.keystore
security.ssl.truststore: /opt/Bigdata/client/Flink/flink/conf/flink.truststore
```

15.6.4 使用不同用户，执行 yarn-session 创建集群失败

问题背景与现象

使用Flink过程中，具有两个相同权限用户testuser和bdpuser。

使用用户testuser创建Flink集群正常，但是切换至bdpuser用户创建Fllink集群时，执行yarn-session.sh命令报错：

```
2019-01-02 14:28:09,098 | ERROR | [main] | Ensure path threw exception |
org.apache.flink.shaded.curator.org.apache.curator.framework.imps.CuratorFrameworkImpl
(CuratorFrameworkImpl.java:566)
org.apache.flink.shaded.zookeeper.org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /flink/application_1545397824912_0022
```

可能原因

高可用配置项未修改。由于在Flink的配置文件中，**high-availability.zookeeper.client.acl**默认为**creator**，仅创建者有权限访问，新用户无法访问ZooKeeper上的目录导致yarn-session.sh执行失败。

解决办法

步骤1 修改客户端配置文件**conf/flink-conf.yaml**中配置项**high-availability.zookeeper.path.root**，例如：

```
high-availability.zookeeper.path.root： flink2
```

步骤2 重新提交任务。

----结束

15.6.5 Flink 业务程序无法读取 NFS 盘上的文件

用户问题

Flink业务程序无法读取集群节点挂载的NFS盘上的文件。

问题现象

用户开发的Flink业务程序中需要读取用户定义的配置文件，该配置文件放在NFS盘上，NFS盘是挂载在集群节点上的，集群的所有节点均可以访问该盘。用户提交flink程序后，业务代码访问不到客户自定义的配置文件，导致业务程序启动失败。

原因分析

该问题的根因是NFS盘上的根目录权限不足，导致Flink程序启动后无法访问该目录。

MRS的Flink任务是在YARN运行，当集群为普通集群时，在YARN上运行任务的用户为yarn_user。用户自定义的配置文件如果在任务启动之后使用，则文件以及文件的父目录（NFS上的文件所在的父目录，非集群节点上的软连接），必须允许yarn_user可以访问，否则程序中无法获取文件内容。当集群为kerberos集群时，则文件的权限必须允许提交程序的用户访问。

处理步骤

步骤1 以root用户登录集群的Master节点。

步骤2 执行如下命令查看用户自定义配置文件所在父目录的权限。

ll <文件所在路径的父目录路径>

步骤3 进入NFS盘待访问文件所在目录，修改用户自定义配置文件所在父目录的权限为755。

chmod 755 -R /<文件所在路径的父目录路径>

步骤4 确认Core或者Task节点是否可以访问到该配置文件。

1. 以root用户登录Core/Task节点。
   如果当前集群已启用Kerberos认证，请以root用户登录Core节点。
2. 执行**su - yarn_user**命令切换到yarn_user用户。
   如果当前集群已启用Kerberos认证，请执行 **su - 提交作业的用户**命令切换用户。
3. 执行如下命令查看用户权限，文件所在路径请使用该文件的绝对路径。
   ll <文件所在路径>

----结束

建议与总结

当用户提交的任务中要访问自定义的配置文件时，特别是挂载NFS盘时，除了确认文件的权限之外，还要确认文件所在父目录的权限是否正确。NFS盘挂载到MRS集群节点上，一般会新建软连接到NFS目录，这个时候需要查看NFS上的目录权限是否正确。

15.6.6 自定义 Flink log4j 日志输出级别

用户问题

MRS 3.1.0集群自定义Flink log4j日志级别不生效。

问题现象

1. 在使用MRS 3.1.0集群Flink数据分析时，将$Flink_HOME/conf目录下的log4j.properties文件中日志级别修改为INFO级别日志。
2. 任务正常提交后，console未打印出INFO级别日志，输出的日志级别还是ERROR级别。

原因分析

修改$Flink_HOME/conf目录下的log4j.properties文件，控制的是JobManager和TaskManager的算子内的日志输出，输出的日志会打印到对应的yarn contain中，可以在yarn web ui查看对应日志。MRS 3.1.0及之后版本的Flink 1.12.0版本开始默认的日志框架是log4j2，配置的方式跟之前log4j的方式有区别，使用如log4j日志规则不会生效。

处理步骤

Log4j2详细日志规格配置参考开源官方文档：**http://logging.apache.org/log4j/2.x/manual/configuration.html#Properties**

15.7 使用 Flume

15.7.1 Flume 向 Spark Streaming 提交作业，提交到集群后报类找不到

用户问题

Flume向Spark Streaming提交作业，提交到集群后报类找不到。

问题现象

Spark Streaming代码打成jar包提交到集群后报类找不到错误，通过以下两种方式依然不生效。

1. 在提交Spark作业的时候使用**--jars** 命令引用类所在的jar包。
2. 将类所在的jar包引入Spark Streaming的jar包。

原因分析

执行Spark作业时无法加载部分jar，导致找不到class。

处理步骤

步骤1 使用 **--jars** 加载flume-ng-sdk-{version}.jar依赖包。

步骤2 同时修改**spark-default.conf**中两个配置项。

spark.driver.extraClassPath=$PWD/*:{加上原来配置的值}

spark.executor.extraClassPath =$PWD/*

步骤3 作业运行成功。如果还有报错，则需要排查还有哪个jar没有加载，再次执行步骤1和步骤2。

----结束

15.7.2 Flume 客户端安装失败

问题现象

安装Flume客户端失败，提示JAVA_HOME is null或flume has been installed。

```
CST 2016-08-31 17:02:51 [flume-client install]: JAVA_HOME is null in current user,please install the JDK and set the JAVA_HOME
CST 2016-08-31 17:02:51 [flume-client install]: check environment failed.
CST 2016-08-31 17:02:51 [flume-client install]: check param failed.
CST 2016-08-31 17:02:51 [flume-client install]: install flume client failed.
```

```
CST 2016-08-31 17:03:58 [flume-client install]: flume has been installed
CST 2016-08-31 17:03:58 [flume-client install]: check path failed.
CST 2016-08-31 17:03:58 [flume-client install]: check param failed.
CST 2016-08-31 17:03:58 [flume-client install]: install flume client failed.
```

原因分析

- Flume客户端安装时会检查环境变量，如果没有可用的JAVA，会报JAVA_HOME is null错误并且退出安装。
- 如果指定的目录下已经安装有flume，客户端安装时会报flume has been installed并退出安装。

解决办法

步骤1 如果报JAVA_HOME is null错误，需要使用命令：

export JAVA_HOME=*java路径*

设置JAVA_HOME，重新运行安装脚本。

步骤2 如果指定的目录下已经安装有Flume客户端，需要先卸载已经存在的Flume客户端，或指定其他目录安装。

----结束

15.7.3 Flume 客户端无法连接服务端

问题现象

安装Flume客户端并设置avro sink与服务端通信，发现无法连接Flume服务端。

原因分析

1. 服务端配置错误，监听端口启动失败，例如服务端avro source配置了错误的IP，或者已经被占用了的端口。查看Flume运行日志：
```
2016-08-31 17:28:42,092 | ERROR | [lifecycleSupervisor-1-9] | Unable to start
EventDrivenSourceRunner: { source:Avro source avro_source: { bindAddress: 10.120.205.7, port:
21154 } } - Exception follows. | org.apache.flume.lifecycle.LifecycleSupervisor
$MonitorRunnable.run(LifecycleSupervisor.java:253)
java.lang.RuntimeException: org.jboss.netty.channel.ChannelException: Failed to bind to: /
192.168.205.7:21154
```
2. 若采用了加密传输，证书或密码错误。
```
2016-08-31 17:15:59,593 | ERROR | [conf-file-poller-0] | Source avro_source has been removed due to
an error during configuration |
org.apache.flume.node.AbstractConfigurationProvider.loadSources(AbstractConfigurationProvider.java:3
88)
org.apache.flume.FlumeException: Avro source configured with invalid keystore: /opt/Bigdata/
MRS_XXX/install/FusionInsight-Flume-1.9.0/flume/conf/flume_sChat.jks
```
3. 客户端与服务端通信异常。
```
PING 192.168.85.55 (10.120.85.55) 56(84) bytes of data.
From 192.168.85.50 icmp_seq=1 Destination Host Unreachable
From 192.168.85.50 icmp_seq=2 Destination Host Unreachable
From 192.168.85.50 icmp_seq=3 Destination Host Unreachable
From 192.168.85.50 icmp_seq=4 Destination Host Unreachable
```

解决办法

步骤1 设置为正确的IP，必须为本机的IP，如果端口被占用，重新配置一个空闲的端口。

步骤2 配置正确的证书路径。

步骤3 联系网络管理员，恢复网络。

----结束

15.7.4 Flume 数据写入组件失败

问题现象

Flume进程启动后，Flume数据无法写入到对应组件。（以下以服务端写入到HDFS为例）

原因分析

1. HDFS未启动或故障。查看Flume运行日志：
```
2019-02-26 11:16:33,564 | ERROR | [SinkRunner-PollingRunner-DefaultSinkProcessor] | opreation the
hdfs file errors. | org.apache.flume.sink.hdfs.HDFSEventSink.process(HDFSEventSink.java:414)
2019-02-26 11:16:33,747 | WARN | [hdfs-CCCC-call-runner-4] | A failover has occurred since the start
of call #32795 ClientNamenodeProtocolTranslatorPB.getFileInfo over
192-168-13-88/192.168.13.88:25000 | org.apache.hadoop.io.retry.RetryInvocationHandler
$ProxyDescriptor.failover(RetryInvocationHandler.java:220)
2019-02-26 11:16:33,748 | ERROR | [hdfs-CCCC-call-runner-4] | execute hdfs error. {} |
org.apache.flume.sink.hdfs.HDFSEventSink$3.call(HDFSEventSink.java:744)
java.net.ConnectException: Call From 192-168-12-221/192.168.12.221 to 192-168-13-88:25000 failed
on connection exception: java.net.ConnectException: Connection refused; For more details see: http://
wiki.apache.org/hadoop/ConnectionRefused
```
2. hdfs sink未启动。查看Flume运行日志，发现“ flume current metrics”中并没有sink信息：
```
2019-02-26 11:46:05,501 | INFO | [pool-22-thread-1] | flume current metrics:{"CHANNEL.BBBB":
{"ChannelCapacity":"10000","ChannelFillPercentage":"0.0","Type":"CHANNEL","ChannelStoreSize":"0","
EventProcessTimedelta":"0","EventTakeSuccessCount":"0","ChannelSize":"0","EventTakeAttemptCount":
"0","StartTime":"1551152734999","EventPutAttemptCount":"0","EventPutSuccessCount":"0","StopTime
```

":"0"},"SOURCE.AAAA":
{"AppendBatchAcceptedCount":"0","EventAcceptedCount":"0","AppendReceivedCount":"0","MonTime":"0","StartTime":"1551152735503","AppendBatchReceivedCount":"0","EventReceivedCount":"0","Type":"SOURCE","TotalFilesCount":"1001","SizeAcceptedCount":"0","UpdateTime":"605410241202740","AppendAcceptedCount":"0","OpenConnectionCount":"0","MovedFilesCount":"1001","StopTime":"0"}} |
org.apache.flume.node.Application.getRestartComps(Application.java:467)

解决办法

步骤1 若Flume数据写入的组件未启动，启动对应组件；若组件异常，请联系服务技术支持。

步骤2 sink未启动，检查配置文件是否配置正确，若配置错误，则正确修改配置文件后重启Flume进程，如果配置正确，则查看日志错误信息，根据具体错误信息指定解决办法。

----结束

15.7.5 Flume 服务端进程故障

问题现象

Flume运行一段时间后，Manager界面Flume实例显示运行状态“故障”。

原因分析

Flume文件或文件夹权限异常，重启后Manager界面提示如下信息：

```
[2019-02-26 13:38:02]RoleInstance prepare to start failure [{ScriptExecutionResult=ScriptExecutionResult [exitCode=126, output=, errMsg=sh: line 1: /opt/Bigdata/MRS_XXX/install/FusionInsight-Flume-1.9.0/flume/bin/flume-manage.sh: Permission denied
```

解决办法

与运行正常的Flume节点进行文件和文件夹权限对比，更改错误文件或文件夹权限。

15.7.6 Flume 数据采集慢

问题现象

Flume启动后，Flume数据采集慢。

原因分析

1. Flume堆内存设置不合理，导致Flume进程一直处于频繁GC。查看Flume运行日志：
   ```
   2019-02-26T13:06:20.666+0800: 1085673.512: [Full GC:[CMS: 3849339k->3843458K(3853568K), 2.5817610 secs] 4153654K->3843458K(4160256K), [CMS Perm : 27335K->27335K(45592K),2.5820080 SECS] [Times: user=2.63, sys0.00, real=2.59 secs]
   ```
2. 用户业务配置的Spooldir source的deletePolicy策略是立即删除（immediate）。

解决办法

步骤1 适当调大堆内存（xmx）的值。

步骤2 将Spooldir source的deletePolicy策略更改为永不删除（never）。

----结束

15.7.7 Flume 启动失败

问题现象

安装Flume服务或重启Flume服务失败。

原因分析

1. Flume堆内存设置的值大于机器剩余内存，查看Flume启动日志：
   ```
   [CST 2019-02-26 13:31:43][INFO] [[checkMemoryValidity:124]] [GC_OPTS is invalid: Xmx(40960000MB) is bigger than the free memory(56118MB) in system.] [9928]
   ```
2. Flume文件或文件夹权限异常，界面或后台会提示如下信息：
   ```
   [2019-02-26 13:38:02]RoleInstance prepare to start failure
   [{ScriptExecutionResult=ScriptExecutionResult [exitCode=126, output=, errMsg=sh: line 1: /opt/Bigdata/MRS_XXX/install/FusionInsight-Flume-1.9.0/flume/bin/flume-manage.sh: Permission denied
   ```
3. JAVA_HOME配置错误，查看Flume agent启动日志：
   ```
   Info: Sourcing environment configuration script /opt/FlumeClient/fusioninsight-flume-1.9.0/conf/flume-env.sh
   + '[' -n '' ']'
   + exec /tmp/MRS-Client/MRS_Flume_ClientConfig/JDK/jdk-8u18/bin/java '-XX:OnOutOfMemoryError=bash /opt/FlumeClient/fusioninsight-flume-1.9.0/bin/out_memory_error.sh /opt/FlumeClient/fusioninsight-flume-1.9.0/conf %p' -Xms2G -Xmx4G -XX:CMSFullGCsBeforeCompaction=1 -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnabled -XX:+UseCMSCompactAtFullCollection -Dkerberos.domain.name=hadoop.hadoop.com -verbose:gc -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:/var/log/Bigdata//flume-client-1/flume/flume-root-20190226134231-%p-gc.log -Dproc_org.apache.flume.node.Application -Dproc_name=client -Dproc_conf_file=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf/properties.properties -Djava.security.krb5.conf=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf//krb5.conf -Djava.security.auth.login.config=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf//jaas.conf -Dzookeeper.server.principal=zookeeper/hadoop.hadoop.com -Dzookeeper.request.timeout=120000 -Dflume.instance.id=884174180 -Dflume.agent.name=clientName1 -Dflume.role=client -Dlog4j.configuration.watch=true -Dlog4j.configuration=log4j.properties -Dflume_log_dir=/var/log/Bigdata//flume-client-1/flume/ -Dflume.service.id=flume-client-1 -Dbeetle.application.home.path=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf/service -Dflume.called.from.service -Dflume.conf.dir=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf -Dflume.metric.conf.dir=/opt/FlumeClient/fusioninsight-flume-1.9.0/conf -Dflume.script.home=/opt/FlumeClient/fusioninsight-flume-1.9.0/bin -cp '/opt/FlumeClient/fusioninsight-flume-1.9.0/conf:/opt/FlumeClient/fusioninsight-flume-1.9.0/lib/*:/opt/FlumeClient/fusioninsight-flume-1.9.0/conf/service/' -Djava.library.path=/opt/FlumeClient/fusioninsight-flume-1.9.0/plugins.d/native/native org.apache.flume.node.Application --conf-file /opt/FlumeClient/fusioninsight-flume-1.9.0/conf/properties.properties --name client
   /opt/FlumeClient/fusioninsight-flume-1.9.0/bin/flume-ng: line 233: /tmp/FusionInsight-Client/Flume/FusionInsight_Flume_ClientConfig/JDK/jdk-8u18/bin/java: No such file or directory
   ```

解决办法

步骤1 适当调大堆内存（xmx）的值。

步骤2 与正常启动Flume的节点进行文件和文件夹权限对比，更改错误文件或文件夹权限。

步骤3 重新配置JAVA_HOME。客户端替换${install_home}/fusioninsight-flume-*flume组件版本号*/conf/ENV_VARS文件中JAVA_HOME的值，服务端替换etc目录下ENV_VARS文件中JAVA_HOME的值。

其中JAVA_HOME的值可通过登录正常启动Flume的节点，执行**echo ${JAVA_HOME}**获取。

说明

${install_home}为Flume客户端的安装路径。

----结束

15.8 使用 HBase

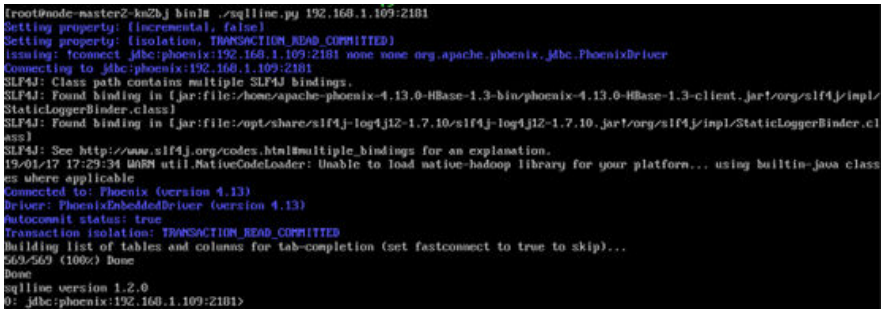
15.8.1 连接到 HBase 响应慢

用户问题

在相同的vpc网络下，外部集群通过Phoenix连接到HBase响应慢。

问题现象

在相同的vpc下，外部集群通过Phoenix连接到HBase时，响应太慢。



原因分析

客户配置了DNS服务，由于客户端连接到HBase先通过DNS来解析服务器端，所以导致响应慢。

处理步骤

步骤1 以root用户登录Master节点。

步骤2 执行**vi /etc/resolv.conf**，打开resolv.conf文件，注释掉DNS服务器地址。例如，#1.1.1.1

----结束

15.8.2 HBase 用户认证失败

用户问题

HBase用户认证失败。

问题现象

客户侧HBase用户认证失败，报错信息如下：

```
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: login failed with hbaseuser and /usr/local/linoseyc/hbase-tomcat/webapps/bigdata_hbase/WEB-INF/classes/user.keytab.
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: perhaps cause 1 is (wrong password) keytab file and user not match, you can kinit -k -t keytab user in client server to check.
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: perhaps cause 2 is (clock skew) time of local server and remote server not match, please check ntp to remote server.
2019-05-13 10:53:09,975 ERROR [localhost-startStop-1] xxxConfig.LoginUtil: perhaps cause 3 is (aes256 not support) aes256 not support by default jdk/jre, need copy local_policy.jar and US_export_policy.jar from remote server in path ${BIGDATA_HOME}/jdk/jre/lib/security.
```

原因分析

客户使用的JDK中的jar包与MRS服务认证的jar包版本不一致。

处理步骤

步骤1 以root登录集群Master1节点。

步骤2 执行如下命令，查看MRS服务认证的jar包。

ll /opt/share/local_policy/local_policy.jar

ll /opt/Bigdata/jdk{version}/jre/lib/security/local_policy.jar

步骤3 将步骤2中的jar包下载到本地。

步骤4 将下载的jar包替换到本地JDK目录/opt/Bigdata/jdk/jre/lib/security。

步骤5 执行**cd /opt/client/HBase/hbase/bin**命令，进入到HBase的bin目录。

步骤6 执行**sh start-hbase.sh**命令，重启HBase组件。

----结束

15.8.3 端口被占用导致 RegionServer 启动失败

问题现象

Manager页面监控发现RegionServer状态为Restoring。

原因分析

1. 通过查看RegionServer日志（/var/log/Bigdata/hbase/rs/hbase-omm-xxx.log）。
2. 使用**lsof -i:21302**（MRS1.7.X及以后端口号是16020）查看到pid，然后根据pid查看到相应的进程，发现RegionServer的端口被DFSZkFailoverController占用。
3. 查看“/proc/sys/net/ipv4/ip_local_port_range”显示为“9000 65500”，临时端口范围与MRS产品端口范围重叠，因为安装时未进行preinstall操作。

解决办法

步骤1 执行**kill -9** *DFSZkFailoverController的pid*， 使得其重启后绑定其它端口，然后重启Restoring的RegionServer。

----结束

15.8.4 节点剩余内存不足导致 HBase 启动失败

问题现象

HBase的RegionServer服务一直是Restoring状态。

原因分析

1. 查看RegionServer的日志（“/var/log/Bigdata/hbase/rs/hbase-omm-XXX.out”），发现显示以下打印信息：
   There is insufficient memory for the Java Runtime Environment to continue.
2. 使用**free**指令查看，该节点确实没有足够内存。

解决办法

步骤1 现场进行排查内存不足原因，确认是否有某些进程占用过多内存，或者由于服务器自身内存不足。

----结束

15.8.5 HDFS 性能差导致 HBase 服务不可用告警

问题现象

HBase组件断断续续上报服务不可用告警。

原因分析

该问题多半为HDFS性能较慢，导致健康检查超时，从而导致监控告警。可通过以下方式判断：

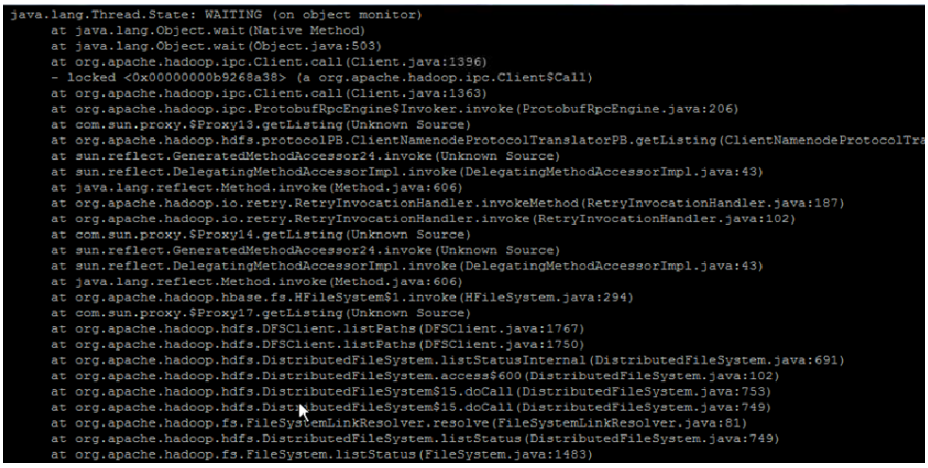
1. 首先查看HMaster日志（“/var/log/Bigdata/hbase/hm/hbase-omm-xxx.log”），确认HMaster日志中没有频繁打印“system pause”或“jvm”等GC相关信息。
2. 然后可以通过下列三种方式确认原因为HDFS性能慢造成告警产生。
   a. 使用客户端验证，通过**hbase shell**进入hbase命令行后，执行**list**验证需要运行多久。
   b. 开启HDFS的debug日志，然后查看下层目录很多的路径（**hadoop fs -ls /XXX/XXX**），验证需要运行多久。
   c. 打印HMaster进程jstack：

      su - omm

      jps

      jstack pid
3. 如下图所示，Jstack显示一直卡在DFSClient.listPaths。

图 15-19 异常



解决办法

步骤1 如果确认是HDFS性能慢导致告警，需要排除是否为旧版本中Impala运行导致HDFS性能慢或者是否为集群最初部署时JournalNode部署不正确（部署过多，大于3个）。

----结束

15.8.6 参数不合理导致 HBase 启动失败

问题现象

修改部分参数后，无法正常启动HBase。

原因分析

1. 查看HMaster日志（/var/log/Bigdata/hbase/hm/hbase-omm-xxx.log）显示，hbase.regionserver.global.memstore.size + hfile.block.cache.size总和大于0.8导致启动不成功，因此需要调整参数配置值总和低于0.8。

2. 查看HMaster和RegionServer的out日志（/var/log/Bigdata/hbase/hm/hbase-omm-xxx.out/var/log/Bigdata/hbase/rs/hbase-omm-xxx.out），提示Unrecognized VM option。

```
Unrecognized VM option
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. Program will exit.
```

检查GC_OPTS相关参数存在多余空格，如-D sun.rmi.dgc.server.gcInterval=0x7FFFFFFFFFFFFFE。

解决办法

步骤1 针对memstore、cache修改配置参数后，重启HBase服务成功。

步骤2 针对GC_OPTS配置错误，修改参数后重启HBase服务成功。

----结束

15.8.7 残留进程导致 Regionsever 启动失败

问题现象

HBase服务启动失败，健康检查报错。

原因分析

查看启动HBase服务时manager页面的详细打印信息，提示the previous process is not quit。

解决办法

步骤1 登录节点，后台通过执行**ps -ef | grep HRegionServer**发现确实存在一个残留的进程。

步骤2 确认进程可以终止后，使用kill命令终止该进程（如果kill无法终止该进程，需要使用kill -9来强制终止该进程）。

步骤3 重新启动HBase服务成功。

----结束

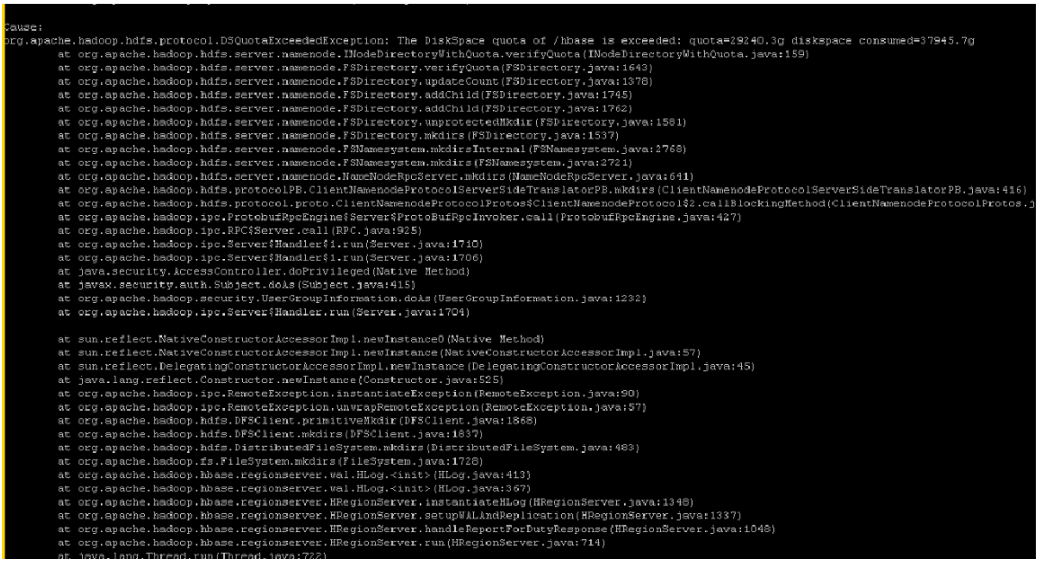
15.8.8 HDFS 上设置配额导致 HBase 启动失败

问题现象

HBase启动失败。

原因分析

查看HMaster日志信息（“/var/log/Bigdata/hbase/hm/hbase-omm-xxx.log”），出现如下异常，The DiskSpace quota of /hbase is exceeded。



解决办法

步骤1 通过后台使用**df -h**命令查看数据盘目录空间已满，因此需要删除无用的数据来进行应急恢复。

步骤2 后续需要扩容节点来解决数据目录空间不足问题。

----结束

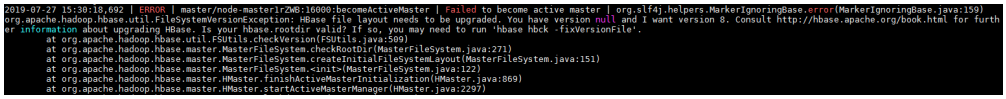
15.8.9 HBase version 文件损坏导致启动失败

问题背景

HBase启动失败。

原因分析

1. HBase启动时会读取hbase.version文件，但是日志显示读取存在异常。



2. 通过**hadoop fs -cat /hbase/hbase.version**命令发现文件不能正常查看，该文件损坏。

解决办法

步骤1 执行**hbase hbck -fixVersionFile**命令修复文件。

步骤2 如**步骤1**不能解决，从同版本的其他集群中获取hbase.version文件上传进行替换。

步骤3 重新启动HBase服务。

----结束

15.8.10 无业务情况下，RegionServer 占用 CPU 高

问题背景

无业务情况下，RegionServer占用CPU较高。

原因分析

1. 通过**top**命令获取RegionServer的进程使用CPU情况信息，查看CPU使用率高的进程号。
2. 根据RegionServer的进程编号，获取该进程下线程使用CPU情况。

   top -H -p <PID>（根据实际RegionServer的进程ID进行替换），具体如下图所示，发现部分线程CPU使用率均达到80%。

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
75706	omm	20	0	6879444	1.0g	25612	S	90.4	1.6	0:00.00	java
75716	omm	20	0	6879444	1.0g	25612	S	90.4	1.6	0:04.74	java
75720	omm	20	0	6879444	1.0g	25612	S	88.6	1.6	0:01.93	java
75721	omm	20	0	6879444	1.0g	25612	S	86.8	1.6	0:01.99	java
75722	omm	20	0	6879444	1.0g	25612	S	86.8	1.6	0:01.94	java
75723	omm	20	0	6879444	1.0g	25612	S	86.8	1.6	0:01.96	java
75724	omm	20	0	6879444	1.0g	25612	S	86.8	1.6	0:01.97	java

```
75725 omm      20   0 6879444   1.0g  25612 S  81.5  1.6   0:02.06 java
75726 omm      20   0 6879444   1.0g  25612 S  79.7  1.6   0:02.01 java
75727 omm      20   0 6879444   1.0g  25612 S  79.7  1.6   0:01.95 java
75728 omm      20   0 6879444   1.0g  25612 S  78.0  1.6   0:01.99 java
```

3. 根据RegionServer的进程编号，获取线程堆栈信息。

   jstack 12345 >allstack.txt（根据实际RegionServer的进程ID进行替换）

4. 将需要的线程ID转换为16进制格式：

   printf "%x\n" 30648

   输出结果TID为77b8。

5. 根据输出16进制TID，在线程堆栈中进行查找，发现在执行compaction操作。

```
"regionserver/ahbd-hbase-dat1/12.2.168.1:21302-longCompactions-1482676601478" #1641 prio=5 os_prio=0 tid=0x00007fa614563000 nid=0x77b8 runnable [0x0
   java.lang.Thread.State: RUNNABLE
    at org.apache.hadoop.io.compress.snappy.SnappyCompressor.compressBytesDirect(Native Method)
    at org.apache.hadoop.io.compress.snappy.SnappyCompressor.compress(SnappyCompressor.java:228)
    at org.apache.hadoop.io.compress.BlockCompressorStream.compress(BlockCompressorStream.java:149)
    at org.apache.hadoop.io.compress.BlockCompressorStream.finish(BlockCompressorStream.java:142)
    at org.apache.hadoop.hbase.io.encoding.HFileBlockDefaultEncodingContext.compressAfterEncoding(HFileBlockDefaultEncodingContext.java:219)
    at org.apache.hadoop.hbase.io.encoding.HFileBlockDefaultEncodingContext.compressAndEncrypt(HFileBlockDefaultEncodingContext.java:132)
    at org.apache.hadoop.hbase.io.hfile.HFileBlock$Writer.finishBlock(HFileBlock.java:989)
    at org.apache.hadoop.hbase.io.hfile.HFileBlock$Writer.ensureBlockReady(HFileBlock.java:961)
    at org.apache.hadoop.hbase.io.hfile.HFileBlock$Writer.finishBlockAndWriteHeaderAndData(HFileBlock.java:1077)
```

6. 对其它线程执行相同操作，发现均为compactions线程。

```
"regionserver/ahbd-hbase-dat1/12.2.168.1:21302-longCompactions-1482676601473" #1629 prio=5 os_prio=0 tid=0x00007fa61454d800 nid=0x77a0 runnable
   java.lang.Thread.State: RUNNABLE
    at org.apache.hadoop.hdfs.DFSOutputStream.writeChunk(DFSOutputStream.java:425)
    - locked <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSOutputSummer.writeChecksumChunks(FSOutputSummer.java:214)
    at org.apache.hadoop.fs.FSOutputSummer.flushBuffer(FSOutputSummer.java:165)
    - locked <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSOutputSummer.flushBuffer(FSOutputSummer.java:146)
    - eliminated <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSOutputSummer.write1(FSOutputSummer.java:137)
    at org.apache.hadoop.fs.FSOutputSummer.write(FSOutputSummer.java:112)
    - locked <0x000000020276ba38> (a org.apache.hadoop.hdfs.DFSOutputStream)
    at org.apache.hadoop.fs.FSDataOutputStream$PositionCache.write(FSDataOutputStream.java:58)
    at java.io.DataOutputStream.write(DataOutputStream.java:107)
    - locked <0x00000004de9535c8> (a org.apache.hadoop.hdfs.client.HdfsDataOutputStream)
    at java.io.FilterOutputStream.write(FilterOutputStream.java:97)
```

解决办法

属于正常现象。

发现消耗CPU较高线程均为HBase的compaction，其中部分线程调用Snappy压缩处理，部分线程调用HDFS读写数据。当前每个Region数据量和数据文件多，且采用Snappy压缩算法，因此执行compaction时会使用大量CPU导致CPU较高。

定位办法

步骤1 使用**top**命令查看 CPU使用率高的进程号。

步骤2 查看此进程中占用CPU高的线程。

使用命令**top -H -p <PID>**即可打印出某进程<PID>下的线程的CPU耗时信息。

一般某个进程如果出现问题，是因为某个线程出现问题了，获取查询到的占用CPU最高的线程号。

或者使用命令**ps -mp <PID> -o THREAD,tid,time | sort -rn**。

观察回显可以得到CPU最高的线程号。

步骤3 获取出现问题的线程的堆栈。

java问题使用jstack工具是最有效，最可靠的。

到java/bin目录下有 jstack工具，获取进程堆栈，并输出到本地文件。

jstack <PID> > allstack.txt

获取线程堆栈，并输出到本地文件。

步骤4 将需要的线程ID转换为16进制格式。

printf "%x\n" <PID>

回显结果为线程ID，即 TID。

步骤5 使用命令获得TID,并输出到本地文件。

jstack <PID> | grep <TID> > Onestack.txt

如果只是在命令行窗口查看，可以使用命名：

jstack <PID> | grep <TID> -A 30

-A 30意思是显示30行。

----结束

15.8.11 HBase 启动失败，RegionServer 日志中提示 FileNotFoundException 异常

问题背景

HBase启动失败，RegionServer一直处于Restoring状态。

原因分析

1. 查看RegionServer的日志（/var/log/Bigdata/hbase/rs/hbase-omm-XXX.log），发现显示以下打印信息：

```
| ERROR | RS_OPEN_REGION-ab-dn01:21302-2 | ABORTING region server ab-
dn01,21302,1487663269375: The coprocessor
org.apache.kylin.storage.hbase.cube.v2.coprocessor.endpoint.CubeVisitService threw
java.io.FileNotFoundException: File does not exist: hdfs://hacluster/kylin/kylin_metadata/coprocessor/
kylin-coprocessor-1.6.0-SNAPSHOT-0.jar |
org.apache.hadoop.hbase.regionserver.HRegionServer.abort(HRegionServer.java:2123)
java.io.FileNotFoundException: File does not exist: hdfs://hacluster/kylin/kylin_metadata/coprocessor/
kylin-coprocessor-1.6.0-SNAPSHOT-0.jar
at org.apache.hadoop.hdfs.DistributedFileSystem$25.doCall(DistributedFileSystem.java:1399)
at org.apache.hadoop.hdfs.DistributedFileSystem$25.doCall(DistributedFileSystem.java:1391)
at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
at org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(DistributedFileSystem.java:1391)
at org.apache.hadoop.fs.FileUtil.copy(FileUtil.java:340)
at org.apache.hadoop.fs.FileUtil.copy(FileUtil.java:292)
at org.apache.hadoop.fs.FileSystem.copyToLocalFile(FileSystem.java:2038)
at org.apache.hadoop.fs.FileSystem.copyToLocalFile(FileSystem.java:2007)
at org.apache.hadoop.fs.FileSystem.copyToLocalFile(FileSystem.java:1983)
at org.apache.hadoop.hbase.util.CoprocessorClassLoader.init(CoprocessorClassLoader.java:168)
at
org.apache.hadoop.hbase.util.CoprocessorClassLoader.getClassLoader(CoprocessorClassLoader.java:250)
at org.apache.hadoop.hbase.coprocessor.CoprocessorHost.load(CoprocessorHost.java:224)
at
org.apache.hadoop.hbase.regionserver.RegionCoprocessorHost.loadTableCoprocessors(RegionCoprocess
orHost.java:365)
at
org.apache.hadoop.hbase.regionserver.RegionCoprocessorHost.<init>(RegionCoprocessorHost.java:227)
at org.apache.hadoop.hbase.regionserver.HRegion.<init>(HRegion.java:783)
at org.apache.hadoop.hbase.regionserver.HRegion.<init>(HRegion.java:689)
at sun.reflect.GeneratedConstructorAccessor22.newInstance(Unknown Source)
at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:4
5)
```

```
at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
at org.apache.hadoop.hbase.regionserver.HRegion.newHRegion(HRegion.java:6312)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6622)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6594)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6550)
at org.apache.hadoop.hbase.regionserver.HRegion.openHRegion(HRegion.java:6501)
at
org.apache.hadoop.hbase.regionserver.handler.OpenRegionHandler.openRegion(OpenRegionHandler.ja
va:363)
at
org.apache.hadoop.hbase.regionserver.handler.OpenRegionHandler.process(OpenRegionHandler.java:12
9)
at org.apache.hadoop.hbase.executor.EventHandler.run(EventHandler.java:129)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
```

2. 使用客户端hdfs命令查看，如下文件不存在。

   hdfs://hacluster/kylin/kylin_metadata/coprocessor/kylin-coprocessor-1.6.0-SNAPSHOT-0.jar

3. HBase在配置协处理器时，一定要保证对应的jar包路径没有问题，否则HBase会无法启动。

解决办法

使用Kylin对接MRS，确保Kylin相关jar包存在。

15.8.12 HBase 启动后原生页面显示 RegionServer 个数多于实际个数

问题背景

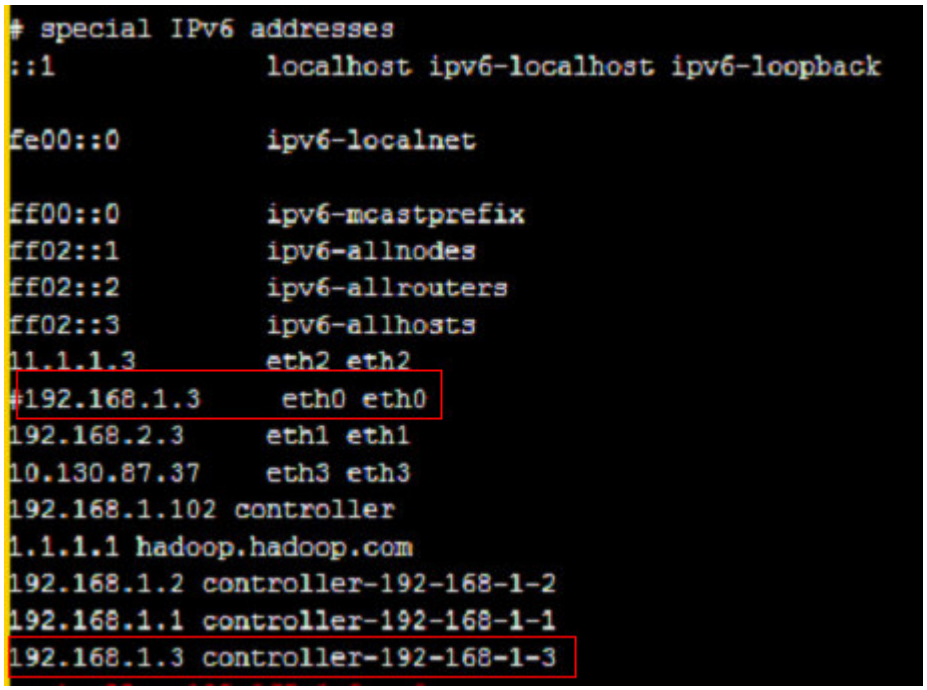
HBase启动后，HMaster原生页面显示RegionServer个数多于实际RegionServer个数。

查看HMaster原生页面，显示有4个RegionServer在线，如下图示：

ServerName	Start time	Requests Per Second	Num. Regions
controller-192-168-1-1,21302,1494933959261	Tue May 16 19:25:59 CST 2017	0	19
controller-192-168-1-2,21302,1494933957536	Tue May 16 19:25:57 CST 2017	0	24
controller-192-168-1-3,21302,1494933958592	Tue May 16 19:25:58 CST 2017	0	16
eth0,21302,1494933958592	Tue May 16 19:25:58 CST 2017	0	0
Total:4		0	59

原因分析

如下图可以看出，第三行hostname为controller-192-168-1-3节点和第四行hostname为eth0节点为同一RegionServer上报的信息，登录相应节点，查看/etc/hosts文件，发现，对应同一ip，配置两个hostname。如下：



解决办法

登录RegionServer所在节点，修改/etc/hosts文件，同一ip只能对应同一hostname。

15.8.13 RegionServer 实例异常，处于 Restoring 状态

问题背景

HBase启动失败，RegionServer一直处于Restoring状态。

原因分析

查看异常的RegionServer实例的运行日志（/var/log/Bigdata/hbase/rs/hbase-omm-XXX.log），发现显示以下打印信息ClockOutOfSyncException...，Reported time is too far out of sync with master

```
2017-09-18 11:16:23,636 | FATAL | regionserver21302 | Master rejected startup because clock is out of sync |
org.apache.hadoop.hbase.regionserver.HRegionServer.reportForDuty(HRegionServer.java:2059)
org.apache.hadoop.hbase.ClockOutOfSyncException: org.apache.hadoop.hbase.ClockOutOfSyncException:
Server nl-bi-fi-datanode-24-65,21302,1505726180086 has been rejected; Reported time is too far out of
sync with master.  Time difference of 152109ms > max allowed of 30000ms
at org.apache.hadoop.hbase.master.ServerManager.checkClockSkew(ServerManager.java:354)
...

...
2017-09-18 11:16:23,858 | ERROR | main | Region server exiting |
org.apache.hadoop.hbase.regionserver.HRegionServerCommandLine.start(HRegionServerCommandLine.java:
70)
java.lang.RuntimeException: HRegionServer Aborted
```

该日志说明异常的RegionServer实例和HMaster实例的时差大于允许的时差值30s（由参数hbase.regionserver.maxclockskew控制，默认30000ms），导致RegionServer实例异常。

解决办法

调整异常节点时间，确保节点间时差小于30s。

15.8.14 新安装的集群 HBase 启动失败

问题背景

新安装的集群HBase启动失败，查看RegionServer日志报如下错误：

```
2018-02-24 16:53:03,863 | ERROR | regionserver/host3/187.6.71.69:21302 | Master passed us a different hostname to use; was=host3, but now=187-6-71-69 | org.apache.hadoop.hbase.regionserver.HRegionServer.handleReportForDutyResponse(HRegionServer.java:1386)
```

原因分析

/etc/hosts中同一个IP地址配置了多个主机名映射关系。

解决办法

步骤1 修改/etc/host中IP与主机名的映射关系，配置正确。

步骤2 重新启动HBase组件。

----结束

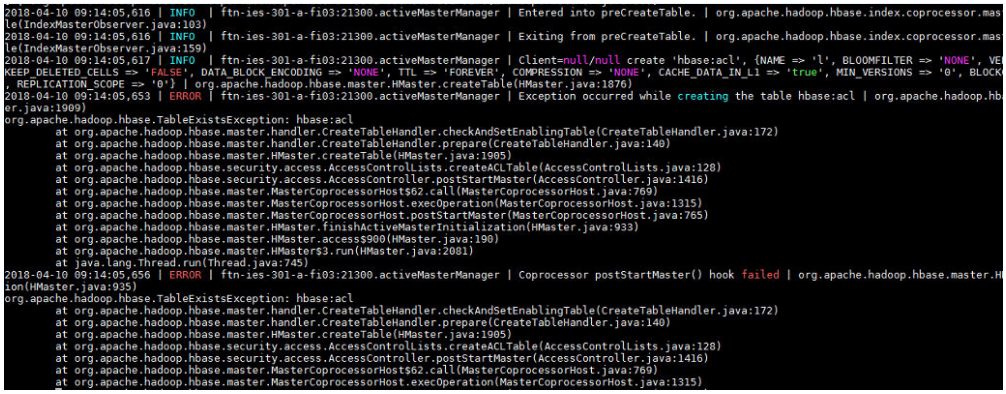
15.8.15 acl 表目录丢失导致 HBase 启动失败

问题背景与现象

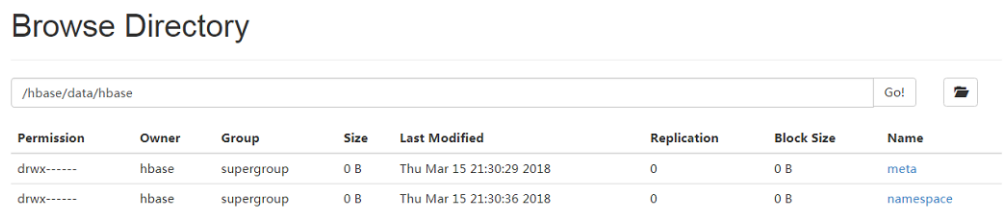
集群HBase启动失败

原因分析

1. 查看HBase的HMaster日志，报如下错误：



2. 检查HDFS上HBase的路径发现acl表路径丢失。



解决办法

步骤1 停止HBase组件。

步骤2 在HBase客户端使用hbase用户登录认证，执行如下命令。

例如：

```
hadoop03:~ # source /opt/client/bigdata_env
hadoop03:~ # kinit hbase
Password for hbase@HADOOP.COM:
hadoop03:~ # hbase zkcli
```

步骤3 删除zk中acl表信息。

例如：

```
[zk: hadoop01:24002,hadoop02:24002,hadoop03:24002(CONNECTED) 0] deleteall /hbase/table/hbase:acl
[zk: hadoop01:24002,hadoop02:24002,hadoop03:24002(CONNECTED) 0] deleteall /hbase/table-lock/
hbase:acl
```

步骤4 启动HBase组件。

----结束

15.8.16 集群上下电之后 HBase 启动失败

问题背景与现象

集群的ECS关机重启后，HBase启动失败。

原因分析

查看HMaster的运行日志，发现有报大量的如下错误：

```
2018-03-26 11:10:54,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:00,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:06,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:10,787 | INFO  | RpcServer.reader=9,bindAddress=hadoopc1h3,port=21300 | Kerberos
principal name is hbase/hadoop.hadoop.com@HADOOP.COM | org.apache.hadoop.hbase
.ipc.RpcServer$Connection.readPreamble(RpcServer.java:1532)
2018-03-26 11:11:12,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
```

```
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
2018-03-26 11:11:18,185 | INFO  | hadoopc1h3,21300,1522031630949_splitLogManager__ChoreService_1 |
total tasks = 1 unassigned = 0 tasks={/hbase/splitWAL/WALs%2Fhadoopc1h1%2C213
02%2C1520214023667-splitting
%2Fhadoopc1h1%252C21302%252C1520214023667.default.1520584926990=last_update =
1522033841041 last_version = 34255 cur_worker_name = hadoopc1h3,21302,
1520943011826 status = in_progress incarnation = 3 resubmits = 3 batch = installed = 1 done = 0 error = 0}
| org.apache.hadoop.hbase.master.SplitLogManager$TimeoutMonitor.chore
(SplitLogManager.java:745)
```

节点上下电，RegionServer的wal分裂失败导致。

解决办法

步骤1 停止HBase组件。

步骤2 通过hdfs fsck命令检查/hbase/WALs文件的健康状态。

hdfs fsck /hbase/WALs

输出如下表示文件都正常，如果有异常则需要先处理异常的文件，再执行后面的操作。

```
The filesystem under path '/hbase/WALs' is HEALTHY
```

步骤3 备份/hbase/WALs文件。

hdfs dfs -mv /hbase/WALs /hbase/WALs_old

步骤4 新建/hbase/WALs目录。

hdfs dfs -mkdir /hbase/WALs

必须保证路径权限是hbase:hadoop。

步骤5 启动HBase组件。

----结束

15.8.17 文件块过大导致 HBase 数据导入失败

问题现象

导入数据到hbase报错：NotServingRegionException。

原因分析

当一个block size大于2G时，hdfs在seek的时候会出现读取异常，持续频繁写入regionserver时出现了full gc，且时间比较久，导致hmaster与regionserver之间的心跳异常，然后hmaster把regionserver标记为dead状态，强制重启了Regionserver，重启后触发servercrash机制开始回滚wal日志。现在这个splitwal的文件已经达到将近2.1G，且其仅有一个block块，导致hdfs seek异常，引起splitwal失败，regionserver检测到当前这个wal日志还需要split，又会触发splitwal日志的机制进行回滚，就这样在split与split失败之间不停循环，导致无法上线该regionserver节点上的region，最后出现查询该RS上某一个region时会报region not online的异常。

处理步骤

步骤1 进入HBase服务页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > HBase”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > HBase”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
- MRS 3.x及后续版本，登录FusionInsight Manager。然后选择“集群 > *待操作的集群名*称 > 服务 > HBase”。

步骤2 在“HMaster Web UI”右侧，单击“HMaster (主)”进入HBase Web UI界面。

步骤3 在“Procedures”页签查看问题节点。

步骤4 以root用户登录问题节并执行**hdfs dfs -ls**命令查看所有块信息。

步骤5 执行**hdfs dfs -mkdir**命令新建目录用于存放问题块。

步骤6 执行**hdfs dfs -mv**将问题块转移至新建目录位置。

----结束

建议与总结

以下两点可供参考：

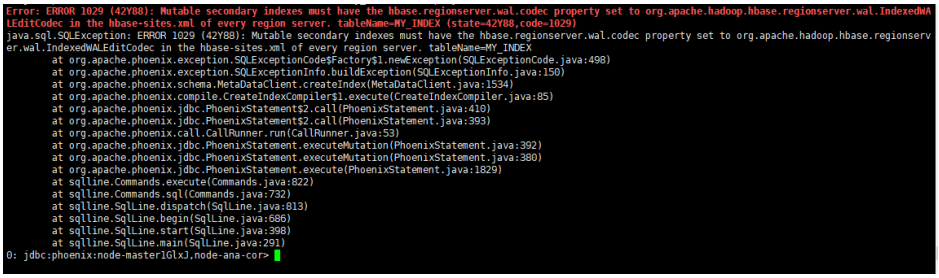
- 数据块损坏，通过**hdfs fsck /tmp -files -blocks -racks**命令检查block数据块健康信息。
- region正在分裂时对数据的操作会抛NotServingRegionException异常。

15.8.18 使用 Phoenix 创建 HBase 表后，向索引表中加载数据报错

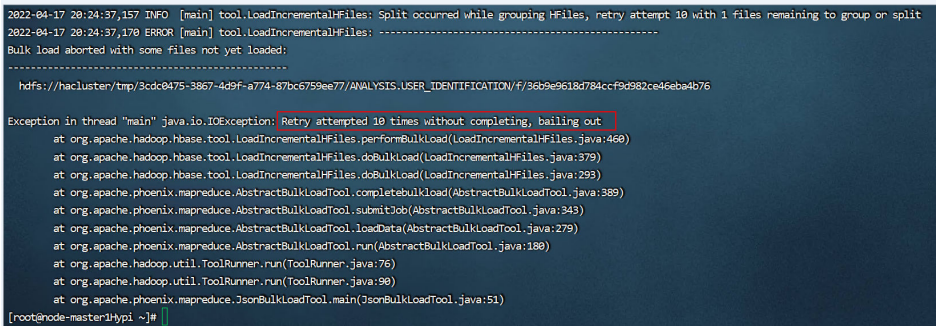
问题背景与现象

使用Phoenix创建HBase表后，使用命令向索引表中加载数据报错：

- MRS 2.x及之前版本：Mutable secondary indexes must have the hbase.regionserver.wal.codec property set to org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec in the hbase-sites.xml of every region server. tableName=MY_INDEX (state=42Y88,code=1029)



- MRS 3.x及之后版本：Exception in thread "main" java.io.IOException: Retry attempted 10 times without completing, bailing out



处理步骤

步骤1 MRS 2.x及之前版本，操作步骤如下：

1. 使用**admin**用户登录MRS Manager界面，选择“服务管理 > HBase > 服务配置”，将“参数类别”的“基础配置”切换为“全部配置”，选择“HMaster > 自定义”，给参数“hbase.hmaster.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项。
2. 选择“RegionServer > 自定义”，给参数“hbase.regionserver.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项，单击“保存配置”，输入当前用户密码，单击“确定”，保存配置。
3. 单击“服务状态”，选择“更多 > 重启服务”，输入当前用户密码，单击“确定”，重启HBase服务。

步骤2 MRS 3.x及之后版本，操作步骤如下：

1. 使用**admin**用户登录FusionInsight Manager，选择“集群 > 服务 > HBase > 配置 > 全部配置 > RegionServer > 自定义”，给参数“hbase.regionserver.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项。
2. 选择“HMaster > 自定义”，给参数“hbase.hmaster.config.expandor”新增名称为“hbase.regionserver.wal.codec”，值为“org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec”的配置项。
3. 单击“保存”，在弹出的对话框中单击“确定”，保存配置。
4. 单击“概览”，选择“更多 > 重启服务”，输入当前用户密码，单击“确定”，重启HBase服务。

----结束

15.8.19 在 MRS 集群客户端无法执行 hbase shell 命令

用户问题

在MRS集群客户端无法执行**hbase shell**命令。

原因分析

- 执行**hbase shell**命令前未配置环境变量。
- 当前MRS集群未安装HBase客户端。

处理步骤

步骤1 使用**root**用户登录安装客户端的节点，切换到客户端安装目录，查看是否安装了HBase客户端。

- 是，执行**步骤2**。
- 否，下载并安装HBase客户端。

步骤2 执行以下命令配置环境变量。

source bigdata_env

步骤3 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

步骤4 执行HBase组件的客户端命令。

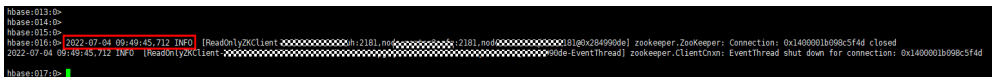
hbase shell

----结束

15.8.20 HBase shell 客户端在使用中有 INFO 信息打印在控制台导致显示混乱

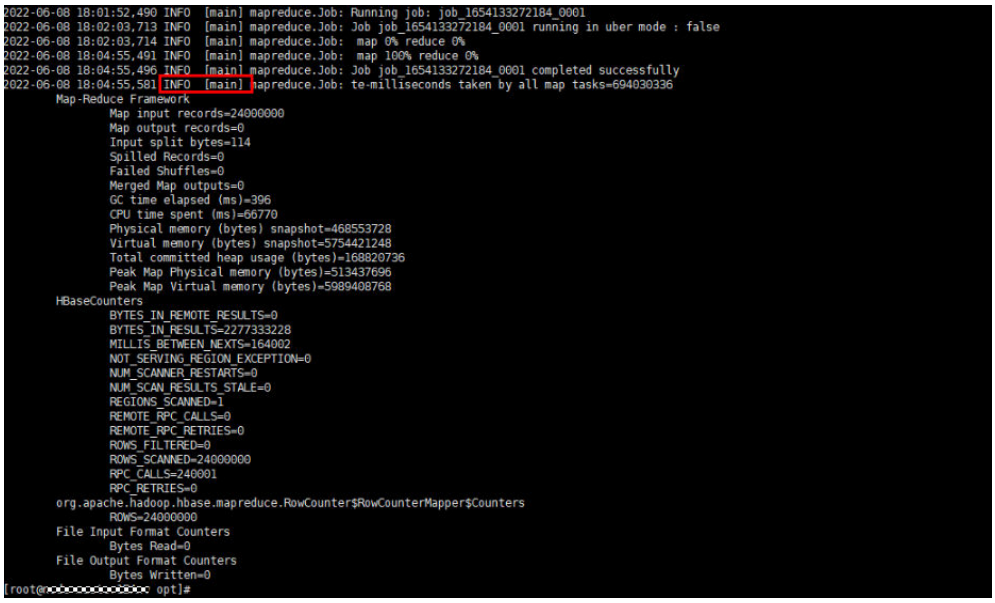
用户问题

在使用HBase shell中会有INFO信息打印在控制台，导致显示混乱，用户在输入命令前需要按回车符清理控制台。如下图所示：



原因分析

- HBase客户端中默认日志打印设置为“INFO,console”，所以在使用期间会有INFO日志输出到控制台，影响HBase shell 窗口的显示。
- HBase客户端命令繁多，例如：**hbase shell**、**hbase hbck**、**hbase org.apache.hadoop.hbase.mapreduce.RowCounter**等，且后续还会增加。部分命令的输出为INFO打印，如果直接把INFO关闭会导致部分命令输出结果丢失。例如：RowCounter输出结果为INFO类型：



处理步骤

步骤1 使用**root**用户登录安装HBase客户端的节点。

步骤2 在“*HBase客户端安装目录*/HBase/component_env”文件中添加如下信息：

export HBASE_ROOT_LOGGER=INFO,RFA

把日志输出到日志文件中，后期如果使用**hbase org.apache.hadoop.hbase.mapreduce.RowCounter**等命令，执行结果请在日志文件“*HBase客户端安装目录*/HBase/hbase/logs/hbase.log”中查看。

步骤3 切换到HBase客户端安装目录，执行以下命令使配置生效。

cd *HBase客户端安装目录*

source HBase/component_env

----结束

15.8.21 RegionServer 剩余内存不足导致 HBase 服务启动失败

用户问题

RegionServer剩余内存不足导致HBase服务启动失败。

原因分析

RegionServer启动时节点剩余内存不足，导致无法启动实例。排查步骤如下：

1. 登录Master节点，到“/var/log/Bigdata”查找HBase相关日志，HMaster的日志中报错“connect regionserver timeout ”。
2. 登录到1中HMaster连接不上的RegionServer节点，到“/var/log/Bigdata”查找HBase相关日志，RegionServer报错“error=’Cannot allocate memory’(errno=12)”。

3. 根据2报错判断由于RegionServer内存不足导致RegionServer启动失败。

处理步骤

步骤1 登录报错的RegionServer节点，执行以下命令查看节点剩余内存：

free -g

步骤2 执行**top**命令查看节点内存使用情况。

步骤3 根据**top**提示结束内存占用多的进程（内存占用多并且非MRS自身组件的进程），并重新启动HBase服务。

说明

集群的Core节点除了MRS组件运行占用外，Yarn上的作业还会被分配到节点运行，占用节点内存。若是由于Yarn作业占用内存多导致组件无法正常启动时，建议扩容Core节点。

----结束

15.9 使用 HDFS

15.9.1 修改集群 HDFS 服务的 NameNode RPC 端口后，NameNode 都变为备状态

用户问题

通过页面更改NameNode的RPC端口，随后重启HDFS服务，出现所有NameNode一直是备状态，导致集群异常。

问题现象

所有NameNode都是备状态，导致集群异常。

原因分析

集群安装启动后，如果修改NameNode的RPC端口，则需要重新格式化Zkfc服务来更新zookeeper上的节点信息。

处理步骤

步骤1 登录Manager，停止HDFS服务。

说明

在停止HDFS时，建议不要停止相关服务。

步骤2 停止成功后，登录到被修改了RPC端口的Master节点。

说明

如果两个Master节点都被修改了RPC端口，则只需登录其中一个修改即可。

步骤3 执行**su - omm**命令切换到omm用户。

说明

如果是安全集群，需要执行**kinit hdfs**命令进行认证。

步骤4 执行如下命令，将环境变量脚本加载到环境中。

cd ${BIGDATA_HOME}/*MRS_X.X.X*/*1_8*_**Zkfc/etc**

source ${BIGDATA_HOME}/*MRS_X.X.X***/install/FusionInsight-Hadoop-3.1.1/hadoop/sbin/exportENV_VARS.sh**

说明

命令中的“*MRS_X.X.X*”和“1_8”根据实际版本而定。

步骤5 加载完成后，执行如下命令，格式化Zkfc。

cd ${HADOOP_HOME}/bin

./hdfs zkfc -formatZK

步骤6 格式化成功后，在Manager页面“重启”HDFS服务。

说明

如果更改了NameNode的RPC端口，则之前安装的所有客户端都需要刷新配置文件。

----结束

15.9.2 通过公网 IP 连接主机，使用 HDFS 客户端报错

用户问题

通过公网IP连接主机，不能使用HDFS客户端，运行HDFS提示**-bash: hdfs: command not found**。

问题现象

通过公网IP连接主机，不能使用HDFS客户端，运行HDFS提示**-bash: hdfs: command not found**。

原因分析

用户登录Master节点执行命令之前，未设置环境变量。

处理步骤

步骤1 以**root**用户登录任意一个Master节点。

步骤2 执行**source /opt/client/bigdata_env**命令，设置环境变量。

步骤3 执行**hdfs**命令即可成功使用HDFS客户端。

----结束

15.9.3 使用 Python 远程连接 HDFS 的端口失败

用户问题

使用Python远程连接HDFS的端口失败，如何解决?

问题现象

用户使用Python远程连接HDFS的50070端口失败。

原因分析

HDFS开源3.0.0以下版本的默认端口为50070，3.0.0及以上的默认端口为9870。用户使用的端口和HDFS版本不匹配导致连接端口失败。

步骤1 登录集群的主Master节点。

步骤2 执行**su - omm**命令，切换到omm用户。

步骤3 执行**/opt/Bigdata/om-0.0.1/sbin/queryVersion.sh**命令，查看集群中的HDFS版本号。

根据版本号确认开源组件的端口号。

步骤4 执行**netstat -anp|grep ${port}**命令，查看组件的默认端口号是否存在。

如果不存在，说明用户修改了默认的端口号。请修改为默认端口，再重新连接HDFS。

如果存在，请联系技术服务。

说明

- ${ port }：表示与组件版本相对应的组件默认端口号。
- 如果用户修改了默认端口号，请使用修改后的端口号连接HDFS。不建议修改默认端口号。

----结束

15.9.4 HDFS 容量使用达到 100%，导致上层服务 HBase、Spark 等上报服务不可用

用户问题

集群的HDFS容量使用达到100%，HDFS服务状态为只读，导致上层服务HBase、Spark等上报服务不可用告警。

问题现象

HDFS使用容量100%，磁盘容量只使用85%左右，HDFS服务状态为只读，导致上层服务HBase、Spark等上报服务不可用。

原因分析

当前NodeManager和DataNode共数据盘使用，MRS默认预留15%的数据磁盘空间给非HDFS使用，可通过HDFS参数**dfs.datanode.du.reserved.percentage**修改百分比来控制具体的磁盘占比。

当HDFS磁盘使用100%之后，可通过降低**dfs.datanode.du.reserved.percentage**百分比来恢复业务，再进行磁盘扩容。

处理步骤

步骤1 登录集群任意Master节点。

步骤2 执行**source /opt/client/bigdata_env**命令，初始化环境变量。

说明

如果是安全集群，则需要执行**kinit -kt <keytab file> <pricipal name>** 进行认证。

步骤3 执行**hdfs dfs –put ./startDetail.log /tmp**命令，测试HDFS写文件失败。

```
19/05/12 10:07:32 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /tmp/startDetail.log._COPYING_ could only be replicated to 0 nodes instead of minReplication (=1).  There are 3 datanode(s) running and no node(s) are excluded in this operation.
```

步骤4 执行**hdfs dfsadmin -report**命令，检查HDFS使用容量，发现已经达到100%。

```
Configured Capacity: 5389790579100 (4.90 TB)
Present Capacity: 5067618628404 (4.61 TB)
DFS Remaining: 133350196 (127.17 MB)
DFS Used: 5067485278208 (4.61 TB)
DFS Used%: 100.00%
Under replicated blocks: 10
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0
Pending deletion blocks: 0
```

步骤5 当HDFS使用容量达到100%时，通过HDFS参数**dfs.datanode.du.reserved.percentage**修改百分比来控制具体的磁盘占比。

1. 登录Manager进入服务配置页面。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > HDFS> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 >HDFS > 配置”。
2. 选择“全部配置”，在搜索框中搜索**dfs.datanode.du.reserved.percentage**。
3. 修改此参数的取值为“10”。

步骤6 修改完成后，扩容Core节点的磁盘个数。

----结束

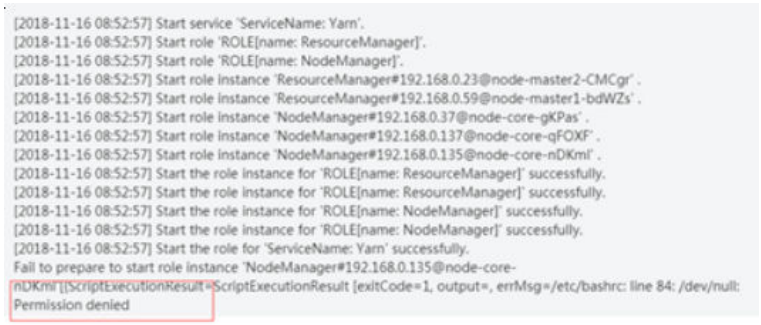
15.9.5 启动 HDFS 和 Yarn 报错

用户问题

启动HDFS和Yarn时报错。

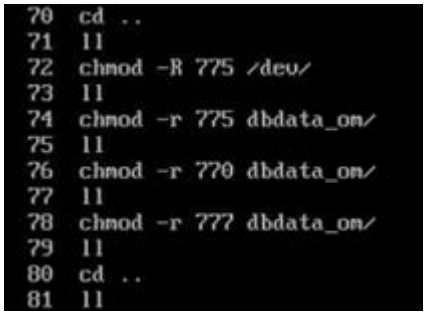
问题现象

无法启动HDFS、Yarn服务组件，报错内容：/dev/null Permission denied。



原因分析

客户修改了虚机系统的/dev/null的权限值为775。



处理步骤

步骤1 以root用户登录集群的任意一个Master节点。

步骤2 登录成功后，执行**chmod 666 /dev/null**命令，修改/dev/null的权限值为666。

步骤3 执行**ls -al /dev/null**命令，查看修改的/dev/null权限值是否为666，如果不是，需要修改为666。

步骤4 修改成功后，重新启动HDFS和Yarn组件。

----结束

15.9.6 HDFS 权限设置问题

用户问题

在使用MRS服务的时候，某个用户可以在其他用户的HDFS目录下面删除或者创建文件。

问题现象

在使用MRS服务时，某个用户可以在其他用户的HDFS目录下面删除或者创建文件。

原因分析

客户配置的用户具有ficommon组的权限，所以可以对HDFS任意操作。需要移除用户的ficommon组权限。

处理步骤

步骤1 以root用户登录集群的Master节点。

步骤2 执行**id ${用户名}**命令，显示用户组信息，确认是否有ficommon组权限。

如果存在ficommon组权限，请执行**步骤3**。如果不存在，请联系技术服务。

说明

${用户名}：出现HDFS权限设置问题的用户名。

步骤3 执行**gpasswd -d ${用户名} ficommon**命令，删除该用户的ficommon组权限。

说明

${用户名}：出现HDFS权限设置问题的用户名。

步骤4 执行成功后，登录Manager修改参数。

MRS Manager界面操作（适用MRS 3.x之前版本）：

1. 登录MRS Manager，在MRS Manager页面，选择“服务管理 > HDFS > 服务配置”。
2. “参数类别”选择“全部配置”，在搜索框中输入“dfs.permissions.enabled”，修改为“true”。
3. 修改完成后，单击“保存配置”，并重启HDFS服务。

FusionInsight Manager界面操作（适用MRS 3.x及之后版本）：

1. 登录FusionInsight Manager。选择“集群 > 服务 > HDFS > 配置 > 全部配置”。
2. 在搜索框中输入“dfs.permissions.enabled”，修改为“true”。
3. 修改完成后，单击“保存”，并重启HDFS服务。

MRS集群详情页操作（适用MRS 2.0.1及之后版本）：

1. 登录MRS控制台，选择“组件管理 > HDFS > 服务配置”。
2. “参数类别”选择“全部配置”，在搜索框中输入“dfs.permissions.enabled”，修改为“true”。
3. 修改完成后，单击“保存配置”，并重启HDFS服务。

----结束

15.9.7 HDFS 的 DataNode 一直显示退服中

用户问题

HDFS的DataNode一直显示退服中。

问题现象

HDFS的某个DataNode退服（或者对Core节点进行缩容）任务失败，但是DataNode在任务失败后一直处于退服中的状态。

原因分析

在对HDFS的某个DataNode进行退服（或者对core节点进行缩容）过程中，因为Master节点重启或者nodeagent进程意外退出等情况出现，使得退服（或缩容）任务

失败，并且没有进行黑名单清理。此时DataNode节点会一直处于退服中的状态，需要人工介入进行黑名单清理。

处理步骤

步骤1 进入服务实例界面。

MRS Manager界面操作：

登录MRS Manager，在MRS Manager页面，选择“服务管理 > HDFS > 实例”。

FusionInsight Manager界面操作：

对于MRS 3.x及后续版本集群：也可登录FusionInsight Manager。选择“集群 > 服务 > HDFS > 实例”。

也可登录MRS控制台，选择“组件管理 > HDFS > 实例”。

对于MRS 2.0.1及之后的版本集群：也可登录MRS控制台，选择“组件管理 > HDFS > 实例”。

步骤2 查看HDFS服务实例状态，找到一直处于退服中的DataNode，复制这个DataNode的IP地址。

步骤3 登录Master1节点的后台，执行**cd ${BIGDATA_HOME}/MRS_*/1_*_NameNode/etc/**命令进入黑名单目录。

步骤4 执行**sed -i "/^*IP*$/d" excludeHosts**命令清理黑名单中的故障DataNode信息，该命令中IP替换为**步骤2**中查询到的故障DataNode的IP地址，其中不能有空格。

步骤5 如果有两个Master节点，请在Master2节点上同样执行**步骤3**和**步骤4**。

步骤6 在Master1节点执行如下命令初始化环境变量。

source /opt/client/bigdata_env

步骤7 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则无需执行此命令。

kinit *MRS集群用户*

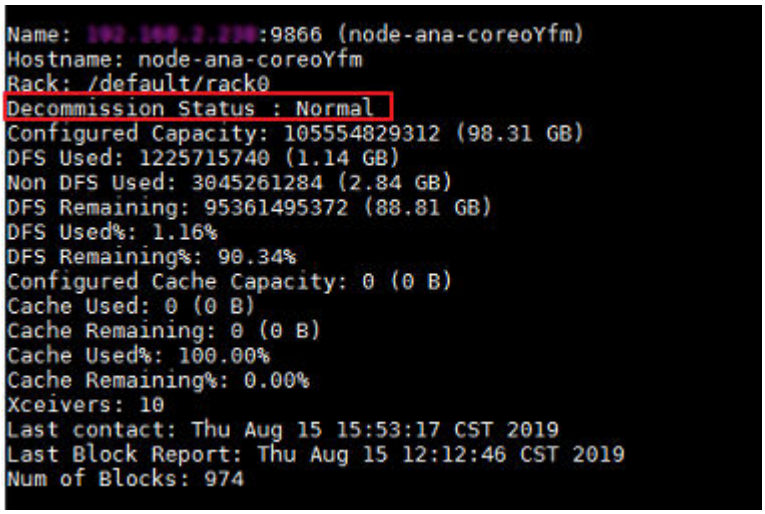
例如, **kinit admin**

步骤8 在Master1节点执行如下命令刷新HDFS的黑名单。

hdfs dfsadmin -refreshNodes

步骤9 使用命令**hdfs dfsadmin -report**来查看各个DataNode的状态，确认中查到的IP对应的DataNode已经恢复为**Normal**状态。

图 15-20 DataNode 的状态



步骤10 进入服务实例界面。

MRS Manager界面操作：

登录MRS Manager，在MRS Manager页面，选择“服务管理 > HDFS > 实例”。

FusionInsight Manager界面操作：

对于MRS 3.x及后续版本集群：可登录FusionInsight Manager。选择“集群 > 服务 > HDFS > 实例”。

对于MRS 2.0.1及之后的版本集群：登录MRS控制台，选择“组件管理 > HDFS > 实例”。

步骤11 勾选一直处于退服中的DataNode实例，单击“更多 > 重启实例”。

步骤12 等待重启完成，确认DataNode是否恢复正常。

----结束

建议与总结

尽量不要在退服（或缩容）过程中重启节点等高危操作。

参考信息

无

15.9.8 内存不足导致 HDFS 启动失败

问题背景与现象

重启HDFS后，HDFS的状态是Bad，且NameNode实例状态常常异常，并且花很久没有退出安全模式。

原因分析

1. 在NameNode运行日志（/var/log/Bigdata/hdfs/nn/hadoop-omm-namendoe-XXX.log）中搜索“WARN”，可以看到有大量时间在垃圾回收，如下例中耗时较长63s。

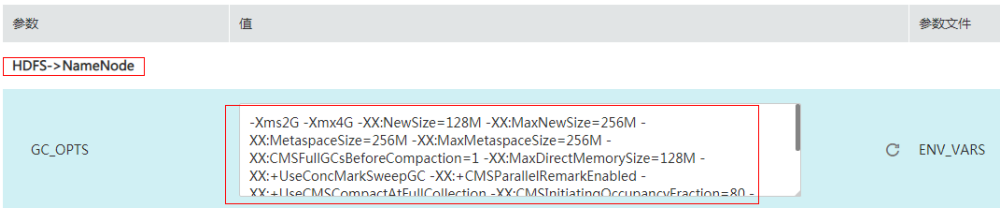
```
2017-01-22 14:52:32,641 | WARN  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@1b39fd82 |
Detected pause in JVM or host machine (eg GC): pause of approximately 63750ms
GC pool 'ParNew' had collection(s): count=1 time=0ms
GC pool 'ConcurrentMarkSweep' had collection(s): count=1 time=63924ms | JvmPauseMonitor.java:189
```

2. 分析NameNode日志“/var/log/Bigdata/hdfs/nn/hadoop-omm-namendoe-XXX.log”，可以看到NameNode在等待块上报，且总的Block个数过多，如下例中是3629万。

```
2017-01-22 14:52:32,641 | INFO  | IPC Server handler 8 on 25000 | STATE* Safe mode ON.
The reported blocks 29715437 needs additional 6542184 blocks to reach the threshold 0.9990 of total
blocks 36293915.
```

3. 打开Manager页面，查看NameNode的GC_OPTS参数配置如下：

图 15-21 查看 NameNode 的 GC_OPTS 参数配置



4. NameNode内存配置和数据量对应关系参考**表15-2**。

表 15-2 NameNode 内存配置和数据量对应关系

文件对象数量	参考值
10,000,000	“-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M”
20,000,000	“-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G”
50,000,000	“-Xms32G -Xmx32G -XX:NewSize=2G -XX:MaxNewSize=3G”
100,000,000	“-Xms64G -Xmx64G -XX:NewSize=4G -XX:MaxNewSize=6G”
200,000,000	“-Xms96G -Xmx96G -XX:NewSize=8G -XX:MaxNewSize=9G”
300,000,000	“-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G”

解决办法

步骤1 按照规格修改NameNode的内存参数，如这里3600万block，将内存参数调整为“-Xms32G -Xmx32G -XX:NewSize=2G -XX:MaxNewSize=3G”。

步骤2 重启一个NameNode，确认该NameNode可以正常启动。

步骤3 重启另一个NameNode，确认页面状态恢复。

----结束

15.9.9 ntpdate 修改时间导致 HDFS 出现大量丢块

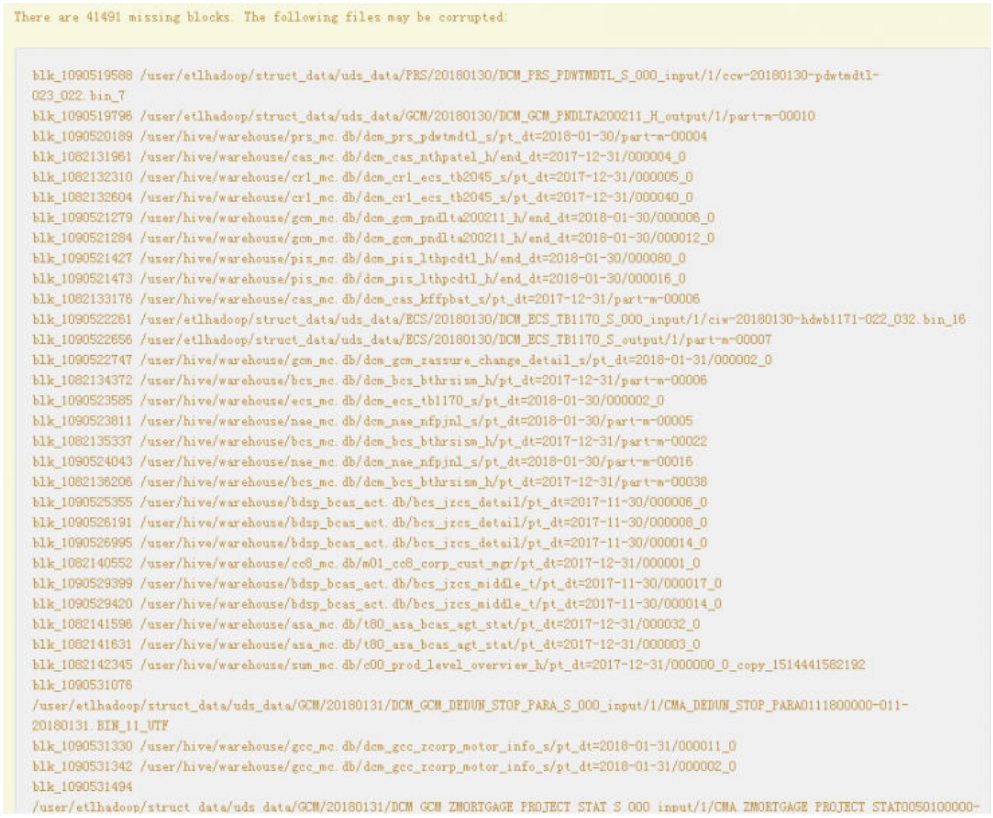
问题背景与现象

1. 用ntpdate修改了集群时间，修改时未停止集群，修改后HDFS进入安全模式，无法启动。
2. 退出安全模式后启动，hfck检查丢了大概1T数据。

原因分析

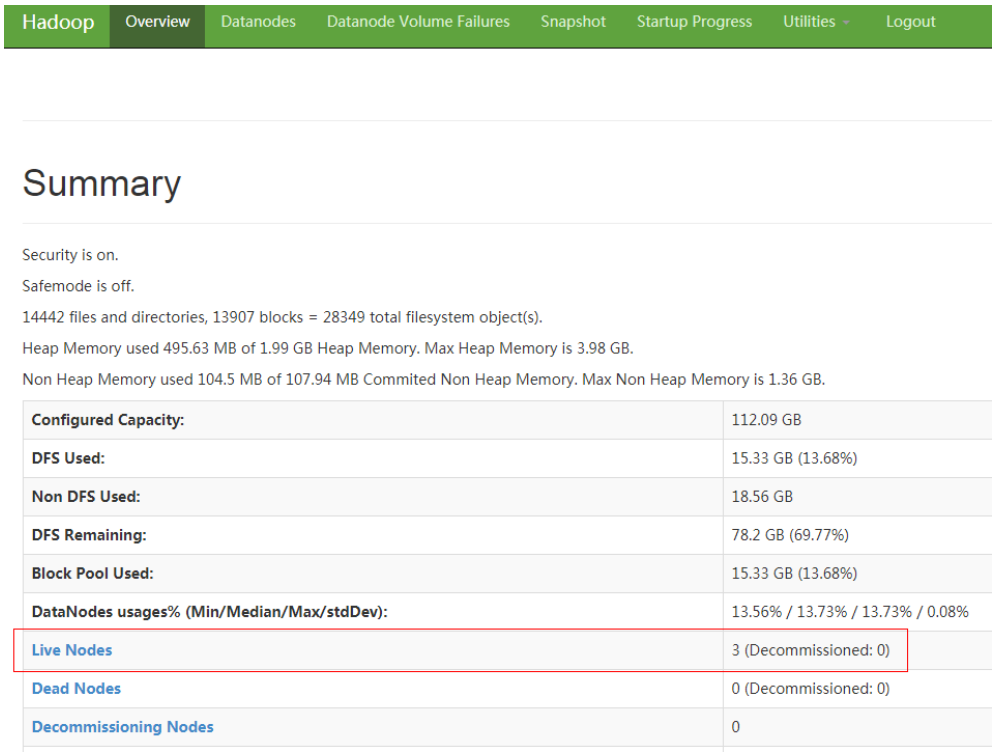
1. 查看NameNode原生页面发现有大量的块丢失。

图 15-22 块丢失



2. 查看原生页面 Datanode Information 发现显示的DataNode节点数和实际的相差10个节点。

图 15-23 查看 DataNode 节点数



3. 查看DateNode运行日志“/var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-主机名.log”，发现如下错误信息。

   重要错误信息 Clock skew too great

图 15-24 DateNode 运行日志错误

```
    at org.apache.hadoop.ipc.Client.call(Client.java:1486)
    at org.apache.hadoop.ipc.Client.call(Client.java:1447)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:229)
    at com.sun.proxy.$Proxy14.versionRequest(Unknown Source)
    at org.apache.hadoop.hdfs.protocolPB.DatanodeProtocolClientSideTranslatorPB.versionRequest(DatanodeProtocolClientSideTranslatorPB.java:273)
    at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.retrieveNamespaceInfo(BPServiceActor.java:187)
    at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.connectToNNAndHandshake(BPServiceActor.java:237)
    at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.run(BPServiceActor.java:689)
    at java.lang.Thread.run(Thread.java:745)
Caused by: GSSException: No valid credentials provided (Mechanism level: Clock skew too great (37))
    at sun.security.jgss.krb5.Krb5Context.initSecContext(Krb5Context.java:770)
    at sun.security.jgss.GSSContextImpl.initSecContext(GSSContextImpl.java:248)
    at sun.security.jgss.GSSContextImpl.initSecContext(GSSContextImpl.java:179)
    at com.sun.security.sasl.gsskerb.GssKrb5Client.evaluateChallenge(GssKrb5Client.java:192)
    ... 20 more
Caused by: KrbException: Clock skew too great (37)
    at sun.security.krb5.KrbKdcRep.check(KrbKdcRep.java:88)
    at sun.security.krb5.KrbTgsRep.<init>(KrbTgsRep.java:87)
    at sun.security.krb5.KrbTgsReq.getReply(KrbTgsReq.java:259)
    at sun.security.krb5.KrbTgsReq.sendAndGetCreds(KrbTgsReq.java:270)
    at sun.security.krb5.internal.CredentialsUtil.serviceCreds(CredentialsUtil.java:302)
    at sun.security.krb5.internal.CredentialsUtil.acquireServiceCreds(CredentialsUtil.java:120)
    at sun.security.krb5.Credentials.acquireServiceCreds(Credentials.java:458)
    at sun.security.jgss.krb5.Krb5Context.initSecContext(Krb5Context.java:693)
```

解决办法

步骤1 修改在原生页面查看不到的10个数据节点的时间。

步骤2 在Manager页面重启对应的DataNode实例。

----结束

15.9.10 DataNode 概率性出现 CPU 占用接近 100%，导致节点丢失（ssh 连得很慢或者连不上）

问题背景与现象

DataNode概率性出现CPU占用接近100%，导致节点丢失。

图 15-25 DataNode 出现 CPU 占用接近 100%

原因分析

1. DataNode有许多写失败的日志。

   图 15-26 DataNode 写失败的日志

   ```
   2015-08-31 11:29:34,184 | ERROR | DataXceiver for client DFSClient_NONMAPREDUCE_1675952887_23 at /192.168.8.40:44514 [Receiving block
   BP-125271511-192.168.8.29-1440656260530:blk_1074766997_1034914] | TSP21:25009:DataXceiver error processing WRITE_BLOCK operation  src:
   /192.168.8.40:44514 dst: /192.168.8.64:25009 | DataXceiver.java:258
   java.io.IOException: Premature EOF from inputStream
           at org.apache.hadoop.io.IOUtils.readFully(IOUtils.java:194)
           at org.apache.hadoop.hdfs.protocol.datatransfer.PacketReceiver.doReadFully(PacketReceiver.java:213)
           at org.apache.hadoop.hdfs.protocol.datatransfer.PacketReceiver.doRead(PacketReceiver.java:134)
           at org.apache.hadoop.hdfs.protocol.datatransfer.PacketReceiver.receiveNextPacket(PacketReceiver.java:109)
           at org.apache.hadoop.hdfs.server.datanode.BlockReceiver.receivePacket(BlockReceiver.java:446)
           at org.apache.hadoop.hdfs.server.datanode.BlockReceiver.receiveBlock(BlockReceiver.java:707)
           at org.apache.hadoop.hdfs.server.datanode.DataXceiver.writeBlock(DataXceiver.java:748)
           at org.apache.hadoop.hdfs.protocol.datatransfer.Receiver.opWriteBlock(Receiver.java:124)
           at org.apache.hadoop.hdfs.protocol.datatransfer.Receiver.processOp(Receiver.java:71)
           at org.apache.hadoop.hdfs.server.datanode.DataXceiver.run(DataXceiver.java:240)
           at java.lang.Thread.run(Thread.java:745)
   2015-08-31 11:29:35,147 | INFO  | DataXceiver for client DFSClient_NONMAPREDUCE_-402997805_1 at /192.168.8.30:59449 [Sending block BP-
   125271511-192.168.8.29-1440656260530:blk_1074181856_446655] | src: /192.168.8.64:25009, dest: /192.168.8.30:59449, bytes: 16826, op:
   HDFS_READ, cliID: DFSClient_NONMAPREDUCE_-402997805_1, offset: 0, srvID: 9d1d30a5-046d-438b-83c9-2c6c54c6bd12, blockid: BP-125271511-
   192.168.8.29-1440656260530:blk_1074181856_446655, duration: 78832 | BlockSender.java:738
   2015-08-31 11:29:35,269 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg
   GC): pause of approximately 7480ms
   No GCs detected | JvmPauseMonitor.java:172
   2015-08-31 11:29:36,985 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg
   GC): pause of approximately 1215ms
   No GCs detected | JvmPauseMonitor.java:172
   2015-08-31 11:29:43,067 | INFO  | DataXceiver for client DFSClient_NONMAPREDUCE_1675952887_23 at /192.168.8.33:35530 [Receiving block
   BP-125271511-192.168.8.29-1440656260530:blk_1074767006_1034923] | Exception for BP-125271511-192.168.8.29-
   1440656260530:blk_1074767006_1034923 | BlockReceiver.java:742
   java.io.IOException: Premature EOF from inputStream
   ```

2. 短时间内写入大量文件导致这种情况，因此DataNode内存不足。

   图 15-27 写入大量文件导致 DataNode 内存不足

   ```
   Line 153101: 2015-08-31 11:24:29,313 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 1199ms
   Line 153132: 2015-08-31 11:24:42,689 | WARN  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 11273ms
   Line 153135: 2015-08-31 11:24:45,810 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 1005ms
   Line 153138: 2015-08-31 11:24:49,801 | INFO  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 1067ms
   Line 153155: 2015-08-31 11:25:10,167 | WARN  | org.apache.hadoop.util.JvmPauseMonitor$Monitor@551bd2a0 | Detected pause in JVM or host machine (eg GC): pause of approximately 12323ms
   ```

解决办法

步骤1 检查DataNode内存配置，以及机器剩余内存是否充足。

步骤2 增加DataNode内存，并重启DataNode。

----结束

15.9.11 单 NameNode 长期故障，如何使用客户端手动 checkpoint

问题背景与现象

在备NameNode长期异常的情况下，会积攒大量的editlog，此时如果重启HDFS或者主NameNode，主NameNode会读取大量的未合并的editlog，导致耗时启动较长，甚至启动失败。

原因分析

备NameNode会周期性做合并editlog，生成fsimage文件的过程叫做checkpoint。备NameNode在新生成fsimage后，会将fsimage传递到主NameNode。

说明

由于“备NameNode会周期性做合并editlog”，因此当备NameNode异常时，无法合并editlog，因此主NameNode在下次启动的时候，需要加载较多editlog，需要大量内存，并且耗时较长。

合并元数据的周期由以下参数确定，即如果NameNode运行30分钟或者HDFS操作100万次，均会执行checkpoint。

- dfs.namenode.checkpoint.period：checkpoint周期，默认1800s。
- dfs.namenode.checkpoint.txns：执行指定操作次数后执行checkpoint，默认1000000。

解决办法

在重启前，主动执行异常checkpoint合并主NameNode的元数据。

步骤1 停止业务。

步骤2 获取主NameNode的主机名。

步骤3 在客户端执行如下命令：

source /opt/client/bigdata_env

kinit *组件用户*

说明：/opt/client 需要换为实际客户端的安装路径。

步骤4 执行如下命令，让主NameNode进入安全模式，其中linux22换为主NameNode的主机名。

hdfs dfsadmin -fs linux22:25000 -safemode enter

```
linux16:/opt/fi_client #  hdfs dfsadmin -fs linux22:25000   -safemode enter
17/04/26 18:38:30 WARN fs.FileSystem: "linux22:25000" is a deprecated filesystem name. Use "hdfs://linux22:25000/" instead.
17/04/26 18:38:32 INFO hdfs.PeerCache: SocketCache disabled.
Safe mode is ON
```

步骤5 执行如下命令，在主NameNode，合并editlog。

hdfs dfsadmin -fs linux22:25000 -saveNamespace

```
linux16:/opt/fi_client #  hdfs dfsadmin -fs linux22:25000   -saveNamespace
17/04/26 18:38:54 WARN fs.FileSystem: "linux22:25000" is a deprecated filesystem name. Use "hdfs://linux22:25000/" instead.
17/04/26 18:38:56 INFO hdfs.PeerCache: SocketCache disabled.
Save namespace successful
```

步骤6 执行如下命令，让主NameNode离开安全模式。

hdfs dfsadmin -fs linux22:25000 -safemode leave

```
linux16:/opt/fi_client #  hdfs dfsadmin -fs linux22:25000  -safemode leave
17/04/26 18:39:07 WARN fs.FileSystem: "linux22:25000" is a deprecated filesystem name. Use "hdfs://linux22:25000/" instead.
17/04/26 18:39:09 INFO hdfs.PeerCache: SocketCache disabled.
Safe mode is OFF
```

步骤7 检查是否真的合并完成。

cd /srv/BigData/namenode/current

检查先产生的fsimage是否是当前时间的，若是则表示已经合并完成

```
-rw-------  1 omm wheel 1048576 Apr 26 18:43 edits_inprogress_0000000000002083018
-rw-------  1 omm wheel  736657 Apr 26 15:46 fsimage_0000000000002071390
-rw-------  1 omm wheel      62 Apr 26 15:46 fsimage_0000000000002071390.md5
-rw-------  1 omm wheel  736657 Apr 26 16:46 fsimage_0000000000002075405
-rw-------  1 omm wheel      62 Apr 26 16:46 fsimage_0000000000002075405.md5
-rw-------  1 omm wheel  736410 Apr 26 17:46 fsimage_0000000000002079398
-rw-------  1 omm wheel      62 Apr 26 17:46 fsimage_0000000000002079398.md5
-rw-------  1 omm wheel       8 Apr 26 18:42 seen_txid
linux-20:/srv/BigData/namenode/current #
linux-20:/srv/BigData/namenode/current #
```

----结束

15.9.12 文件读写常见故障

问题背景与现象

当用户在HDFS上执行写操作时，出现“Failed to place enough replicas:expected…”信息。

原因分析

- DataNode的数据接受器不可用。

  此时DataNode会有如下日志：

  ```
  2016-03-17 18:51:44,721 | WARN |
  org.apache.hadoop.hdfs.server.datanode.DataXceiverServer@5386659f |
  hadoopc1h2:25009:DataXceiverServer: | DataXceiverServer.java:158
  java.io.IOException: Xceiver count 4097 exceeds the limit of concurrent xcievers: 4096
  at org.apache.hadoop.hdfs.server.datanode.DataXceiverServer.run(DataXceiverServer.java:140)
  at java.lang.Thread.run(Thread.java:745)
  ```

- DataNode的磁盘空间不足。
- DataNode的心跳有延迟。

解决办法

- 如果DataNode的数据接收器不可用，通过在Manager页面，增加HDFS参数“dfs.datanode.max.transfer.threads”的值解决。
- 如果没有足够的硬盘空间或者CPU，试着增加新的数据节点或确保资源是可用的（磁盘空间或CPU）。
- 如果网络问题，确保网络是可用的。

15.9.13 文件最大打开句柄数设置太小导致读写文件异常

问题背景与现象

文件最大打开句柄数设置太小，导致文件句柄不足。写文件到HDFS很慢，或者写文件失败。

原因分析

1. DataNode日志“/var/log/Bigdata/hdfs/dn/hadoop-omm-datanode-XXX.log”，存在异常提示java.io.IOException: Too many open files。

```
2016-05-19 17:18:59,126 | WARN  |
org.apache.hadoop.hdfs.server.datanode.DataXceiverServer@142ff9fa |
YSDN12:25009:DataXceiverServer:  |
org.apache.hadoop.hdfs.server.datanode.DataXceiverServer.run(DataXceiverServer.java:160)
java.io.IOException: Too many open files
at sun.nio.ch.ServerSocketChannelImpl.accept0(Native Method)
at sun.nio.ch.ServerSocketChannelImpl.accept(ServerSocketChannelImpl.java:241)
at sun.nio.ch.ServerSocketAdaptor.accept(ServerSocketAdaptor.java:100)
at org.apache.hadoop.hdfs.net.TcpPeerServer.accept(TcpPeerServer.java:134)
at org.apache.hadoop.hdfs.server.datanode.DataXceiverServer.run(DataXceiverServer.java:137)
at java.lang.Thread.run(Thread.java:745)
```

2. 如果某个DataNode日志中打印“Too many open files”，说明该节点文件句柄不足，导致打开文件句柄失败，然后就会重试往其他DataNode节点写数据，最终表现为写文件很慢或者写文件失败。

解决办法

步骤1 执行**ulimit -a**命令查看有问题节点文件句柄数最多设置是多少，如果很小，建议修改成640000。

图 15-28 查看文件句柄数

```
[omm@189-39-150-167 ~]$ ulimit -a
core file size          (blocks, -c) 0
data seg size           (kbytes, -d) unlimited
scheduling priority             (-e) 0
file size               (blocks, -f) unlimited
pending signals                 (-i) 256551
max locked memory       (kbytes, -l) 64
max memory size         (kbytes, -m) unlimited
open files                      (-n) 640000
pipe size            (512 bytes, -p) 8
POSIX message queues     (bytes, -q) 819200
real-time priority              (-r) 0
stack size              (kbytes, -s) 10240
cpu time               (seconds, -t) unlimited
max user processes              (-u) 60000
virtual memory          (kbytes, -v) unlimited
file locks                      (-x) unlimited
```

步骤2 执行**vi /etc/security/limits.d/90-nofile.conf**命令编辑这个文件，修改文件句柄数设置。如果没有这个文件，可以新建一个文件，并按照下图内容修改。

图 15-29 修改文件句柄数

```
*          hard    nofile    640000
*          soft    nofile    640000
~
```

步骤3 重新打开一个终端窗口，用**ulimit -a**命令查看是否修改成功，如果没有，请重新按照上述步骤重新修改。

步骤4 从Manager页面重启DataNode实例。

----结束

15.9.14 客户端写文件 close 失败

问题背景与现象

客户端写文件close失败，客户端提示数据块没有足够副本数。

客户端日志：

```
2015-05-27 19:00:52.811 [pool-2-thread-3] ERROR: /tsp/nedata/collect/UGW/ugwufdr/
20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp143272380633
8 close hdfs sequence file fail (SequenceFileInfoChannel.java:444)
java.io.IOException: Unable to close file because the last block does not have enough number of replicas.
at org.apache.hadoop.hdfs.DFSOutputStream.completeFile(DFSOutputStream.java:2160)
at org.apache.hadoop.hdfs.DFSOutputStream.close(DFSOutputStream.java:2128)
at org.apache.hadoop.fs.FSDataOutputStream$PositionCache.close(FSDataOutputStream.java:70)
at org.apache.hadoop.fs.FSDataOutputStream.close(FSDataOutputStream.java:103)
at com.xxx.pai.collect2.stream.SequenceFileInfoChannel.close(SequenceFileInfoChannel.java:433)
at com.xxx.pai.collect2.stream.SequenceFileWriterToolChannel
$FileCloseTask.call(SequenceFileWriterToolChannel.java:804)
at com.xxx.pai.collect2.stream.SequenceFileWriterToolChannel
$FileCloseTask.call(SequenceFileWriterToolChannel.java:792)
at java.util.concurrent.FutureTask.run(FutureTask.java:262)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
at java.lang.Thread.run(Thread.java:745)
```

原因分析

1. HDFS客户端开始写Block。

   例如：HDFS客户端是在2015-05-27 18:50:24,232开始写/20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp1432723806338的。其中分配的块是blk_1099105501_25370893。

   ```
   2015-05-27 18:50:24,232 | INFO  | IPC Server handler 30 on 25000 | BLOCK* allocateBlock: /
   20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp1432723
   806338. BP-1803470917-192.168.57.33-1428597734132
   blk_1099105501_25370893{blockUCState=UNDER_CONSTRUCTION, primaryNodeIndex=-1,
   replicas=[ReplicaUnderConstruction[[DISK]DS-b2d7b7d0-f410-4958-8eba-6deecbca2f87:NORMAL|
   RBW], ReplicaUnderConstruction[[DISK]DS-76bd80e7-ad58-49c6-bf2c-03f91caf750f:NORMAL|RBW]]}
   |
   org.apache.hadoop.hdfs.server.namenode.FSNamesystem.saveAllocatedBlock(FSNamesystem.java:3166
   )
   ```

2. 写完之后HDFS客户端调用了fsync。

   ```
   2015-05-27 19:00:22,717 | INFO  | IPC Server handler 22 on 25000 | BLOCK* fsync:
   20150527/10/6_20150527105000_20150527105500_SR5S14_1432723806338_128_11.pkg.tmp1432723
   806338 for DFSClient_NONMAPREDUCE_-120525246_15 |
   org.apache.hadoop.hdfs.server.namenode.FSNamesystem.fsync(FSNamesystem.java:3805)
   ```

3. HDFS客户端调用close关闭文件，NameNode收到客户端的close请求之后就会检查最后一个块的完成状态，只有当有足够的DataNode上报了块完成才可用关闭文件，检查块完成的状态是通过checkFileProgress函数检查的，打印如下：

   ```
   2015-05-27 19:00:27,603 | INFO  | IPC Server handler 44 on 25000 | BLOCK* checkFileProgress:
   blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
   replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
   RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
   ```

```
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:28,005 | INFO  | IPC Server handler 45 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:28,806 | INFO  | IPC Server handler 63 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:30,408 | INFO  | IPC Server handler 43 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:33,610 | INFO  | IPC Server handler 37 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
2015-05-27 19:00:40,011 | INFO  | IPC Server handler 37 on 25000 | BLOCK* checkFileProgress:
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
has not reached minimal replication 1 |
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkFileProgress(FSNamesystem.java:3197)
```

4. NameNode打印了多次checkFileProgress是由于HDFS客户端多次尝试close文件，但是由于当前状态不满足要求，导致close失败， HDFS客户端retry的次数是由参数dfs.client.block.write.locateFollowingBlock.retries决定的，该参数默认是5，所以在NameNode的日志中看到了6次checkFileProgress打印。
5. 但是再过0.5s之后，DataNode就上报块已经成功写入。

```
2015-05-27 19:00:40,608 | INFO  | IPC Server handler 60 on 25000 | BLOCK* addStoredBlock:
blockMap updated: 192.168.10.21:25009 is added to
blk_1099105501_25370893{blockUCState=COMMITTED, primaryNodeIndex=-1,
replicas=[ReplicaUnderConstruction[[DISK]DS-ef5fd3c9-5088-4813-ae9a-34a0714ec3a3:NORMAL|
RBW], ReplicaUnderConstruction[[DISK]DS-f863e30f-ce5b-48cc-9cca-72f64c558adc:NORMAL|RBW]]}
size 11837530 |
org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.logAddStoredBlock(BlockManager.java
:2393)
2015-05-27 19:00:48,297 | INFO  | IPC Server handler 37 on 25000 | BLOCK* addStoredBlock:
blockMap updated: 192.168.10.10:25009 is added to blk_1099105501_25370893 size 11837530 |
org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.logAddStoredBlock(BlockManager.java
:2393)
```

6. DataNode上报块写成功通知延迟的原因可能有：网络瓶颈导致、CPU瓶颈导致。
7. 如果此时再次调用close或者close的retry的次数增多，那么close都将返回成功。建议适当增大参数dfs.client.block.write.locateFollowingBlock.retries的值，默认值为5次，尝试的时间间隔为400ms、800ms、1600ms、3200ms、6400ms，12800ms，那么close函数最多需要25.2秒才能返回。

解决办法

步骤1 规避办法：

可以通过调整客户端参数dfs.client.block.write.locateFollowingBlock.retries的值来增加retry的次数，可以将值设置为6，那么中间睡眠等待的时间为400ms、800ms、

1600ms、3200ms、6400ms、12800ms，也就是说close函数最多要50.8秒才能返回。

----结束

备注说明

一般出现上述现象，说明集群负载很大，通过调整参数只是临时规避这个问题，建议还是降低集群负载。例如：避免把所有CPU都分配MR跑任务。

15.9.15 文件错误导致上传文件到 HDFS 失败

问题背景与现象

用**hadoop dfs -put**把本地文件拷贝到HDFS上，有报错。

上传部分文件后，报错失败，从NameNode原生页面看，临时文件大小不再变化。

原因分析

1. 查看NameNode日志“/var/log/Bigdata/hdfs/nn/hadoop-omm-namenode-主机名.log”，发现该文件一直在被尝试写，直到最终失败。

   ```
   2015-07-13 10:05:07,847 | WARN  | org.apache.hadoop.hdfs.server.namenode.LeaseManager
   $Monitor@36fea922 | DIR* NameSystem.internalReleaseLease: Failed to release lease for file /hive/
   order/OS_ORDER._8.txt._COPYING_. Committed blocks are waiting to be minimally replicated. Try
   again later. | FSNamesystem.java:3936
   2015-07-13 10:05:07,847 | ERROR | org.apache.hadoop.hdfs.server.namenode.LeaseManager
   $Monitor@36fea922 | Cannot release the path /hive/order/OS_ORDER._8.txt._COPYING_ in the lease
   [Lease.  Holder: DFSClient_NONMAPREDUCE_-1872896146_1, pendingcreates: 1] |
   LeaseManager.java:459
   org.apache.hadoop.hdfs.protocol.AlreadyBeingCreatedException: DIR*
   NameSystem.internalReleaseLease: Failed to release lease for file /hive/order/
   OS_ORDER._8.txt._COPYING_. Committed blocks are waiting to be minimally replicated. Try again
   later.
   at FSNamesystem.internalReleaseLease(FSNamesystem.java:3937)
   ```

2. 根因分析：被上传的文件损坏，因此会上传失败。
3. 验证办法：cp或者scp被拷贝的文件，也会失败，确认文件本身已损坏。

解决办法

步骤1 文件本身损坏造成的此问题，采用正常文件进行上传。

----结束

15.9.16 界面配置 dfs.blocksize 后 put 数据，block 大小还是原来的大小

问题背景与现象

界面配置“dfs.blocksize”，将其设置为268435456，put数据，block大小还是原来的大小。

原因分析

客户端的“hdfs-site.xml”文件中的dfs.blocksize大小没有更改，以客户端配置为准。

解决办法

步骤1 确保“dfs.blocksize”为512的倍数。

步骤2 重新下载安装客户端或者更改客户端配置。

步骤3 dfs.blocksize是客户端配置，以客户端为准。若客户端不配置，以服务端为准。

----结束

15.9.17 读取文件失败，FileNotFoundException

问题背景与现象

有MapReduce任务所有map任务均成功，但reduce任务失败，查看日志发现报异常“FileNotFoundException...No lease on...File does not exist”。

```
Error: org.apache.hadoop.ipc.RemoteException(java.io.FileNotFoundException): No lease on /user/sparkhive/warehouse/daas/dsp/output/_temporary/1/_temporary/attempt_1479799053892_17075_r_000007_0/part-r-00007 (inode 6501287): File does not exist. Holder DFSClient_attempt_1479799053892_17075_r_000007_0_-1463597952_1 does not have any open files.
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkLease(FSNamesystem.java:3350)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.completeFileInternal(FSNamesystem.java:3442)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.completeFile(FSNamesystem.java:3409)
at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.complete(NameNodeRpcServer.java:789)
```

原因分析

FileNotFoundException...No lease on...File does not exist，该日志说明文件在操作的过程中被删除了。

1. 搜索HDFS的NameNode的审计日志（Active NameNode的/var/log/Bigdata/audit/hdfs/nn/hdfs-audit-namenode.log）搜索文件名，确认文件的创建时间。
2. 搜索文件创建到出现异常时间范围的NameNode的审计日志，搜索该文件是否被删除或者移动到其他目录。
3. 如果该文件没有被删除或者移动，可能是该文件的父目录，或者更上层目录被删除或者移动，需要继续搜索上层目录。如本样例中，是文件的父目录的父目录被删除。

   ```
   2017-05-31 02:04:08,286 | INFO  | IPC Server handler 30 on 25000 | allowed=true ugi=appUser@HADOOP.COM (auth:TOKEN) ip=/192.168.1.22  cmd=delete  src=/user/sparkhive/warehouse/daas/dsp/output/_temporary  dst=null  perm=null proto=rpc | FSNamesystem.java:8189
   ```

   说明

   - 如上日志说明：192.168.1.22 节点的appUser用户删除了/user/sparkhive/warehouse/daas/dsp/output/_temporary。
   - 可以使用**zgrep "文件名" *.zip**命令搜索zip包的内容。

解决办法

步骤1 需要排查业务，确认为何该文件或者文件的父目录被删除。

----结束

15.9.18 HDFS 写文件失败，item limit of / is exceeded

问题背景与现象

客户端或者上层组件日志报往HDFS的某目录写文件失败，报错为

The directory item limit of /tmp is exceeded: limit=5 items=5。

原因分析

1. 查看客户端或者NameNode运行日志“/var/log/Bigdata/hdfs/nn/hadoop-omm-namenode-XXX.log”存在异常提示The directory item limit of /tmp is exceeded:。该错误的含义为/tmp目录的文件数超过1048576的限制。
   ```
   2018-03-14 11:18:21,625 | WARN  | IPC Server handler 62 on 25000 | DIR* NameSystem.startFile: /tmp/test.txt The directory item limit of /tmp is exceeded: limit=1048576 items=1048577 | FSNamesystem.java:2334
   ```
2. 该限制是dfs.namenode.fs-limits.max-directory-items参数，定义单个目录下不含递归的最大目录数或者文件数，默认值1048576，取值范围1～6400000。

解决办法

步骤1 确认该目录不含递归拥有100万以上文件目录是否正常，如果正常，可以将HDFS的参数dfs.namenode.fs-limits.max-directory-items调大并且重启HDFS NameNode生效。

步骤2 如果该目录下拥有100万文件不正常，需要清理不需要的文件。

----结束

15.9.19 调整 shell 客户端日志级别

- **临时调整**，关闭该shell客户端窗口后，日志会还原为默认值。
  a. 执行**export HADOOP_ROOT_LOGGER**命令可以调整客户端日志级别。
  b. 执行**export HADOOP_ROOT_LOGGER=*日志级别*,console**，可以调整shell客户端的日志级别。

     export HADOOP_ROOT_LOGGER=DEBUG,console，调整为DEBUG。

     export HADOOP_ROOT_LOGGER=ERROR,console，调整为ERROR。
- **永久调整**
  a. 在HDFS客户端环境变量配置文件“/opt/client/HDFS/component_env”（其中“/opt/client”需要改为实际客户端路径）增加“export HADOOP_ROOT_LOGGER=*日志级别*,console”。
  b. 执行**source /opt/client/bigdata_env**。
  c. 重新执行客户端命令。

15.9.20 读文件失败 No common protection layer

问题背景与现象

shell客户端或者其他客户端操作HDFS失败，报“**No common protection layer between client and server**”。

在集群外的机器，执行任意hadoop命令，如**hadoop fs -ls /**均失败，最底层的报错为"**No common protection layer between client and server**"。

```
2017-05-13 19:14:19,060 | ERROR | [pool-1-thread-1] |  Server startup failure  |
org.apache.sqoop.core.SqoopServer.initializeServer(SqoopServer.java:69)
org.apache.sqoop.common.SqoopException: MAPRED_EXEC_0028:Failed to operate HDFS - Failed to get the
file /user/loader/etl_dirty_data_dir status
        at org.apache.sqoop.job.mr.HDFSClient.fileExist(HDFSClient.java:85)
...
        at java.lang.Thread.run(Thread.java:745)
Caused by: java.io.IOException: Failed on local exception: java.io.IOException: Couldn't setup connection for
loader/hadoop@HADOOP.COM to loader37/10.162.0.37:25000; Host Details : local host is:
"loader37/10.162.0.37"; destination host is: "loader37":25000;
        at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:776)
...
        ... 10 more
Caused by: java.io.IOException: Couldn't setup connection for loader/hadoop@HADOOP.COM to
loader37/10.162.0.37:25000
        at org.apache.hadoop.ipc.Client$Connection$1.run(Client.java:674
        ... 28 more
Caused by: javax.security.sasl.SaslException: No common protection layer between client and server
        at com.sun.security.sasl.gsskerb.GssKrb5Client.doFinalHandshake(GssKrb5Client.java:251)
...
        at org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:720)
```

原因分析

1. HDFS的客户端和服务端数据传输走的rpc协议，该协议有多种加密方式，由hadoop.rpc.protection参数控制。
2. 如果客户端和服务端的hadoop.rpc.protection参数的配置值不一样，即会报**No common protection layer between client and server**错误。

   说明

   hadoop.rpc.protection参数表示数据可通过以下任一方式在节点间进行传输。

   - privacy：指数据在鉴权及加密后再传输。这种方式会降低性能。
   - authentication：指数据在鉴权后直接传输，不加密。这种方式能保证性能但存在安全风险。
   - integrity：指数据直接传输，即不加密也不鉴权。 为保证数据安全，请谨慎使用这种方式。

解决办法

步骤1 重新下载客户端，如果是应用程序，更新应用程序中的配置文件。

----结束

15.9.21 HDFS 目录配额（quota）不足导致写文件失败

问题背景与现象

给某目录设置quota后，往目录中写文件失败，出现如下问题“**The DiskSpace quota of** /tmp/tquota2 is **exceeded**”

```
[omm@189-39-150-115 client]$ hdfs dfs -put switchuser.py  /tmp/tquota2
put: The DiskSpace quota of /tmp/tquota2 is exceeded: quota = 157286400 B = 150 MB but diskspace
consumed = 402653184 B = 384 MB
```

可能原因

目录配置的剩余的空间小于写文件实际需要的空间。

原因分析

1. HDFS支持设置某目录的配额，即限制某目录下的文件最多占用空间大小，例如如下命令是设置/tmp/tquota 目录最多写入150MB的文件（文件大小*副本数）。

   hadoop dfsadmin -setSpaceQuota 150M /tmp/tquota2

2. 使用如下命令可以查看目录设置的配额情况，SPACE_QUOTA是设置的空间配额，REM_SPACE_QUOTA是当前剩余的空间配额。

   hdfs dfs -count -q -h -v /tmp/tquota2

   图 15-30 查看目录设置的配额

   ```
   hdfs dfs -count -q -h -v /tmp/tquota2
   QUOTA REM_QUOTA SPACE_QUOTA REM_SPACE_QUOTA  DIR_COUNT   FILE_COUNT CONTENT_SIZE PATHNAME
   none  inf       150M        150M                1           0            0  /tmp/tquota2
   ```

3. 日志分析，如下日志说明写入文件需要消耗384M，但是当前的空间配额是150M，因此空间不足。写文件前，需要的剩余空间是：块大小*副本数，128M*3副本=384M。

   ```
   [omm@189-39-150-115 client]$
   [omm@189-39-150-115 client]$ hdfs dfs -put switchuser.py  /tmp/tquota2
   put: The DiskSpace quota of /tmp/tquota2 is exceeded: quota = 157286400 B = 150 MB but diskspace consumed = 402653184 B = 384 MB
   ```

解决办法

步骤1 增加配额大小，即重新设置目录的配额大小。

hadoop dfsadmin -setSpaceQuota 150G /*目录名*

步骤2 清空配额。

hdfs dfsadmin -clrSpaceQuota /*目录名*

----结束

15.9.22 执行 balance 失败，Source and target differ in block-size

问题背景与现象

执行distcp跨集群拷贝文件时，出现部分文件拷贝失败“ Source and target differ in block-size. Use -pb to preserve block-sizes during copy. ”

```
Caused by: java.io.IOException: Check-sum mismatch between hdfs://10.180.144.7:25000/kylin/kylin_default_instance_prod/parquet/f2e72874-f01c-45ff-b219-207f3a5b3fcb/c769cd2d-575a-4459-837b-a19dd7b20c27/339114721280/0.parquettar and hdfs://10.180.180.194:25000/kylin/kylin_default_instance_prod/parquet/f2e72874-f01c-45ff-b219-207f3a5b3fcb/.distcp.tmp.attempt_1523424430246_0004_m_000019_2. Source and target differ in block-size. Use -pb to preserve block-sizes during copy. Alternatively, skip checksum-checks altogether, using -skipCrc. (NOTE: By skipping checksums, one runs the risk of masking data-corruption during file-transfer.)        at org.apache.hadoop.tools.mapred.RetriableFileCopyCommand.compareCheckSums(RetriableFileCopyCommand.java:214)
```

可能原因

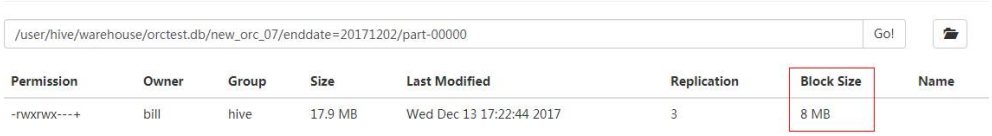
distcp默认拷贝文件时不记录原block大小导致在原文件block.size不是128M时校验失败，需要在distcp命令增加-pb参数。

原因分析

1. HDFS在写的时候有设置块大小，默认128M，某些组件或者业务程序写入的文件可能不是128M，如8M。

```
<name>dfs.blocksize</name>
<value>134217728</value>
```

图 15-31 某些组件或者业务程序写入的文件大小



2. distcp 从源集群读文件后写入新集群，默认是使用的MapReduce任务中的dfs.blocksize，默认128M。
3. 在distcp写完文件后，会基于块的物理大小做校验，因为该文件在新旧集群中block.size不一致，因此拆分大小不一致，导致校验失败。

   如以上文件，在旧集群是17.9/8MB = 3个block， 在新集群 17.9/128M = 1个block. 因此实际在磁盘的物理大小因分割而导致校验失败。

解决办法

distcp时，增加**-pb**参数。该参数作用为distcp时候保留block大小，确保新集群写入文件blocksize和老集群一致。

图 15-32 distcp 时保留 block 大小

```
[root@189-39-235-118 clientu10]#hadoop distcp -pb  hdfs://haclusterX/user hdfs://hacluster/tmp/test
```

15.9.23 查询或者删除文件失败，父目录可以看见此文件（不可见字符）

问题背景与现象

使用HDFS的shell客户端查询或者删除文件失败，父目录可以看见此文件。

图 15-33 父目录文件列表

```
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 01:44 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 16:45 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp2
[root@dggtsp355-or FusionInsight_client]# hadoop fs -ls /user/hive/warehouse/datalake_dwi_barpsit.db
Found 4 items
drwxrwxr-x   - datalab90020_639_w hive          0 2018-04-11 12:05 /user/hive/warehouse/datalake_dwi_barpsit.db/bak_v_tp_mp_aut_input
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-11 11:16 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 01:44 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp
drwxrwx---+  - datalab90020_639_w hive          0 2018-04-10 16:45 /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input_tmp2
[root@dggtsp355-or FusionInsight_client]# hadoop fs -rm -r /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
rm: `/user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input': No such file or directory
[root@dggtsp355-or FusionInsight_client]# hdfs dfs -rm -r /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
rm: `/user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input': No such file or directory
[root@dggtsp355-or FusionInsight_client]#
[root@dggtsp355-or FusionInsight_client]#
[root@dggtsp355-or FusionInsight_client]# hdfs dfs -ls /user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input
ls: `/user/hive/warehouse/datalake_dwi_barpsit.db/v_tp_mp_aut_input': No such file or directory
[root@dggtsp355-or FusionInsight_client]#
[root@dggtsp355-or FusionInsight_client]#
```

原因分析

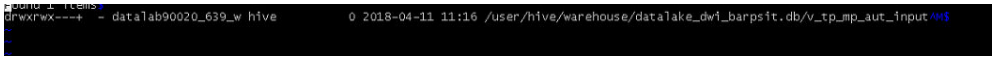
可能是该文件写入时有异常，写入了不可见字符。可以将该文件名重定向写入本地文本中，使用**vi**命令打开。

hdfs dfs -ls *父目录* **> /tmp/t.txt**

vi /tmp/t.txt

然后输入命令“**:set list**”将文件名的不可见字符显示出来。如这里显示出文件名中包含“**^M**”不可见字符。

图 15-34 显示不可见字符

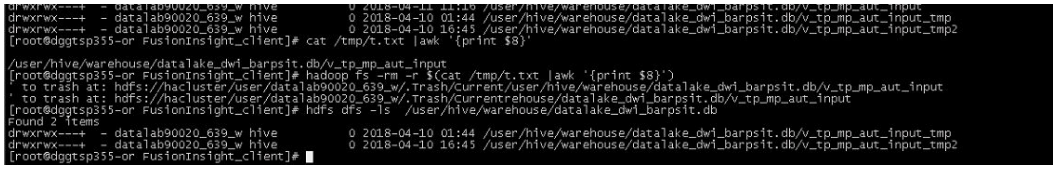


解决办法

步骤1 使用shell命令读到文本中记录的该文件名，确认如下命令输出的是该文件在HDFS中的全路径。

cat /tmp/t.txt |awk '{print $8}'

图 15-35 文件路径



步骤2 使用如下命令删除该文件。

hdfs dfs -rm $(cat /tmp/t.txt |awk '{print $8}')

步骤3 查看确认该文件已被删除。

hdfs dfs -ls *父目录*

----结束

15.9.24 非 HDFS 数据残留导致数据分布不均衡

问题背景与现象

数据出现不均衡，某磁盘过满而其他磁盘未写满。

HDFS DataNode数据存储目录配置为“/export/data1/dfs--/export/data12/dfs”，看到的现象是大量数据都是存储到了“/export/data1/dfs”，其他盘的数据比较均衡。

原因分析

磁盘为卸载重装，有一个目录在上次卸载时未卸载干净，即添加的磁盘，未格式化，残留历史垃圾数据。

解决办法

手动清理未卸载干净的数据。

15.9.25 客户端安装在数据节点导致数据分布不均衡

问题背景与现象

HDFS的DataNode数据分布不均匀，在某节点上磁盘使用率很高，甚至达到100%，其他节点空闲很多。

原因分析

客户端安装在该节点，根据HDFS数据副本机制，第一个副本会存放在本地机器，最终导致节点磁盘被占满，而其他节点空闲很多。

解决办法

步骤1 针对已有不平衡的数据，执行balance脚本均衡数据。

/opt/client/HDFS/hadoop/sbin/start-balancer.sh -threshold 10

其中 /opt/client是实际的客户端安装目录。

步骤2 针对新写入数据，将客户端安装在没有安装DataNode的节点。

----结束

15.9.26 节点内 DataNode 磁盘使用率不均衡处理指导

问题背景与现象

单个节点内DataNode的各磁盘使用率不均匀。

例如：

```
189-39-235-71:~ # df -h
Filesystem  Size  Used Avail Use% Mounted on
/dev/xvda  360G  92G  250G  28% /
/dev/xvdb  700G  900G  200G  78% /srv/BigData/hadoop/data1
/dev/xvdc  700G  900G  200G  78% /srv/BigData/hadoop/data2
/dev/xvdd  700G  900G  200G  78% /srv/BigData/hadoop/data3
/dev/xvde  700G  900G  200G  78% /srv/BigData/hadoop/data4
/dev/xvdf  10G  900G  890G  2% /srv/BigData/hadoop/data5
189-39-235-71:~ #
```

可能原因

部分磁盘故障，更换为新盘，因此新盘使用率低。

增加了磁盘个数，如原先4个数据盘，现扩容为5个数据盘。

原因分析

DataNode节点内写block磁盘时，有两种策略“轮询”和“优先写剩余磁盘空间多的磁盘”，默认是“轮询”。

参数说明：dfs.datanode.fsdataset.volume.choosing.policy

可选值：

- 轮询：
  org.apache.hadoop.hdfs.server.datanode.fsdataset.RoundRobinVolumeChoosingPolicy
- 优先写剩余空间多的磁盘：
  org.apache.hadoop.hdfs.server.datanode.fsdataset.AvailableSpaceVolumeChoosingPolicy

解决办法

将DataNode选择磁盘策略的参数 dfs.datanode.fsdataset.volume.choosing.policy 的值改为：
org.apache.hadoop.hdfs.server.datanode.fsdataset.AvailableSpaceVolumeChoosingPolicy，保存并重启受影响的服务或实例。

让DataNode根据磁盘剩余空间大小，优先选择磁盘剩余空间多的节点存储数据副本。

说明

- 针对新写入到本DataNode的数据会优先写磁盘剩余空间多的磁盘。
- 部分磁盘使用率较高，依赖业务逐渐删除在HDFS中的数据（老化数据）来逐渐降低。

15.9.27 执行 balance 常见问题定位方法

问题 1：报没权限（Access denied）执行 balance

问题详细：执行start-balancer.sh，“hadoop-root-balancer-主机名.out”日志显示“Access denied for user test1. Superuser privilege is required”

```
cat /opt/client/HDFS/hadoop/logs/hadoop-root-balancer-host2.out
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  Bytes Being Moved
INFO: Watching file:/opt/client/HDFS/hadoop/etc/hadoop/log4j.properties for changes with interval : 60000
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.AccessControlException): Access denied for user test1.
Superuser privilege is required
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkSuperuserPrivilege(FSPermissionChecker.java:122)
at
org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkSuperuserPrivilege(FSNamesystem.java:5916)
```

问题根因：

执行balance需要使用管理员账户

解决方法

- 安全版本
  使用hdfs或者其他属于supergroup组的用户认证后，执行balance
- 普通版本
  执行HDFS的balance命令前，需要在客户端执行su - hdfs命令。

问题 2：执行 balance 失败，/system/balancer.id 文件异常

问题详细：

在HDFS客户端启动一个Balance进程，该进程被异常停止后，再次执行Balance操作，操作会失败。

```
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.protocol.RecoveryInProgressException): Failed to APPEND_FILE /system/balancer.id for DFSClient because lease recovery is in progress. Try again later.
```

问题根因：

通常，HDFS执行Balance操作结束后，会自动释放“/system/balancer.id”文件，可再次正常执行Balance。

但在上述场景中，由于第一次的Balance操作是被异常停止的，所以第二次进行Balance操作时，“/system/balancer.id”文件仍然存在，则会触发append /system/balancer.id操作，进而导致Balance操作失败。

解决方法

方法1：等待硬租期超过1小时后，原有客户端释放租约，再执行第二次Balance操作。

方法2：删除HDFS中的“/system/balancer.id”文件，再执行下次Balance操作。

15.9.28 HDFS 显示磁盘空间不足，其实还有 10%磁盘空间

问题背景与现象

1. 出现“HDFS磁盘空间使用率超过阈值”告警。
2. 查看HDFS页面，查看磁盘空间使用率非常高。

原因分析

HDFS中配置了dfs.datanode.du.reserved.percentage参数：每个磁盘的保留空间所占磁盘百分比。DataNode会保留这么多可用空间，以备其他组件如Yarn的NodeManager运行计算时，或者预留升级时使用。

因为预留了10%的磁盘，当磁盘使用率达到90%的时候，HDFS的DataNode即会认为没有可用磁盘空间。

解决办法

步骤1 扩容，在HDFS DataNode磁盘到80%，即需要及时扩容。

步骤2 如不能及时扩容，需要删除HDFS中的不需要数据，释放磁盘空间。

----结束

15.9.29 普通集群在 Core 节点安装 hdfs 客户端，使用时报错

用户问题

普通集群在Core节点新建用户安装使用客户端报错。

问题现象

普通集群在Core节点新建用户安装使用客户端报错如下：

```
2020-03-14 19:16:17,166 WARN shortcircuit.DomainSocketFactory: error creating DomainSocket
java.net.ConnectException: connect(2) error: Permission denied when trying to connect to '/var/run/MRS-
```

```
HDFS/dn_socket'
at org.apache.hadoop.net.unix.DomainSocket.connect0(Native Method)
at org.apache.hadoop.net.unix.DomainSocket.connect(DomainSocket.java:256)
at org.apache.hadoop.hdfs.shortcircuit.DomainSocketFactory.createSocket(DomainSocketFactory.java:168)
at org.apache.hadoop.hdfs.client.impl.BlockReaderFactory.nextDomainPeer(BlockReaderFactory.java:799)
...
```

原因分析

用户使用 **useradd** 命令来创建用户，此用户默认用户组不包含“ficommmon”用户组，导致在使用hdfs的get命令的时候出现上述报错。

处理步骤

使用命令**usermod -a -G ficommon username** 为用户添加用户组“ficommon”。

15.9.30 集群外节点安装客户端使用 hdfs 上传文件失败

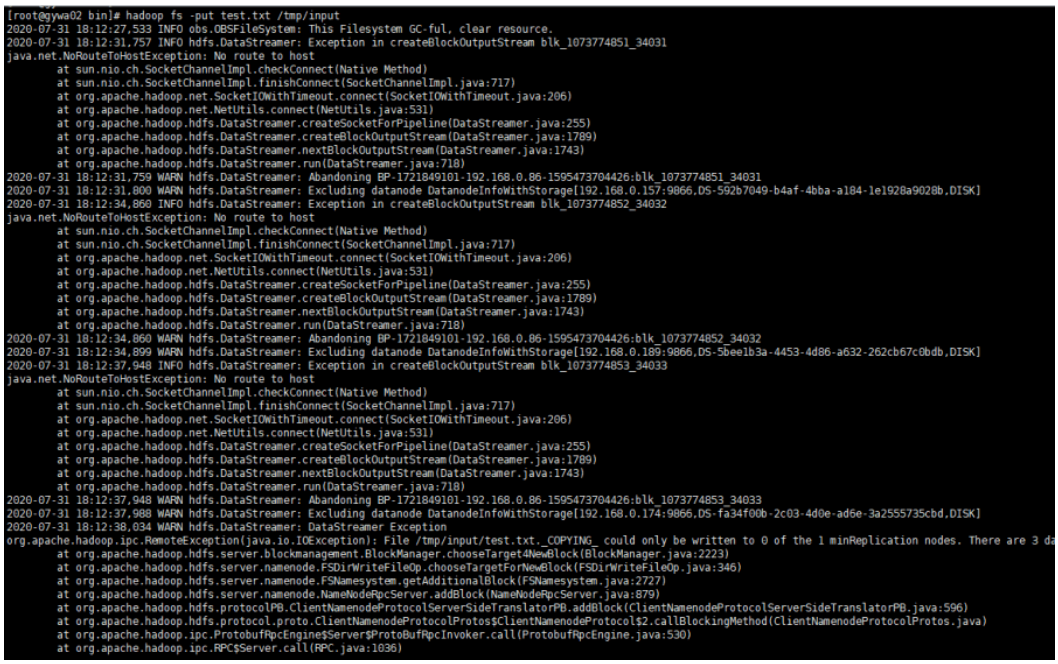
用户问题

集群外节点安装客户端使用hdfs上传文件失败

问题现象

在集群节点上安装客户端，在该客户端使用hdfs命令上传一个文件，报如下错误：

图 15-36 上传文件报错



原因分析

从错误截图可以看到报错是no route to host，且报错信息里面有192.168的ip，也即客户端节点到集群的DN节点的内网路由不通，导致上传文件失败。

处理步骤

在客户端节点的客户端目录下，找到HDFS的客户端配置目录hdfs-site.xml文件，在配置文件中增加配置项dfs.client.use.datanode.hostname，并将该配置设置为true。

15.9.31 HDFS 写并发较大时，报副本不足的问题

问题背景与现象

用户运行作业时写文件到HDFS，偶现写文件失败的情况。

操作日志如下：

```
105 | INFO  | IPC Server handler 23 on 25000 | IPC Server handler 23 on 25000, call
org.apache.hadoop.hdfs.protocol.ClientProtocol.addBlock from 192.168.1.96:47728 Call#1461167 Retry#0 |
Server.java:2278
java.io.IOException: File /hive/warehouse/000000_0.835bf64f-4103 could only be replicated to 0 nodes
instead of minReplication (=1).  There are 3 datanode(s) running and 3 node(s) are excluded in this
operation.
```

原因分析

- HDFS写文件的预约机制：无论文件是10M还是1G，开始写的每个块都会被预约128M。如果需要写入一个10M的文件，HDFS会预约一个块来写，当文件写完后，这个块只占实际大小10M，释放多余预约的118M空间。如果需要写入一个1G的文件，HDFS还是会预约一个块来写，这个块写完后再开启下一个块，文件写完后，实际占用1G磁盘，释放多余预约的空间。
- 该异常通常是因为业务写文件的并发量太高，预约写Block的磁盘空间不足，导致写文件失败。

解决办法

步骤1 登录HDFS的WebUI页面，进入DataNode的JMX页面。

1. 在HDFS原生界面，选择Datanodes页面。
2. 找到对应的DataNode节点，单击Http Address地址进入DataNode详情。
3. 将url的“datanode.html”改为“jmx”就能获取到DataNode的JMX信息。

步骤2 搜索“XceiverCount”指标，当该指标的值*Block块的大小超过DataNode磁盘的容量，就说明预约写Block的磁盘空间不足。

步骤3 发生该问题，通常有以下两种方法来解决：

方法一：降低业务的并发度。

方法二：减少业务写文件的数目，将多个文件合并成一个文件来写。

----结束

15.9.32 HDFS 客户端无法删除超长目录

问题背景与现象

执行**hadoop fs -rm -r -f obs://***<obs_path>*命令，删除OBS超长目录出现如下报错：

```
2022-02-28 17:12:45,605 INFO internal.RestStorageService: OkHttp cost 19 ms to apply http request
2022-02-28 17:12:45,606 WARN internal.RestStorageService: Request failed, Response code: 400; Request
```

```
ID: 0000017F3F9A8545401491602FC8CAD9; Request path: http://wordcount01-fcq.obs.xxx.***.xxx.com/user
%2Froot%2F.Trash%2FCurrent
%2Ftest1%2F12345678901234567890123456789012345678901234567890123456789012345678901234567
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
789012345678901234567890123456789012345678901234567890123456789012345678901234567890123
456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
123456789012345678901234567890123456789012345678901234567890123456789012345678901234567
890123456789012345678901234567890123456789012345678901234567890123456
2022-02-28 17:12:45,606 WARN services.AbstractClient: Storage|1|HTTP+XML|getObjectMetadata||||
2022-02-28 17:12:45|2022-02-28 17:12:45|||400|
2022-02-28 17:12:45,607 INFO log.AccessLogger: 2022-02-28 17:12:45 605|
com.obs.services.internal.RestStorageService|executeRequest|560|OkHttp cost 19 ms to apply http request
2022-02-28 17:12:45 606|com.obs.services.internal.RestStorageService|handleThrowable|221|Request failed,
Response code: 400; Request ID: 0000017F3F9A8545401491602FC8CAD9; Request path: http://wordcount01-
fcq.obs.xxx.***.xxx.com/user%2Froot%2F.Trash%2FCurrent
%2Ftest1%2F12345678901234567890123456789012345678901234567890123456789012345678901234567
890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
012345678901234567890123456789012345678901234567890123456789012345678901234567890123456
789012345678901234567890123456789012345678901234567890123456789012345678901234567890123
456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
123456789012345678901234567890123456789012345678901234567890123456789012345678901234567
890123456789012345678901234567890123456789012345678901234567890123456
2022-02-28 17:12:45 606|com.obs.services.AbstractClient|doActionWithResult|404|Storage|1|HTTP+XML|
getObjectMetadata||||2022-02-28 17:12:45|2022-02-28 17:12:45|||400|
```

原因分析

使用rm命令从HDFS中删除某些内容时，文件或目录不会立即被清除，它们将被移动到回收站Current目录（ /user/*${username}*/.Trash/current ）中。

解决办法

使用skipTrash命令可以跳过HDFS回收站，直接删除。使用前先设置HDFS客户端配置项“dfs.client.skipTrash.enabled=true”。

步骤1 以**root**用户登录集群任一Master节点。

步骤2 执行如下命令编辑HDFS用到的hdfs-site.xml文件。

vim *客户单安装目录*/**HDFS/hadoop/etc/hadoop/hdfs-site.xml**

步骤3 在hdfs-site.xml文件中增加如下内容。

```
<property>
<name>dfs.client.skipTrash.enabled</name>
<value>true</value>
</property>
```

步骤4 执行以下命令直接删除OBS超长目录。

hadoop fs -rm -r -f –skipTrash obs://*<obs_path>*

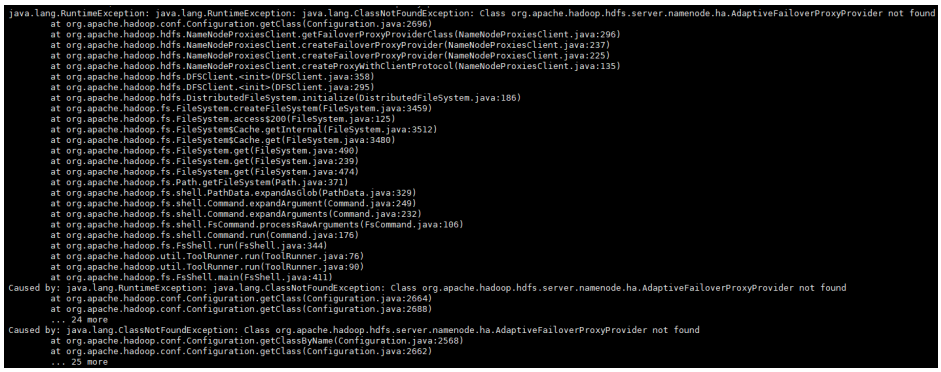
步骤5 登录集群其他Master节点，执行**步骤2**~**步骤4**，直到集群所有Master节点操作完成。

----结束

15.9.33 集群外节点访问 MRS HDFS 报错

问题背景与现象

集群外节点访问MRS HDFS的时候报错：Class org.apache.hadoop.hdfs.server.namenode.ha.AdaptiveFailoverProxyProvider not found。



原因分析

出现这个报错可能的场景有：

- 开源HDFS客户端访问MRS集群的HDFS时报错。
- 使用jar包连接MRS集群的HDFS（包括提交任务时连接HDFS）时报错。

解决办法

方法一：

步骤1 找到命令或者jar包使用的HDFS配置文件hdfs-site.xml。

步骤2 修改“dfs.client.failover.proxy.provider.hacluster”参数配置项如下。

```
<property>
<name>dfs.client.failover.proxy.provider.hacluster</name>
<value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider</value>
</property>
```

说明

也可以将上述配置项删除。

步骤3 保存文件后，重新访问MRS HDFS。

----结束

方法二：

步骤1 从Maven库下载MRS集群版本对应的hadoop-plugins版本。

步骤2 将下载的jar包添加到命令或者jar包的依赖中。

----结束

15.10 使用 Hive

15.10.1 Hive 各个日志里都存放了什么信息?

审计日志

首先，对于审计日志来说，记录了某个时间点某个用户从哪个IP发起对HiveServer或者MetaStore的请求以及记录执行的语句是什么。

如下的HiveServer审计日志，表示在2016-02-01 14:51:22 用户user_chen向HiveServer发起了show tables请求，客户端IP为192.168.1.18。

```
2016-02-01 14:51:22,335 | INFO  | HiveServer2-Handler-Pool: Thread-37815 | UserN
ame=user_chen  UserIP=192.168.1.18  Time=2016/02/01 14:51:22  Opertai
on=ExecuteStatement  stmt={show tables}  Resource=  Result= Detail=
| org.apache.hive.service.cli.thrift.ThriftCLIService.logAuditEvent(ThriftCLISer
vice.java:350)
```

如下面MetaStore审计日志，表示在2016-01-29 11:31:15 用户hive向MetaStore发起shutdown请求，客户端ip为192.168.1.18。

```
2016-01-29 11:31:15,451 | INFO  | pool-6-thread-70648 | ugi=hive/hadoop.hadoop.c
om@HADOOP.COM  ip=/192.168.1.18  cmd=Shutting down the object store...
 | org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.logAuditEvent(HiveM
etaStore.java:375)
```

通常情况下，审计日志对我们定位实际错误信息并无太大帮助。但在遇到诸如下述类问题时，需要查看审计日志：

1. 如客户端发起请求，但是迟迟未得到响应。由于不确定到底是任务是卡在客户端还是服务端，可以通过审计日志查看。如果审计日志根本没有相关信息，那么说明卡死在客户端；如审计日志有相关打印，那么就需要去运行日志里看到底程序卡在哪一步了。
2. 查看指定时间段的任务请求个数。可通过审计日志查看在指定时间段到底有多少个请求。

HiveServer 运行日志

简言之，HiveServer负责接收客户端请求（SQL语句），然后编译、执行（提交到YARN或运行local MR）、与MetaStore交互获取元数据信息等。HiveServer运行日志记录了一个SQL完整的执行过程。

通常情况下，当我们遇到SQL语句运行失败，我们首先要看的就是HiveServer运行日志。

MetaStore 运行日志

通常情况下，当我们遇到查看HiveServer运行日志时，如遇到MetaException或者连接MetaStore失败，我们就需要去看MetaStore运行日志了。

GC 日志查看

HiveServer和MetaStore均有GC日志，当我们遇到GC问题可以查看GC日志以快速定位是否是GC导致。如，当遇到HiveServer或MetaStore频繁重启就需要去看下对应的GC日志了。

15.10.2 Hive 启动失败问题的原因有哪些?

Hive启动失败最常见的原因是metastore实例无法连接上DBservice。可以查看metastore日志中具体的错误信息。目前总结连不上DBservice原因主要有：

可能原因 1

DBservice没有初始化好Hive的元数据库hivemeta。

处理步骤 1

步骤1 执行以下命令：

source /opt/Bigdata/MRS_*XXX*/install/dbservice/.dbservice_profile

gsql -h *DBservice浮动IP* **-p 20051 -d hivemeta -U hive -W HiveUser@**

步骤2 如果不能正确进入交互界面，说明数据库初始化失败。如果报如下错误说明在DBservice所在的节点的配置文件可能丢失了hivemeta的配置。

```
org.postgresql.util.PSQLException: FATAL: no pg_hba.conf entry for host "192.168.0.146", database "HIVEMETA"。
```

步骤3 编辑“/srv/BigData/dbdata_service/data/pg_hba.conf”，在文件最后面追加**host hivemeta hive 0.0.0.0/0 sha256**配置。

步骤4 执行**source /opt/Bigdata/MRS_*XXX*/install/dbservice/.dbservice_profile**命令配置环境变量。

步骤5 执行**gs_ctl -D $GAUSSDATA reload #**命令使修改后的配置生效。

----结束

可能原因 2

DBservice的浮动IP配置有误，导致metastore节点IP无法正确连接浮动IP，或者是在与该ip建立互信的时候失败导致metastore启动失败。

处理步骤 2

DBservice的浮动IP配置需要同网段内没有被使用过的ip，也就是在配置前ping不通的ip，请修改DBService浮动IP配置。

15.10.3 安全集群执行 set 命令的时候报 Cannot modify xxx at runtime.

问题现象

执行set命令时报以下错误：

```
0: jdbc:hive2://192.168.1.18:21066/> set mapred.job.queue.name=QueueA;
 Error: Error while processing statement: Cannot modify mapred.job.queue.name at list of params that are allowed to be modified at runtime (state=42000,code=1)
```

处理步骤

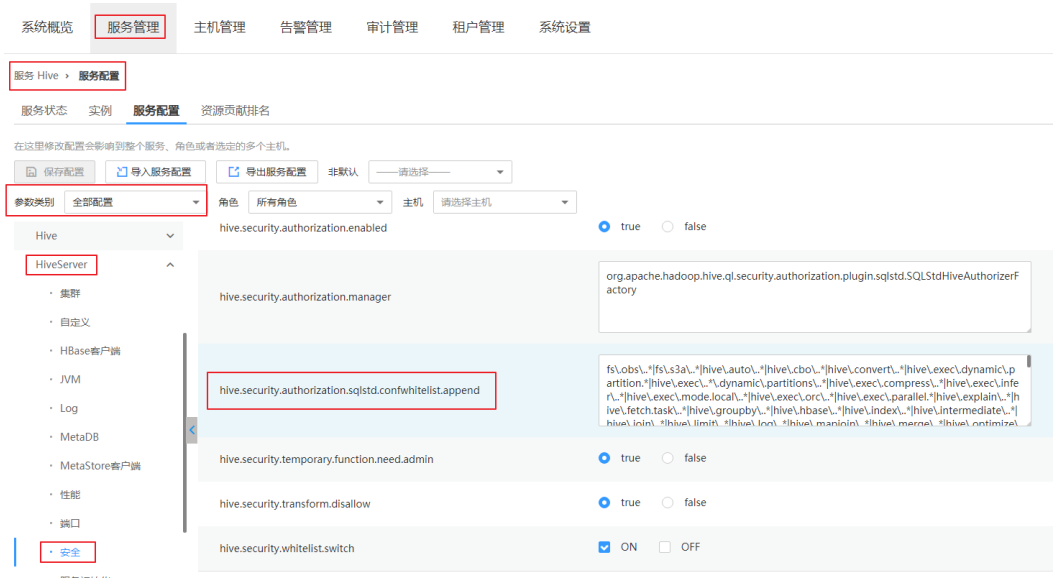
方案1：

步骤1 登录Manager界面，修改Hive参数。

- MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Hive > 服务配置 > 全部配置 > HiveServer > 安全”。
- FusionInsight Manager界面操作：登录FusionInsight Manager页面，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置 > 全部配置 > HiveServer > 安全”。

步骤2 将需要执行的命令参数添加到配置项hive.security.authorization.sqlstd.confwhitelist.append中。

步骤3 单击保存并重启HiveServer后即可。如下图所示：



----结束

方案2：

步骤1 登录Manager界面，修改Hive参数。

- MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Hive > 服务配置 > 全部配置 > HiveServer > 安全”。
- FusionInsight Manager界面操作：登录FusionInsight Manager页面，选择“集群 > *待操作集群的名称* > 服务 > Hive > 配置 > 全部配置 > HiveServer > 安全”。

步骤2 找到选项hive.security.whitelist.switch，选择OFF，单击保存并重启HiveServer即可。

----结束

15.10.4 怎样在 Hive 提交任务的时候指定队列?

问题现象

怎样在Hive提交任务的时候指定队列?

处理步骤

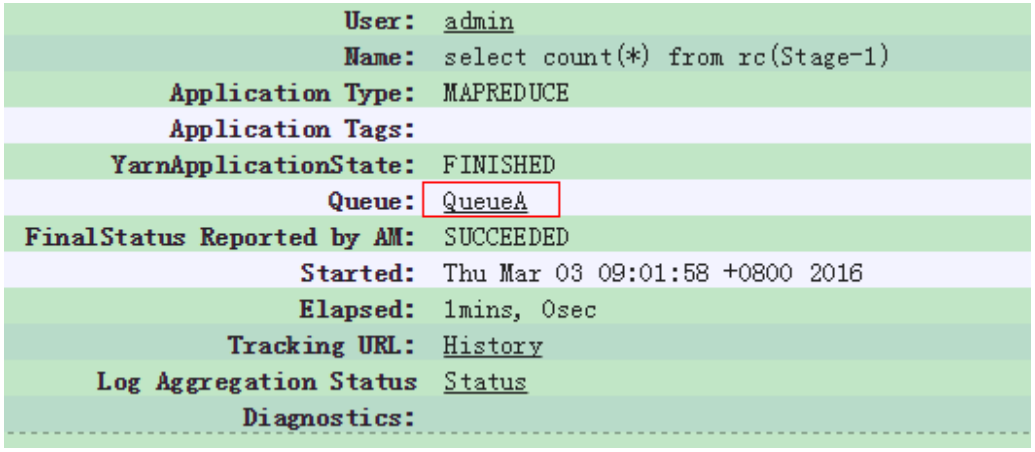
步骤1 在执行语句前通过如下参数设置任务队列，例如，提交任务至队列QueueA。

```
set mapred.job.queue.name=QueueA;
select count(*) from rc;
```

说明

队列的名称区分大小写，如写成queueA，Queuea均无效；且该队列为叶子队列，不支持提交任务到非叶子队列。

步骤2 提交任务后，可在Yarn页面看到，如下任务已经提交到队列QueueA。



----结束

15.10.5 客户端怎么设置 Map/Reduce 内存?

问题现象

客户端怎么设置Map/Reduce内存?

处理步骤

Hive在执行SQL语句前，可以通过set命令来设置Map/Reduce相关客户端参数。

以下为与Map/Reduce内存相关的参数：

```
set mapreduce.map.memory.mb=4096;// 每个Map Task需要的内存量
set mapreduce.map.java.opts=-Xmx3276M; // 每个Map Task 的JVM最大使用内存
set mapreduce.reduce.memory.mb=4096; // 每个Reduce Task需要的内存量
set mapreduce.reduce.java.opts=-Xmx3276M; // 每个Reduce Task 的JVM最大使用内存
set mapred.child.java.opts=-Xms1024M -Xmx3584M;//此参数为全局参数，即对Map和Reduce统一设置
```

说明

参数设置只对当前session有效。

15.10.6 如何在导入表时指定输出的文件压缩格式

问题现象

如何在导入表时指定输出的文件压缩格式?

处理步骤

当前Hive支持以下几种压缩格式:

```
org.apache.hadoop.io.compress.BZip2Codec
org.apache.hadoop.io.compress.Lz4Codec
org.apache.hadoop.io.compress.DeflateCodec
org.apache.hadoop.io.compress.SnappyCodec
org.apache.hadoop.io.compress.GzipCodec
```

- 如需要全局设置，即对所有表都进行压缩，可以在Manager页面对Hive的服务配置参数进行如下全局配置:
  - hive.exec.compress.output设置为true
  - mapreduce.output.fileoutputformat.compress.codec设置为org.apache.hadoop.io.compress.BZip2Codec

    说明

    hive.exec.compress.output参数必须设置为true，才能使下边的参数选项生效。
- 如需在session级设置，只需要在执行命令前增加如下设置即可:

  ```
  set hive.exec.compress.output=true;
  set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.SnappyCodec;
  ```

15.10.7 desc 描述表过长时，无法显示完整

问题现象

desc描述表过长时，如何让描述显示完整?

处理步骤

步骤1 启动Hive的beeline时，设置参数maxWidth=20000即可，例如:

```
[root@192-168-1-18 logs]# beeline --maxWidth=20000
scan complete in 3ms
Connecting to
……
Beeline version 1.1.0 by Apache Hive
```

步骤2 （可选）通过**beeline -help**命令查看关于客户端显示的设置。如下:

```
 -u <database url>              the JDBC URL to connect to
 -n <username>                  the username to connect as
 -p <password>                  the password to connect as
 -d <driver class>              the driver class to use
 -i <init file>                 script file for initialization
 -e <query>                     query that should be executed
 -f <exec file>                 script file that should be executed
 --hiveconf property=value      Use value for given property
 --color=[true/false]           control whether color is used for display
 --showHeader=[true/false]      show column names in query results
 --headerInterval=ROWS;         the interval between which heades are displayed
 --fastConnect=[true/false]     skip building table/column list for tab-completion
 --autoCommit=[true/false]      enable/disable automatic transaction commit
 --verbose=[true/false]         show verbose error messages and debug info
```

```
 --showWarnings=[true/false]     display connection warnings
 --showNestedErrs=[true/false]    display nested errors
 --numberFormat=[pattern]         format numbers using DecimalFormat pattern
 --force=[true/false]             continue running script even after errors
 --maxWidth=MAXWIDTH              the maximum width of the terminal
 --maxColumnWidth=MAXCOLWIDTH     the maximum width to use when displaying columns
 --silent=[true/false]            be more silent
 --autosave=[true/false]          automatically save preferences
 --outputformat=[table/vertical/csv2/tsv2/dsv/csv/tsv]  format mode for result display
                                  Note that csv, and tsv are deprecated - use csv2, tsv2 instead
 --truncateTable=[true/false]     truncate table column when it exceeds length
 --delimiterForDSV=DELIMITER      specify the delimiter for delimiter-separated values output format
(default: |)
 --isolation=LEVEL                set the transaction isolation level
 --nullemptystring=[true/false]  set to true to get historic behavior of printing null as empty string
 --socketTimeOut=n                socket connection timeout interval, in second. The default value is 300.
```

----结束

15.10.8 增加分区列后再 insert 数据显示为 NULL

问题现象

1. 执行如下命令创建表
   ```
   create table test_table(
   col1 string,
   col2 string
   )
   PARTITIONED BY(p1 string)
   STORED AS orc tblproperties('orc.compress'='SNAPPY');
   ```
2. 修改表结构，添加分区并插入数据
   ```
   alter table test_table add partition(p1='a');
   insert into test_table partition(p1='a') select col1,col2 from temp_table;
   ```
3. 修改表结构，添加列并插入数据
   ```
   alter table test_table add columns(col3 string);
   insert into test_table partition(p1='a') select col1,col2,col3 from temp_table;
   ```
4. 查询test_table表数据，返回结果中列col3的值全为NULL
   ```
   select * from test_table where p1='a'
   ```
5. 新添加表分区，并插入数据
   ```
   alter table test_table add partition(p1='b');
   insert into test_table partition(p1='b') select col1,col2,col3 from temp_table;
   ```
6. 查询test_table表数据，返回结果中列col3有不为NULL的值
   ```
   select * from test_table where p1='b'
   ```

原因分析

在alter table时默认选项为RESTRICT，RESTRICT只会更改元数据，不会修改在此操作之前创建的partition的表结构，而只会修改在此之后创建的新的partition，所以查询时旧的partition中的值全为NULL。

处理步骤

add column时加入cascade关键字即可，例如：

```
alter table test_table add columns(col3 string) cascade;
```

15.10.9 创建新用户，执行查询时报无权限

问题现象

创建了新用户，但是执行查询的时候报无权限的错。

```
Error: Error while compiling statement: FAILED: HiveAccessControlException Permission denied: Principal [name=hive, type=USER] does not have following privileges for operation QUERY [[SELECT] on Object [type=TABLE_OR_VIEW, name=default.t1]] (state=42000,code=40000)
```

原因分析

创建的新用户没有Hive组件的操作权限。

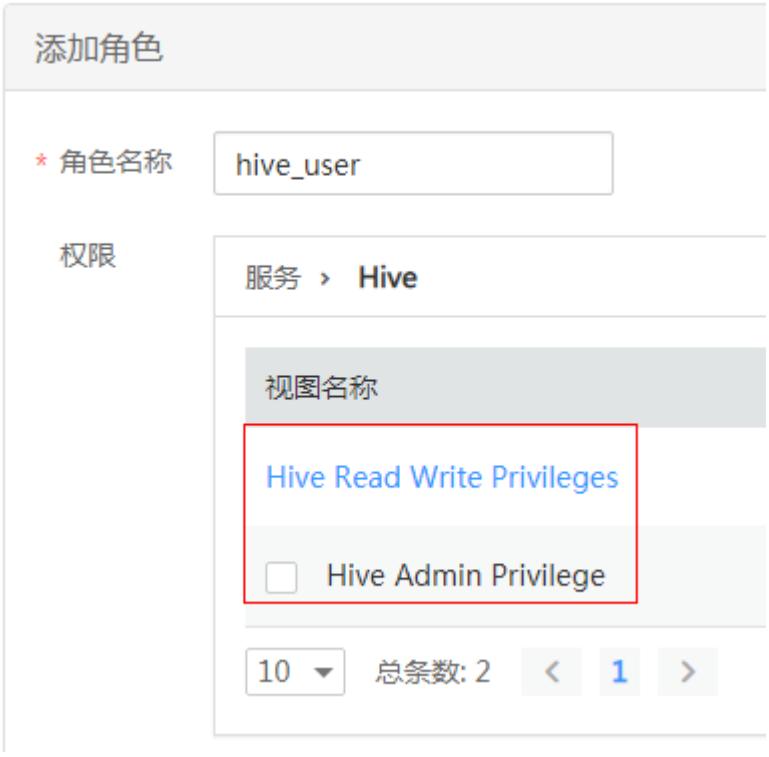
解决方案

MRS Manager界面操作：

步骤1 登录MRS Manager页面，选择“系统配置 > 角色管理 > 添加角色”。

步骤2 输入角色名称。

步骤3 在“权限”区域选择Hive，出现Hive管理员权限和Hive表的读写权限。



步骤4 选择“Hive Read Write Privileges”Hive表的读写权限，此时显示列Hive中的所有数据库。

步骤5 勾选角色需要的权限并单击“确定”完成角色创建。

步骤6 在MRS Manager页面，选择“系统配置 > 用户管理”。

步骤7 在已创建的新用户对应的“操作”列单击“修改”。

步骤8 单击“选择添加的用户组”，如需使用Hive服务，必须添加Hive组。

步骤9 单击“选择并绑定角色”，勾选**步骤5**中已创建的角色。

步骤10 单击“确定”完成用户权限的配置。

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。选择“系统 > 权限 > 角色”。

步骤2 单击“添加角色”，输入“角色名称”和“描述”。

步骤3 设置角色“配置资源权限”，选择“Hive读写权限”Hive表的读写权限，此时显示列Hive中的所有数据库。

步骤4 勾选角色需要的权限并单击“确定”完成角色创建。

步骤5 在FusionInsight Manager页面，选择“系统 > 权限 > 用户”。

步骤6 在已创建的新用户对应的“操作”列单击“修改”。

步骤7 单击“用户组”右侧的“添加”，如需使用Hive服务，必须添加Hive组。

步骤8 单击“角色”右侧的“添加”，勾选**4**中已创建的角色。

步骤9 单击“确定”完成用户权限的配置。

----结束

15.10.10 执行 SQL 提交任务到指定队列报错

问题现象

执行SQL提交任务到Yarn报如下错误：

```
Failed to submit application_1475400939788_0033 to YARN :
org.apache.hadoop.security.AccessControlException: User newtest cannot submit applications to queue
root.QueueA
```

原因分析

当前登录的用户无YARN队列提交权限。

解决方案

用户无YARN队列提交权限，需要赋予YARN相应队列的提交权限。在Manager页面的“系统 >权限 > 用户”中给用户绑定队列提交权限的角色。

15.10.11 执行 load data inpath 命令报错

问题现象

执行load data inpath报如下错误：

- 错误1：

  ```
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have
  following privileges on Object [type=DFS_URI, name=hdfs://hacluster/tmp/input/mapdata] for
  operation LOAD : [OBJECT OWNERSHIP]
  ```

- 错误2：
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have following privileges on Object [type=DFS_URI, name=hdfs://hacluster/tmp/input/mapdata] for operation LOAD : [INSERT, DELETE]
- 错误3：
  SemanticException [Error 10028]: Line 1:17 Path is not legal ''file:///tmp/input/mapdata'': Move from: file:/tmp/input/mapdata to: hdfs://hacluster/user/hive/warehouse/tmp1 is not valid. Please check that values for params "default.fs.name" and "hive.metastore.warehouse.dir" do not conflict.

原因分析

当前登录的用户不具备操作此目录的权限或者文件目录格式不正确。

解决方案

Hive对load data inpath命令有如下权限要求，请对照下述要求是否满足：

- 文件的owner需要为执行命令的用户。
- 当前用户需要对该文件有读、写权限。
- 当前用户需要对该文件的目录有执行权限。
- 由于load操作会将该文件移动到表对应的目录中，所以要求当前用户需要对表的对应目录有写权限。
- 要求文件的格式与表指定的存储格式相同。如创建表时指定stored as rcfile，但是文件格式为txt，则不符合要求。
- 文件必须是HDFS上的文件，不可以用file://的形式指定本地文件系统上的文件。
- 文件名不能以下横线（ _ ）或点（ . ）开头，以这些开头的文件会被忽略。

  如下所示，如果用户test_hive load数据，正确的权限如下：

  ```
  [root@192-168-1-18 duan]# hdfs dfs -ls /tmp/input2
  16/03/21 14:45:07 INFO hdfs.PeerCache: SocketCache disabled.
  Found 1 items
  -rw-r--r--   3 test_hive hive          6 2016-03-21 14:44 /tmp/input2/input.txt
  ```

15.10.12 执行 load data local inpath 命令报错

问题现象

执行load data local inpath报如下错误：

- 错误1：
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have following privileges on Object [type=LOCAL_URI, name=file:/tmp/input/mapdata] for operation LOAD : [SELECT, INSERT, DELETE]
- 错误2：
  HiveAccessControlException Permission denied. Principal [name=user1, type=USER] does not have following privileges on Object [type=LOCAL_URI, name=file:/tmp/input/mapdata] for operation LOAD : [OBJECT OWNERSHIP]
- 错误3：
  SemanticException Line 1:23 Invalid path ''/tmp/input/mapdata'': No files matching path file:/tmp/input/mapdata

原因分析

当前登录的用户不具备操作此目录的权限或者在HiveServer所在节点上没有此目录。

解决方案

说明

通常不建议使用本地文件加载数据到hive表。 建议先将本地文件放入HDFS，然后从集群中加载数据。

Hive对load data local inpath命令有如下权限要求，请对照下述要求是否满足：

- 由于所有的命令都是发送到主HiveServer上去执行的，所以要求此文件在HiveServer节点上。
- HiveServer进程是以操作系统上的omm用户启动的，所以要求omm用户对此文件有读权限，对此文件的目录有读、执行权限。
- 文件的owner需要为执行命令的用户。
- 当前用户需要对该文件有读、写权限。
- 要求文件的格式与表指定的存储格式相同。如创建表时指定stored as rcfile，但是文件格式为txt，则不符合要求。
- 文件名不能以下横线（ _ ）或点（ . ）开头，以这些开头的文件会被忽略。

15.10.13 执行 create external table 报错

问题现象

执行命令：create external table xx(xx int) stored as textfile location '/tmp/aaa/aaa'，报以下错误：

```
Permission denied. Principal [name=fantasy, type=USER] does not have following privileges on Object [type=DFS_URI, name=/tmp/aaa/aaa] for operation CREATETABLE : [SELECT, INSERT, DELETE, OBJECT OWNERSHIP] (state=42000,code=40000)
```

原因分析

当前登录的用户不具备该目录或者其父目录的读写权限。创建外部表时，会判断当前用户对指定的目录以及该目录下其它目录和文件是否有读写权限，如果该目录不存在，会去判断其父目录，依次类推。如果一直不满足就会报权限不足。而不是报指定的目录不存在。

解决方案

请确认当前用户为路径“/tmp/aaa/aaa”的owner有读写权限，如果该路径不存在，确认对其父路径有读写权限。

15.10.14 在 beeline 客户端执行 dfs -put 命令报错

问题现象

执行命令：

dfs -put /opt/kv1.txt /tmp/kv1.txt

报以下错误：

```
Permission denied. Principal [name=admin, type=USER] does not have following privileges onObject[type=COMMAND_PARAMS,name=[-put, /opt/kv1.txt, /tmp/kv1.txt]] for operation DFS : [ADMIN PRIVILEGE] (state=,code=1)
```

原因分析

当前登录的用户不具备操作此命令的权限。

解决方案

如果登录的当前用户具有admin角色，请用set role admin来切换成admin角色操作。
如果不具备admin角色，在Manager页面中给用户绑定对应角色的权限。

15.10.15 执行 set role admin 报无权限

问题现象

执行命令：

set role admin

报下述错误：

```
0: jdbc:hive2://192.168.42.26:21066/> set role admin;
Error: Error while processing statement: FAILED: Execution Error, return code 1 from
org.apache.hadoop.hive.ql.exec.DDLTask. dmp_B doesn't belong to role admin (state=08S01,code=1)
```

原因分析

当前登录的用户不具有Hive的admin角色的权限。

解决方案

步骤1 登录Manager。

- MRS 3.x之前版本，执行**步骤7**。
- MRS 3.x及之后版本，选择“集群 > 服务 > Hive”，在服务“概览”页面右上角单击“更多”，查看“启用Ranger鉴权”是否置灰。
  - 是，执行**步骤2**。
  - 否，执行**步骤7**。

步骤2 选择“集群 > 服务 > Ranger”，单击“基本信息”区域中的“RangerAdmin”，进入Ranger WebUI界面。

步骤3 单击右上角用户名后，选择“Log Out”，退出当前用户后使用**rangeradmin**用户登录。

步骤4 在首页中单击“Settings”，选择“Roles”。

步骤5 单击“Role Name”为“admin”的角色，在“Users”区域，单击“Select User”，选择指定用户名。

步骤6 单击Add Users按钮，在对应用户名所在行勾选“Is Role Admin”，单击“Save”保存配置，操作结束。

步骤7 选择“系统 > 权限 > 角色”，添加一个拥有Hive管理员权限的角色。

步骤8 在FusionInsight Manager页面，选择“系统 > 权限 > 用户 ”。

步骤9 在指定用户对应的“操作”列单击“修改”。

步骤10 为用户绑定拥有Hive管理员权限的角色，并单击“确定”完成权限添加。

----结束

15.10.16 通过 beeline 创建 UDF 时候报错

问题现象

执行命令：

create function fn_test3 as 'test.MyUDF' using jar 'hdfs:///tmp/udf2/MyUDF.jar'

报以下错误：

```
Error: Error while compiling statement: FAILED: HiveAccessControlException Permission denied: Principal [name=admin, type=USER] does not have following privileges for operation CREATEFUNCTION [[ADMIN PRIVILEGE] on Object [type=DATABASE, name=default], [ADMIN PRIVILEGE] on Object [type=FUNCTION, name=default.fn_test3]] (state=42000,code=40000)
```

原因分析

Hive中创建永久函数需要特殊的role admin。

解决方案

在执行语句前执行**set role admin命令**即可解决。

15.10.17 Hive 服务健康状态和 Hive 实例健康状态的区别

问题现象

Hive服务健康状态和Hive实例健康状态的区别是什么？

解决方案

Hive服务的健康状态（也就是在services界面看到的健康状态）有Good，Bad，Partially Healthy，Unknown四种状态，四种状态除了取决于Hive本身服务的可用性（会用简单的sql来检测Hive服务的可用性），还取决于Hive服务所依赖的其他组件的服务状态。

Hive实例分为Hiveserver和Metastore两种，健康状态有Good，Concerning，Unknown三种状态，这三种状态是通过jmx通信来判定，与实例通信正常时为Good，通信异常时为Concerning，无法通信时为Unknown。

15.10.18 Hive 中的告警有哪些以及触发的场景

Hive 中的告警

告警ID	告警级别	可自动清除	告警名称	告警类型
16000	Minor	TRUE	Percentage of Sessions Connected to the HiveServer to Maximum Number Allowed Exceeds the Threshold	故障告警
16001	Minor	TRUE	Hive Warehouse Space Usage Exceeds the Threshold	故障告警
16002	Minor	TRUE	The Successful Hive SQL Operations Lower than The Threshold	故障告警
16004	Critical	TRUE	Hive Service Unavailable	故障告警

告警触发场景

- 16000：当连接HiveServer的session数占允许连接总数的比率超过设定的阈值的时候触发告警。如连接的session数为9，总连接数为12，设定的阈值为70%，9/12 > 70%便触发告警。
- 16001：当Hive使用的HDFS容量占分配给Hive的HDFS总容量的比率超过设定的阈值时触发告警。如分配给Hive的是500G，Hive已经使用400G，设定的阈值时75%，400/500 > 75%便触发告警。
- 16002：当执行SQL的成功率低于设定的阈值时变触发告警。如你执行了4条失败了2条，设定的阈值为60%，成功率2/4 < 60%便触发告警。
- 16004：Hive服务的健康状态变为Bad时触发告警。

说明

- MRS Manager界面操作：告警的阈值和告警的级别以及触发告警的时间段可以在MRS Manager界面的“系统设置 > 阈值配置”中设定。FusionInsight Manager界面操作：告警的阈值和告警的级别以及触发告警的时间段可以在FusionInsight Manager界面的“运维 > 告警 > 阈值设置”中设定。
- Hive运行相关的指标可以在Hive监控界面查看。

15.10.19 Shell 客户端连接提示"authentication failed"

问题现象

安全集群中，HiveServer服务正常的情况下，Shell客户端中执行beeline命令失败，界面提示“authentication failed”，如下：

```
Debug is true storeKey false useTicketCache true useKeyTab false doNotPrompt false ticketCache is null isInitiator true KeyTab is null refreshKrb5Config is false principal is null tryFirstPass is false useFirstPass is false storePass is false clearPass is false
Acquire TGT from Cache
Credentials are no longer valid
Principal is null
null credentials from Ticket Cache
[Krb5LoginModule] authentication failed
No password provided
```

可能原因

- 客户端用户没有进行安全认证
- kerberos认证超期

解决方案

步骤1 登录Hive客户端所在节点。

步骤2 执行**source** *集群客户端安装目录*/**bigdata_env**命令。

可通过**klist**命令查看本地是否有有效票据，如下信息表明票据在16年12月24日14:11:42生效，将在16年12月25日14:11:40失效。在此期间可以使用该票据，其他时间则该票据无效。

```
klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: xxx@HADOOP.COM
Valid starting     Expires            Service principal
12/24/16 14:11:42  12/25/16 14:11:40  krbtgt/HADOOP.COM@HADOOP.COM
```

步骤3 执行**kinit** *username*进行认证，然后再使用客户端。

----结束

15.10.20 客户端提示访问 ZooKeeper 失败

问题现象

安全集群中，HiveServer服务正常的情况下，通过jdbc接口连接HiveServer执行sql时报出ZooKeeper认证异常"The ZooKeeper client is AuthFailed"，如下：

```
14/05/19 10:52:00 WARN utils.HAClientUtilDummyWatcher: The ZooKeeper client is AuthFailed
 14/05/19 10:52:00 INFO utils.HiveHAClientUtil: Exception thrown while reading data from znode.The
```

```
possible reason may be connectionless. This is recoverable. Retrying..
 14/05/19 10:52:16 WARN utils.HAClientUtilDummyWatcher: The ZooKeeper client is AuthFailed
 14/05/19 10:52:32 WARN utils.HAClientUtilDummyWatcher: The ZooKeeper client is AuthFailed
 14/05/19 10:52:32 ERROR st.BasicTestCase: Exception: Could not establish connection to active hiveserver
 java.sql.SQLException: Could not establish connection to active hiveserver
```

或者报出无法读取"Hiveserver2 configs from ZooKeeper"，如下：

```
Exception in thread "main" java.sql.SQLException: org.apache.hive.jdbc.ZooKeeperHiveClientException:
Unable to read HiveServer2 configs from ZooKeeper
at org.apache.hive.jdbc.HiveConnection.<init>(HiveConnection.java:144)
at org.apache.hive.jdbc.HiveDriver.connect(HiveDriver.java:105)
at java.sql.DriverManager.getConnection(DriverManager.java:664)
at java.sql.DriverManager.getConnection(DriverManager.java:247)
at JDBCExample.main(JDBCExample.java:82)
Caused by: org.apache.hive.jdbc.ZooKeeperHiveClientException: Unable to read HiveServer2 configs from
ZooKeeper
at
org.apache.hive.jdbc.ZooKeeperHiveClientHelper.configureConnParams(ZooKeeperHiveClientHelper.java:100)
at org.apache.hive.jdbc.Utils.configureConnParams(Utils.java:509)
at org.apache.hive.jdbc.Utils.parseURL(Utils.java:429)
at org.apache.hive.jdbc.HiveConnection.<init>(HiveConnection.java:142)
... 4 more
Caused by: org.apache.zookeeper.KeeperException$ConnectionLossException: KeeperErrorCode =
ConnectionLoss for /hiveserver2
at org.apache.zookeeper.KeeperException.create(KeeperException.java:99)
at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
at org.apache.zookeeper.ZooKeeper.getChildren(ZooKeeper.java:2374)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl$3.call(GetChildrenBuilderImpl.java:214)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl$3.call(GetChildrenBuilderImpl.java:203)
at org.apache.curator.RetryLo，op.callWithRetry(RetryLoop.java:107)
at
org.apache.curator.framework.imps.GetChildrenBuilderImpl.pathInForeground(GetChildrenBuilderImpl.java:2
00)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl.forPath(GetChildrenBuilderImpl.java:191)
at org.apache.curator.framework.imps.GetChildrenBuilderImpl.forPath(GetChildrenBuilderImpl.java:38)
```

可能原因

- 客户端连接HiveServer时，HiveServer的地址是从ZooKeeper中自动获取，当ZooKeeper连接认证异常时，无法从ZooKeeper中获取正确的HiveServer地址。
- 在连接zookeeper认证时，需要客户端传入krb5.conf，principal，keytab等相关信息。认证失败有如下几种：
  - user.keytab路径写错。
  - user.principal写错。
  - 集群做过切换域名操作但客户端拼接url时使用旧的principal。
  - 有防火墙相关设置，导致客户端本身无法通过kerberos认证，Kerberos需要开放的端口有21730(TCP)、21731(TCP/UDP)、21732(TCP/UDP)。

解决方案

步骤1 确保用户可以正常读取客户端节点相关路径下的user.keytab文件。

步骤2 确保用户的user.principal与指定的keytab文件对应。

可通过**klist -kt keytabpath/user.keytab**查看。

步骤3 如果集群有做过切换域名操作，需要保证url中使用的principal字段是新域名。

如默认为hive/hadoop.hadoop.com@HADOOP.COM，当集群有切换域名的操作时，该字段需要进行相关修改。如域名为abc.com时，则此处应填写hive/hadoop.abc.com@ABC.COM。

步骤4 确保可以正常的认证连接HiveServer。

在客户端执行以下命令

source *客户端安装目录*/**bigdata_env**

kinit *username*

然后再使用客户端执行**beeline**，确保可以正常运行。

----结束

15.10.21 使用 udf 函数提示"Invalid function"

问题现象

在 Hive客户端中使用Spark创建UDF函数时，报出"ERROR 10011","invalid function"的异常，如下：

```
Error: Error while compiling statement: FAILED: SemanticException [Error 10011]: Line 1:7 Invalid function 'test_udf' (state=42000,code=10011)
```

在多个HiveServer之间使用UDF也存在上述问题。例如，在HiveServer1中使用HiverServer2创建的UDF，如果不及时同步元数据信息，连接HiveServer1的客户端也会提示上述错误信息。

可能原因

多个HiveServer之间或者Hive与Spark之间共用的元数据未同步，导致不同HiveServer实例内存数据不一致，造成UDF不生效。

解决方案

需要将新建的UDF信息同步到HiveServer中，执行reload function操作即可。

15.10.22 Hive 服务状态为 Unknown 总结

可能原因

Hive服务停止。

解决方案

重启Hive服务。

15.10.23 Hiveserver 或者 Metastore 实例的健康状态为 unknown

问题现象

hiveserver或者metastore实例的健康状态为unknown。

可能原因

hiveserver或者metastore实例被停止。

解决方案

重启hiveserver或者metastore实例。

15.10.24 Hiveserver 或者 Metastore 实例的健康状态为 Concerning

问题现象

Hiveserver或者Metastore实例的健康状态为Concerning。

可能原因

hiveserver或者metastore实例在启动的时候发生异常，无法正常启动。如，当修改MetaStore/HiveServer GC参数时，可通过查看对应进程的启动日志，如hiveserver.out(hadoop-omm-jar-192-168-1-18.out)文件排查。 如下异常：

Error: Could not find or load main class Xmx2048M

说明java虚拟机启动时，将Xmx2048M 作为java进程的启动参数而不是JVM的启动参数了，如下将符号‘-’误删掉。

```
METASTORE_GC_OPTS=Xms1024M Xmx2048M -DIgnoreReplayReqDetect
-XX\:CMSFullGCsBeforeCompaction\=1 -XX\:+UseConcMarkSweepGC
-XX\:+CMSParallelRemarkEnabled -XX\:+UseCMSCompactAtFullCollection
-XX\:+ExplicitGCInvokesConcurrent -server -XX\:MetaspaceSize\=128M
-XX\:MaxMetaspaceSize\=256M
```

解决方案

因此遇到此类异常应该检查最近的变更项，以确认是否配置有误。

```
METASTORE_GC_OPTS=Xms1024M -Xmx2048M -DIgnoreReplayReqDetect
-XX\:CMSFullGCsBeforeCompaction\=1 -XX\:+UseConcMarkSweepGC
-XX\:+CMSParallelRemarkEnabled -XX\:+UseCMSCompactAtFullCollection
-XX\:+ExplicitGCInvokesConcurrent -server -XX\:MetaspaceSize\=128M
-XX\:MaxMetaspaceSize\=256M
```

15.10.25 TEXTFILE 类型文件使用 ARC4 压缩时 select 结果乱码

问题现象

Hive查询结果表做压缩存储（ARC4），对结果表做select * 查询时返回结果为乱码。

可能原因

Hive默认压缩格式不是ARC4格式或者未开启输出压缩。

解决方案

步骤1 在select结果乱码时，在beeline中进行如下设置。

set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.encryption.arc4.ARC4BlockCodec;

set hive.exec.compress.output=true;

步骤2 使用块解压的方式先将表导入一个新表中。

insert overwrite table tbl_result select * from tbl_source;

步骤3 再进行查询。

select * from tbl_result;

----结束

15.10.26 hive 任务运行过程中失败，重试成功

问题现象

当hive任务在正常运行时失败，在客户端报出错误，类似的错误打印：

```
Error:Invalid OperationHandler:OperationHander [opType=EXECUTE_STATEMENT,getHandleIdentifier()=XXX] (state=,code=0)
```

而此任务提交到yarn上的mapreduce任务运行成功。

```
0: jdbc:hive2://189.120.204.104:21066/> select count(*) from test1;
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1484563934624_0003
INFO  : Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 7 for admin)
INFO  : Kind: HIVE_DELEGATION_TOKEN, Service: HiveServer2ImpersonationToken, Ident: 00 05 61 64 6d 69 6e 05 61 64 6d 69 6e 21 68 69 76 65 2f 68 61 64 6f 6f 70 2e 68
 85 ce e4 8a 01 59 ce 92 52 e4 8e 07 d8 0c
INFO  : The url to track the job: https://189-120-204-104:26001/proxy/application_1484563934624_0003/
INFO  : Starting Job = job_1484563934624_0003, Tracking URL = https://189-120-204-104:26001/proxy/application_1484563934624_0003/
INFO  : Kill Command = /opt/huawei/Bigdata/FusionInsight-Hive-1.1.0/hadoop/bin/hadoop job  -kill job_1484563934624_0003
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2017-01-17 11:46:12,579 Stage-1 map = 0%,  reduce = 0%
INFO  : 2017-01-17 11:46:23,243 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.32 sec
Error: Invalid OperationHandle: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=386323de-dfla-4299-826e-96368d4baf80] (state=,code=0)
0: jdbc:hive2://189.120.204.215:21066/>
```

原因分析

出错的集群有两个hiveserver实例，首先查看其中一个hiveserver日志发现里面的报错与客户端中的错误一样均是Error:Invalid OperationHandler，查看另一个hiveserver发现在出错的时间段此实例有如下类似START_UP的打印，说明那段时间进程被停止过，后来又启动成功，提交的任务本来连接的是重启过的hiveserver实例，当这个实例被停止后，任务进程连接到另一个健康的hiveserver上导致报错。

```
2017-02-15 14:40:11,309 | INFO  | main | STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting HiveServer2
STARTUP_MSG:   host = XXX-120-85-154/XXX.120.85.154
STARTUP_MSG:   args = []
STARTUP_MSG:   version = 1.3.0
```

解决办法

重新提交一次任务即可，保证在任务执行期间不手动重启hiveserver进程。

15.10.27 执行 select 语句报错

问题现象

执行语句select count(*) from XXX;时客户端报错：Error:Error while processing statement :FAILED:Execution Error,return code 2 from ...

这个报错return code2说明是在执行mapreduce任务期间报错导致任务失败。

```
0: jdbc:hive2://134.160.37.21:21066/> select count(*) from src.gn_data_info_gz where day_id='18' and timenap='10';
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:496
INFO  : Submitting tokens for job: job_1482323187492_57815
INFO  : Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 1083948 for boncusermm)
INFO  : Kind: HIVE_DELEGATION_TOKEN, Service: HiveServer2ImpersonationToken, Ident: 00 0a 62 6f 6e 63 75 73 65 72 6d 6d 0a 62 6f 6e 63 75 73 65 72 6d 6d 21 68 69
74 55 8a 01 59 d4 b5 f8 55 8d 02 50 ae 8e 03 65
INFO  : The url to track the job: https://hnocno3:26001/proxy/application_1482323187492_57815/
INFO  : Starting Job = job_1482323187492_57815, Tracking URL = https://hnocno3:26001/proxy/application_1482323187492_57815/
INFO  : Kill Command = /opt/huawei/Bigdata/FusionInsight_V100R002C60U10/FusionInsight-Hive-1.3.0/hive-1.3.0/bin/../../hadoop/bin/hadoop job  -kill job_148232318
INFO  : Hadoop job information for Stage-1: number of mappers: 496; number of reducers: 1
INFO  : 2017-01-18 16:21:00,906 Stage-1 map = 0%,  reduce = 0%
INFO  : 2017-01-18 16:21:18,957 Stage-1 map = 1%,  reduce = 0%, Cumulative CPU 50.53 sec
INFO  : 2017-01-18 16:21:32,526 Stage-1 map = 2%,  reduce = 0%, Cumulative CPU 416.23 sec
INFO  : 2017-01-18 16:21:35,035 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 1421.09 sec
INFO  : 2017-01-18 16:21:36,331 Stage-1 map = 7%,  reduce = 0%, Cumulative CPU 2159.35 sec
INFO  : 2017-01-18 16:21:37,810 Stage-1 map = 9%,  reduce = 0%, Cumulative CPU 2548.77 sec
INFO  : 2017-01-18 16:21:39,126 Stage-1 map = 15%,  reduce = 0%, Cumulative CPU 3264.95 sec
INFO  : 2017-01-18 16:21:40,509 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 3621.79 sec
INFO  : 2017-01-18 16:21:41,710 Stage-1 map = 26%,  reduce = 0%, Cumulative CPU 3913.79 sec
INFO  : 2017-01-18 16:21:42,890 Stage-1 map = 32%,  reduce = 0%, Cumulative CPU 4202.18 sec
INFO  : 2017-01-18 16:21:44,037 Stage-1 map = 41%,  reduce = 0%, Cumulative CPU 4595.63 sec
INFO  : 2017-01-18 16:21:45,119 Stage-1 map = 49%,  reduce = 0%, Cumulative CPU 4822.15 sec
INFO  : 2017-01-18 16:21:46,213 Stage-1 map = 57%,  reduce = 0%, Cumulative CPU 5107.44 sec
INFO  : 2017-01-18 16:21:47,300 Stage-1 map = 68%,  reduce = 0%, Cumulative CPU 5405.71 sec
INFO  : 2017-01-18 16:21:48,407 Stage-1 map = 76%,  reduce = 0%, Cumulative CPU 5611.73 sec
INFO  : 2017-01-18 16:21:49,483 Stage-1 map = 85%,  reduce = 0%, Cumulative CPU 5804.64 sec
INFO  : 2017-01-18 16:21:50,565 Stage-1 map = 92%,  reduce = 0%, Cumulative CPU 5958.81 sec
INFO  : 2017-01-18 16:21:51,641 Stage-1 map = 96%,  reduce = 0%, Cumulative CPU 6041.06 sec
INFO  : 2017-01-18 16:21:52,744 Stage-1 map = 98%,  reduce = 0%, Cumulative CPU 6073.82 sec
INFO  : 2017-01-18 16:22:08.352 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6078.4 sec
INFO  : MapReduce Total cumulative CPU time: 0 days 1 hours 41 minutes 18 seconds 400 msec
ERROR : Ended Job = job_1482323187492_57815 with errors
Error: Error while processing statement: FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask (state=08S01,code=2)
0: jdbc:hive2://134.160.37.21:21066/>
```

原因分析

1. 进入yarn原生页面查看mapreduce任务的日志看到报错是无法识别到压缩方式导致错误，看文件后缀是gzip压缩，堆栈却报出是zlib方式。

```
2017-01-18 16:22:07,566 INFO [main] org.apache.hadoop.hive.ql.exec.MapOperator: 4 Close done
2017-01-18 16:22:07,572 WARN [main] org.apache.hadoop.mapred.YarnChild: Exception running child : java.io.IOException: java.io.IOException: unknown compression method
        at org.apache.hadoop.hive.io.HiveIOExceptionHandlerChain.handleRecordReaderNextException(HiveIOExceptionHandlerChain.java:121)
        at org.apache.hadoop.hive.io.HiveIOExceptionHandlerUtil.handleRecordReaderNextException(HiveIOExceptionHandlerUtil.java:77)
        at org.apache.hadoop.hive.ql.io.HiveContextAwareRecordReader.doNext(HiveContextAwareRecordReader.java:355)
        at org.apache.hadoop.hive.ql.io.HiveRecordReader.doNext(HiveRecordReader.java:79)
        at org.apache.hadoop.hive.ql.io.HiveRecordReader.doNext(HiveRecordReader.java:33)
        at org.apache.hadoop.hive.ql.io.HiveContextAwareRecordReader.next(HiveContextAwareRecordReader.java:116)
        at org.apache.hadoop.mapred.MapTask$TrackedRecordReader.moveToNext(MapTask.java:199)
        at org.apache.hadoop.mapred.MapTask$TrackedRecordReader.next(MapTask.java:185)
        at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:52)
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:453)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:180)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1726)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:174)
Caused by: java.io.IOException: unknown compression method
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.inflateBytesDirect(Native Method)
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.decompress(ZlibDecompressor.java:225)
        at org.apache.hadoop.io.compress.DecompressorStream.decompress(DecompressorStream.java:91)
        at org.apache.hadoop.io.compress.DecompressorStream.read(DecompressorStream.java:85)
        at java.io.InputStream.read(InputStream.java:101)
        at org.apache.hadoop.util.LineReader.fillBuffer(LineReader.java:180)
        at org.apache.hadoop.util.LineReader.readDefaultLine(LineReader.java:216)
        at org.apache.hadoop.util.LineReader.readLine(LineReader.java:174)
        at org.apache.hadoop.mapred.LineRecordReader.next(LineRecordReader.java:248)
        at org.apache.hadoop.mapred.LineRecordReader.next(LineRecordReader.java:48)
        at org.apache.hadoop.hive.ql.io.HiveContextAwareRecordReader.doNext(HiveContextAwareRecordReader.java:350)
        ... 13 more

2017-01-18 16:22:07,576 INFO [main] org.apache.hadoop.mapred.Task: Runnning cleanup for the task
```

2. 因此怀疑此语句查询的表对应的HDFS上的文件有问题，map日志中打印出了解析的对应的文件名，将其从HDFS上下载到本地，看到是gz结尾的文件，使用**tar**命令解压报错，格式不正确无法解压。使用file命令查看文件属性发现此文件来自于FAT系统的压缩而非unix。

```
[root@hn0dn01 ~]# ls -l *.txt.gz
-rw-r--r-- 1 root root 101966463 Jan 18 20:13 20170118095958920074010l.txt.gz
-rw-r--r-- 1 root root  90448283 Jan 18 19:55 20170118104000000074002O.txt.gz
[root@hn0dn01 ~]# file 20170118095958920074010l.txt.gz
20170118095958920074010l.txt.gz: gzip compressed data, was "20170118095958920074010l.txt", from Unix, last modified: Wed Jan 18 09:59:52 2017
[root@hn0dn01 ~]# file  20170118104000000074002O.txt.gz
20170118104000000074002O.txt.gz: gzip compressed data, from FAT filesystem (MS-DOS, OS/2, NT)
[root@hn0dn01 ~]# tar -zxvf  20170118104000000074002O.txt.gz
tar: This does not look like a tar archive
tar: Skipping to next header

gzip: stdin: decompression OK, trailing garbage ignored
tar: Child returned status 2
tar: Error is not recoverable: exiting now
[root@hn0dn01 ~]#
```

解决办法

将格式不正确的文件移除hdfs目录或者替换为正确的格式的文件。

15.10.28 drop partition 操作，有大量分区时操作失败

问题背景与现象

执行drop partitions 操作，执行异常：

```
MetaStoreClient lost connection. Attempting to reconnect. |
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.invoke(RetryingMetaStoreClient.java:187)
org.apache.thrift.transport.TTransportException
at org.apache.thrift.transport.TIOStreamTransport.read(TIOStreamTransport.java:132)
at org.apache.thrift.transport.TTransport.xxx(TTransport.java:86)
at org.apache.thrift.transport.TSaslTransport.readLength(TSaslTransport.java:376)
at org.apache.thrift.transport.TSaslTransport.readFrame(TSaslTransport.java:453)
at org.apache.thrift.transport.TSaslTransport.read(TSaslTransport.java:435)
...
```

查看对应metaStore日志，有StackOverFlow异常

```
2017-04-22 01:00:58,834 | ERROR | pool-6-thread-208 | java.lang.StackOverflowError
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:330)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
at org.datanucleus.store.rdbms.sql.SQLText.toSQL(SQLText.java:339)
```

原因分析

drop partition的处理逻辑是将找到所有满足条件的分区，将其拼接起来，最后统一删除。由于分区数过多，拼删元数据堆栈较深，出现StackOverFlow异常。

解决办法

分批次删除分区。

15.10.29 localtask 启动失败

问题背景与现象

1. 执行join等操作，数据量较小时，会启动localtask执行，执行过程会报错：

   ```
   jdbc:hive2://10.*.*.*:21066/> select a.name ,b.gender from student a join student1 b on (a.name =
   b.name);
   ERROR : Execution failed with exit status: 1
   ERROR : Obtaining error information
   ERROR :
   Task failed!
   Task ID:
     Stage-4
   ...
   Error: Error while processing statement: FAILED: Execution Error, return code 1 from
   org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask (state=08S01,code=1)
   ...
   ```

2. 查看对应hiveserver日志，发现是启动localtask失败

   ```
   2018-04-25 16:37:19,296 | ERROR | HiveServer2-Background-Pool: Thread-79 | Execution failed with
   exit status: 1 | org.apache.hadoop.hive.ql.session.SessionState
   $LogHelper.printError(SessionState.java:1016)
   2018-04-25 16:37:19,296 | ERROR | HiveServer2-Background-Pool: Thread-79 | Obtaining error
   information | org.apache.hadoop.hive.ql.session.SessionState
   $LogHelper.printError(SessionState.java:1016)
   2018-04-25 16:37:19,297 | ERROR | HiveServer2-Background-Pool: Thread-79 |
   Task failed!
   Task ID:
     Stage-4
   ```

```
Logs:
 | org.apache.hadoop.hive.ql.session.SessionState$LogHelper.printError(SessionState.java:1016)
2018-04-25 16:37:19,297 | ERROR | HiveServer2-Background-Pool: Thread-79 | /var/log/Bigdata/hive/
hiveserver/hive.log | org.apache.hadoop.hive.ql.session.SessionState
$LogHelper.printError(SessionState.java:1016)
2018-04-25 16:37:19,297 | ERROR | HiveServer2-Background-Pool: Thread-79 | Execution failed with
exit status: 1 |
org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask.executeInChildVM(MapredLocalTask.java:342)
2018-04-25 16:37:19,309 | ERROR | HiveServer2-Background-Pool: Thread-79 | FAILED: Execution
Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask |
org.apache.hadoop.hive.ql.session.SessionState$LogHelper.printError(SessionState.java:1016)
...
2018-04-25 16:37:36,438 | ERROR | HiveServer2-Background-Pool: Thread-88 | Error running hive
query:  | org.apache.hive.service.cli.operation.SQLOperation$1$1.run(SQLOperation.java:248)
org.apache.hive.service.cli.HiveSQLException: Error while processing statement: FAILED: Execution
Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask
        at org.apache.hive.service.cli.operation.Operation.toSQLException(Operation.java:339)
        at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:169)
        at org.apache.hive.service.cli.operation.SQLOperation.access$200(SQLOperation.java:75)
        at org.apache.hive.service.cli.operation.SQLOperation$1$1.run(SQLOperation.java:245)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1710)
        at org.apache.hive.service.cli.operation.SQLOperation$1.run(SQLOperation.java:258)
        at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)
```

3. 查看对应hiveserver日志目录(/var/log/Bigdata/hive/hiveserver)下，hs_err_pid_*****.log，发现有内存不够的错误

```
# There is insufficient memory for the Java Runtime Environment to continue.
# Native memory allocation (mmap) failed to map 20776943616 bytes for committing reserved
memory.
 ...
```

原因分析

Hive在执行join操作，数据量小时会生成MapJoin，执行MapJoin时会生成localtask任务，localtask启动的jvm内存继承了父进程的内存。

当有多个join执行的时候，启动多个localtask，如果机器内存不够，就会导致启动localtask失败。

解决办法

步骤1 进入Hive服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 搜索“hive.auto.convert.join”参数并修改hive的配置hive.auto.convert.join 为false，保存配置并重启服务。

该参数修改后会对业务性能有一定影响。继续执行后续步骤可不影响业务性能。

步骤3 搜索“HIVE_GC_OPTS”参数并修改 hive的HIVE_GC_OPTS，把Xms调小，具体要根据业务评估，最小设置为Xmx的一半，修改完后保存配置并重启服务。

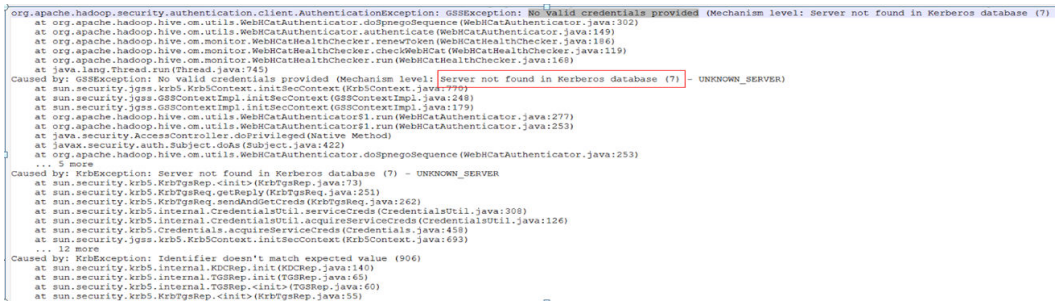
----结束

15.10.30 WebHCat 启动失败

问题背景与现象

用户修改hostname导致WebHCat启动失败。

查看对应节点WebHCat启动日志（ /var/log/Bigdata/hive/webhcat/hive.log ），发现报如下错误：



原因分析

1. MRSwebhcat角色的服务端账户中涉及到hostname，如果安装完后再修改hostname，就会导致启动失败。
2. /etc/hosts中配置了一对多或者多对一的主机名和IP对应关系，导致在执行**hostname**和**hostname -i**获取不到正确的IP和hostname。

解决办法

步骤1 将修改了节点的hostname全部修改为集群安装前的hostname。

步骤2 排查WebHCat所在节点的/etc/hosts是否配置正确。

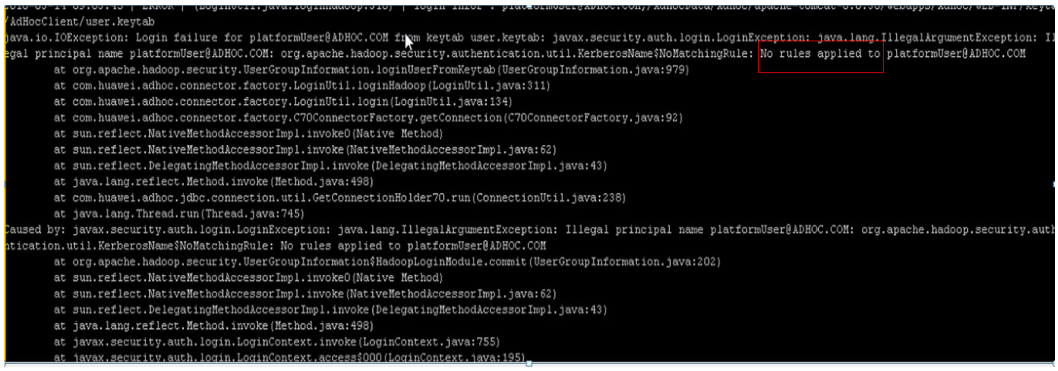
步骤3 重启WebHCat。

----结束

15.10.31 切域后 Hive 二次开发样例代码报错

问题背景与现象

hive的二次开发代码样例运行报No rules applied to ****的错误：



原因分析

1. hive的二次开发样例代码会加载core-site.xml，此文件默认是通过classload加载，所以使用的时候要把此配置文件放到启动程序的classpath路径下面。
2. 如果修改了集群的域名，那么core-site.xml将发生变化，需要下载最新的core-site.xml并放入到打包hive二次开发样例代码进程的classpath路径下面。

解决办法

步骤1 下载集群Hive最新的客户端，获取最新的core-site.xml。

步骤2 将core-site.xml放入到打包hive二次开发样例代码进程的classpath路径下面。

----结束

15.10.32 DBService 超过最大连接数，导致 metastore 异常

问题背景与现象

DBService默认最大连接数是300，如果当业务量比较大，导致连接DBService的最大连接数超过300时，metastore会出现异常，并报slots are reserved for non-replication superuser connections的错误：

```
2018-04-26 14:58:55,657 | ERROR | BoneCP-pool-watch-thread | Failed to acquire connection to jdbc:postgresql://10.*.*.*:20051/hivemeta?socketTimeout=60. Sleeping for 1000 ms. Attempts left: 9 | com.jolbox.bonecp.BoneCP.obtainInternalConnection(BoneCP.java:292)
org.postgresql.util.PSQLException: FATAL: remaining connection slots are reserved for non-replication superuser connections
    at org.postgresql.core.v3.ConnectionFactoryImpl.readStartupMessages(ConnectionFactoryImpl.java:643)
    at org.postgresql.core.v3.ConnectionFactoryImpl.openConnectionImpl(ConnectionFactoryImpl.java:184)
    at org.postgresql.core.ConnectionFactory.openConnection(ConnectionFactory.java:64)
    at org.postgresql.jdbc2.AbstractJdbc2Connection.<init>(AbstractJdbc2Connection.java:124)
    at org.postgresql.jdbc3.AbstractJdbc3Connection.<init>(AbstractJdbc3Connection.java:28)
    at org.postgresql.jdbc3g.AbstractJdbc3gConnection.<init>(AbstractJdbc3gConnection.java:20)
    at org.postgresql.jdbc4.AbstractJdbc4Connection.<init>(AbstractJdbc4Connection.java:30)
    at org.postgresql.jdbc4.Jdbc4Connection.<init>(Jdbc4Connection.java:22)
    at org.postgresql.Driver.makeConnection(Driver.java:392)
    at org.postgresql.Driver.connect(Driver.java:266)
    at java.sql.DriverManager.getConnection(DriverManager.java:664)
    at java.sql.DriverManager.getConnection(DriverManager.java:208)
    at com.jolbox.bonecp.BoneCP.obtainRawInternalConnection(BoneCP.java:361)
    at com.jolbox.bonecp.BoneCP.obtainInternalConnection(BoneCP.java:269)
    at com.jolbox.bonecp.ConnectionHandle.<init>(ConnectionHandle.java:242)
    at com.jolbox.bonecp.PoolWatchThread.fillConnections(PoolWatchThread.java:115)
    at com.jolbox.bonecp.PoolWatchThread.run(PoolWatchThread.java:82)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)
```

原因分析

业务量大导致连接DBService的最大连接数超过了300，需要修改DBService的最大连接数。

解决办法

步骤1 进入DBService服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > DBService > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > DBService > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 > DBService > 配置 > 全部配置”。

步骤2 搜索dbservice.database.max.connections配置项，并修改dbservice.database.max.connections配置的值到合适值，不能超过1000。

步骤3 保存配置，并重启受影响的服务或者实例。

步骤4 如果调整完还报超过最大连接数，需要排查业务代码，是否有连接泄露。

----结束

15.10.33 beeline 报 Failed to execute session hooks: over max connections 错误

问题背景与现象

HiveServer连接的最大连接数默认为200，当超过200时，beeline会报Failed to execute session hooks: over max connections

```
beeline> [root@172-27-16-38 c70client]# beeline
Connecting to
jdbc:hive2://129.188.82.38:24002,129.188.82.36:24002,129.188.82.35:24002/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2;sasl.qop=auth-conf;auth=KERBEROS;principal=hive/hadoop.hadoop.com@HADOOP.COM
Debug is  true storeKey false useTicketCache true useKeyTab false doNotPrompt false ticketCache is null isInitiator true KeyTab is null refreshKrb5Config is false principal is null tryFirstPass is false useFirstPass is false storePass is false clearPass is false
Acquire TGT from Cache
Principal is xxx@HADOOP.COM
Commit Succeeded

Error: Failed to execute session hooks: over max connections. (state=,code=0)
Beeline version 1.2.1 by Apache Hive
```

查看hiveserver日志(/var/log/Bigdata/hive/hiveserver/hive.log)报over max connections错误

```
2018-05-03 04:31:56,728 | WARN  | HiveServer2-Handler-Pool: Thread-137 | Error opening session:  | org.apache.hive.service.cli.thrift.ThriftCLIService.OpenSession(ThriftCLIService.java:542)
org.apache.hive.service.cli.HiveSQLException: Failed to execute session hooks: over max connections.
```

```
    at org.apache.hive.service.cli.session.SessionManager.openSession(SessionManager.java:322)
    at org.apache.hive.service.cli.CLIService.openSessionWithImpersonation(CLIService.java:189)
    at org.apache.hive.service.cli.thrift.ThriftCLIService.getSessionHandle(ThriftCLIService.java:663)
    at org.apache.hive.service.cli.thrift.ThriftCLIService.OpenSession(ThriftCLIService.java:527)
    at org.apache.hive.service.cli.thrift.TCLIService$Processor$OpenSession.getResult(TCLIService.java:1257)
    at org.apache.hive.service.cli.thrift.TCLIService$Processor$OpenSession.getResult(TCLIService.java:1242)
    at org.apache.thrift.ProcessFunction.process(ProcessFunction.java:39)
    at org.apache.thrift.TBaseProcessor.process(TBaseProcessor.java:39)
    at org.apache.hadoop.hive.thrift.HadoopThriftAuthBridge$Server
$TUGIAssumingProcessor.process(HadoopThriftAuthBridge.java:710)
    at org.apache.thrift.server.TThreadPoolServer$WorkerProcess.run(TThreadPoolServer.java:286)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)
Caused by: org.apache.hive.service.cli.HiveSQLException: over max connections.
    at
org.apache.hadoop.hive.transporthook.SessionControllerTsaslTransportHook.checkTotalSessionNumber(Sessi
onControllerTsaslTransportHook.java:208)
    at
org.apache.hadoop.hive.transporthook.SessionControllerTsaslTransportHook.postOpen(SessionControllerTsasl
TransportHook.java:163)
    at
org.apache.hadoop.hive.transporthook.SessionControllerTsaslTransportHook.run(SessionControllerTsaslTransp
ortHook.java:134)
    at org.apache.hive.service.cli.session.SessionManager.executeSessionHooks(SessionManager.java:432)
    at org.apache.hive.service.cli.session.SessionManager.openSession(SessionManager.java:314)
    ... 12 more
```

原因分析

业务量大导致连接HiveServer单个节点最大连接数超过了200，需要调大连接HiveServer实例的最大连接数。

解决办法

步骤1 进入Hive服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 搜索hive.server.session.control.maxconnections配置项，并修改hive.server.session.control.maxconnections配置的值到合适值，不能超过1000。

步骤3 保存配置并重启受影响的服务或者实例。

----结束

15.10.34 beeline 报 OutOfMemoryError 错误

问题背景与现象

beeline客户端查询大量数据时，报OutOFMemoryError：Java heap space，具体报错信息如下：

```
org.apache.thrift.TException: Error in calling method FetchResults
    at org.apache.hive.jdbc.HiveConnection$SynchronizedHandler.invoke(HiveConnection.java:1514)
    at com.sun.proxy.$Proxy4.FetchResults(Unknown Source)
    at org.apache.hive.jdbc.HiveQueryResultSet.next(HiveQueryResultSet.java:358)
    at org.apache.hive.beeline.BufferedRows.<init>(BufferedRows.java:42)
    at org.apache.hive.beeline.BeeLine.print(BeeLine.java:1856)
    at org.apache.hive.beeline.Commands.execute(Commands.java:873)
    at org.apache.hive.beeline.Commands.sql(Commands.java:714)
    at org.apache.hive.beeline.BeeLine.dispatch(BeeLine.java:1035)
    at org.apache.hive.beeline.BeeLine.execute(BeeLine.java:821)
    at org.apache.hive.beeline.BeeLine.begin(BeeLine.java:778)
    at org.apache.hive.beeline.BeeLine.mainWithInputRedirection(BeeLine.java:486)
    at org.apache.hive.beeline.BeeLine.main(BeeLine.java:469)
Caused by: java.lang.OutOfMemoryError: Java heap space
    at com.sun.crypto.provider.CipherCore.doFinal(CipherCore.java:959)
    at com.sun.crypto.provider.CipherCore.doFinal(CipherCore.java:824)
    at com.sun.crypto.provider.AESCipher.engineDoFinal(AESCipher.java:436)
    at javax.crypto.Cipher.doFinal(Cipher.java:2223)
    at sun.security.krb5.internal.crypto.dk.AesDkCrypto.decryptCTS(AesDkCrypto.java:414)
    at sun.security.krb5.internal.crypto.dk.AesDkCrypto.decryptRaw(AesDkCrypto.java:291)
    at sun.security.krb5.internal.crypto.Aes256.decryptRaw(Aes256.java:86)
    at sun.security.jgss.krb5.CipherHelper.aes256Decrypt(CipherHelper.java:1397)
    at sun.security.jgss.krb5.CipherHelper.decryptData(CipherHelper.java:576)
    at sun.security.jgss.krb5.WrapToken_v2.getData(WrapToken_v2.java:130)
    at sun.security.jgss.krb5.WrapToken_v2.getData(WrapToken_v2.java:105)
    at sun.security.jgss.krb5.Krb5Context.unwrap(Krb5Context.java:1058)
    at sun.security.jgss.GSSContextImpl.unwrap(GSSContextImpl.java:403)
    at com.sun.security.sasl.gsskerb.GssKrb5Base.unwrap(GssKrb5Base.java:77)
    at org.apache.thrift.transport.TSaslTransport$SaslParticipant.unwrap(TSaslTransport.java:559)
    at org.apache.thrift.transport.TSaslTransport.readFrame(TSaslTransport.java:462)
    at org.apache.thrift.transport.TSaslTransport.read(TSaslTransport.java:435)
    at org.apache.thrift.transport.TSaslClientTransport.read(TSaslClientTransport.java:37)
    at org.apache.thrift.transport.TTransport.xxx(TTransport.java:86)
    at org.apache.hadoop.hive.thrift.TFilterTransport.xxx(TFilterTransport.java:62)
    at org.apache.thrift.protocol.TBinaryProtocol.xxx(TBinaryProtocol.java:429)
    at org.apache.thrift.protocol.TBinaryProtocol.readI32(TBinaryProtocol.java:318)
    at org.apache.thrift.protocol.TBinaryProtocol.readMessageBegin(TBinaryProtocol.java:219)
    at org.apache.thrift.TServiceClient.receiveBase(TServiceClient.java:77)
    at org.apache.hive.service.cli.thrift.TCLIService$Client.recv_FetchResults(TCLIService.java:505)
    at org.apache.hive.service.cli.thrift.TCLIService$Client.FetchResults(TCLIService.java:492)
    at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.apache.hive.jdbc.HiveConnection$SynchronizedHandler.invoke(HiveConnection.java:1506)
    at com.sun.proxy.$Proxy4.FetchResults(Unknown Source)
    at org.apache.hive.jdbc.HiveQueryResultSet.next(HiveQueryResultSet.java:358)
Error: Error retrieving next row (state=,code=0)
```

原因分析

- 客户查询大量数据，数据量过大。
- 客户在检索数据时使用**select * from table_name;**，进行全表查询，表内数据过多。
- beeline默认启动内存128M，查询时返回结果集过大，导致beeline无法承载导致。

解决办法

步骤1 执行**select count(*) from table_name;**前确认需要查询的数据量大小，确认是否需要在beeline中显示如此数量级的数据。

步骤2 如数量在一定范围内需要显示，请调整hive客户端的jvm参数， 在hive客户端目录/Hive下的component_env中添加export HIVE_OPTS=-Xmx1024M(具体数值请根据业务调整)，并重新执行**source** *客户端目录*/**bigdata_env**配置环境变量。

----结束

15.10.35 输入文件数超出设置限制导致任务执行失败

问题背景与现象

Hive执行查询操作时报Job Submission failed with exception 'java.lang.RuntimeException(input file number exceeded the limits in the conf;input file num is: 2380435,max heap memory is: 16892035072,the limit conf is: 500000/4)'，此报错中具体数值根据实际情况会发生变化，具体报错信息如下:

```
ERROR : Job Submission failed with exception 'java.lang.RuntimeException(input file numbers exceeded the limits in the conf;
 input file num is: 2380435 ,
 max heap memory is: 16892035072 ,
 the limit conf is: 500000/4)'
java.lang.RuntimeException: input file numbers exceeded the limits in the conf;
 input file num is: 2380435 ,
 max heap memory is: 16892035072 ,
 the limit conf is: 500000/4
    at org.apache.hadoop.hive.ql.exec.mr.ExecDriver.checkFileNum(ExecDriver.java:545)
    at org.apache.hadoop.hive.ql.exec.mr.ExecDriver.execute(ExecDriver.java:430)
    at org.apache.hadoop.hive.ql.exec.mr.MapRedTask.execute(MapRedTask.java:137)
    at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:158)
    at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:101)
    at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:1965)
    at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:1723)
    at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1475)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1283)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1278)
    at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:167)
    at org.apache.hive.service.cli.operation.SQLOperation.access$200(SQLOperation.java:75)
    at org.apache.hive.service.cli.operation.SQLOperation$1$1.run(SQLOperation.java:245)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1710)
    at org.apache.hive.service.cli.operation.SQLOperation$1.run(SQLOperation.java:258)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)

Error: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask (state=08S01,code=1)
```

原因分析

MapReduce 任务提交前对输入文件数的检查策略：在提交的MapReduce 任务中，允许的最大输入文件数和HiveServer最大堆内存的比值，例如500000/4（默认值），表示每4GB堆内存最大允许500000个输入文件。在输入的文件数超出此限制时则会发生此错误。

解决办法

步骤1 进入Hive服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名称* > 服务 > Hive > 配置 > 全部配置”。

----结束

步骤1 搜索hive.mapreduce.input.files2memory配置项，并修改hive.mapreduce.input.files2memory配置的值到合适值，根据实际内存和任务情况对此值进行调整。

步骤2 保存配置并重启受影响的服务或者实例。

步骤3 如调整后问题仍未解决，请根据业务情况调整HiveServer的GC参数至合理的值。

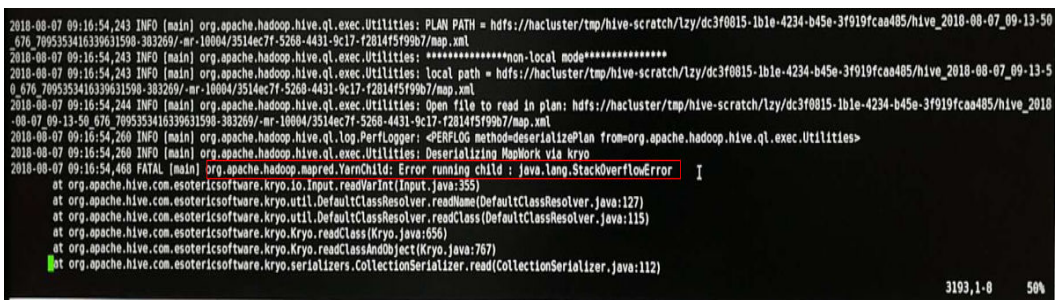
----结束

15.10.36 任务执行中报栈内存溢出导致任务执行失败

问题背景与现象

Hive执行查询操作时报错**Error running child : java.lang.StackOverflowError**，具体报错信息如下：

```
FATAL [main] org.apache.hadoop.mapred.YarnChild: Error running child : java.lang.StackOverflowError
at org.apache.hive.come.esotericsoftware.kryo.io.Input.readVarInt(Input.java:355)
at
org.apache.hive.come.esotericsoftware.kryo.util.DefautClassResolver.readName(DefautClassResolver.java:127)
at
org.apache.hive.come.esotericsoftware.kryo.util.DefautClassResolver.readClass(DefautClassResolver.java:115)
at org.apache.hive.come.esotericsoftware.kryo.Kryo.readClass(Kryo.java.656)
at org.apache.hive.come.esotericsoftware.kryo.kryo.readClassAnd0bject(Kryo.java:767)
at
org.apache.hive.come.esotericsoftware.kryo.serializers.collectionSerializer.read(CollectionSerializer.java:112)
```



原因分析

java.lang.StackOverflowError这是内存溢出错误的一种，即线程栈的溢出，方法调用层次过多（比如存在无限递归调用）或线程栈太小都会导致此报错。

解决办法

通过调整mapreduce阶段的map和reduce子进程JVM参数中的栈内存解决此问题，主要涉及参数为mapreduce.map.java.opts（调整map的栈内存）和mapreduce.reduce.java.opts（调整reduce的栈内存），调整方法如下（以mapreduce.map.java.opts参数为例）。

- 临时增加map内存（只针对此次beeline生效）：

  在beeline中执行如下命令**set mapreduce.map.java.opts=-Xss8G;**（具体数值请结合实际业务情况进行调整）。
- 永久增加map内存mapreduce.map.memory.mb和mapreduce.map.java.opts的值：

  a. 进入Hive服务配置页面：

  - MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
  - MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

    说明

    若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
  - MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名称* > 服务 > Hive > 配置 > 全部配置”。

  b. 在hiveserver自定义参数界面添加自定义参数mapreduce.map.java.opts及相应的值。

  c. 保存配置并重启受影响的服务或者实例。

  修改配置后需要保存，请注意参数在HiveServer自定义参数处修改，保存重启后生效（重启期间hive服务不可用），请注意执行时间窗口。

15.10.37 对同一张表或分区并发写数据导致任务失败

问题背景与现象

Hive执行插入语句时，报错HDFS上文件或目录已存在或被清除，具体报错如下：

```
2019-03-18 14:34:23,016 | WARN  | HiveServer2-Background-Pool: Thread-1179606 | Failed to move to trash: hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f/000000_0; Force to delete it. | org.apache.hadoop.hive.common.FileUtils.moveToTrash(FileUtils.java:651)
2019-03-18 14:34:23,017 | INFO  | HiveServer2-Background-Pool: Thread-1179604 | Moved to trash: hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp6_f/000000_0 | org.apache.hadoop.hive.common.FileUtils.moveToTrash(FileUtils.java:644)
2019-03-18 14:34:23,017 | ERROR | HiveServer2-Background-Pool: Thread-1179606 | Failed to delete hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f/000000_0 | org.apache.hadoop.hive.common.FileUtils.moveToTrash(FileUtils.java:660)
2019-03-18 14:34:23,017 | ERROR | HiveServer2-Background-Pool: Thread-1179606 | Failed with exception Destination directory hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f has not be cleaned up.
org.apache.hadoop.hive.ql.metadata.HiveException: Destination directory hdfs://hacluster/user/hive/warehouse/frpdb.db/dw_fixed_cost_xn_temp5_f has not be cleaned up.
at org.apache.hadoop.hive.ql.metadata.Hive.replaceFiles(Hive.java:2974)
at org.apache.hadoop.hive.ql.metadata.Hive.loadTable(Hive.java:1664)
at org.apache.hadoop.hive.ql.exec.MoveTask.execute(MoveTask.java:374)
at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:158)
at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:101)
```

原因分析

1. 根据HiveServer的审计日志，确认该任务的开始时间和结束时间。

2. 在上述时间区间内，查找是否有对同一张表或分区进行插入数据的操作。
3. Hive不支持对同一张表或分区进行并发数据插入，这样会导致多个任务操作同一个数据临时目录，一个任务将另一个任务的数据移走，导致任务失败。

解决办法

修改业务逻辑，单线程插入数据到同一张表或分区。

15.10.38 Hive 任务失败，报没有 HDFS 目录的权限

问题背景与现象

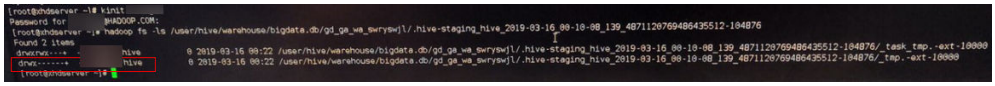
Hive任务报错，提示执行用户没有HDFS目录权限

```
2019-04-09 17:49:19,845 | ERROR | HiveServer2-Background-Pool: Thread-3160445 | Job Submission failed
with exception 'org.apache.hadoop.security.AccessControlException(Permission denied: user=hive_quanxian,
access=READ_EXECUTE, inode="/user/hive/warehouse/bigdata.db/
gd_ga_wa_swryswjl":zhongao:hive:drwx------
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkAccessAcl(FSPermissionChecker.java:426
)
at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:329)
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkSubAccess(FSPermissionChecker.java:30
0)
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:24
1)
at
com.xxx.hadoop.adapter.hdfs.plugin.HWAccessControlEnforce.checkPermission(HWAccessControlEnforce.java:
69)
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:19
0)
at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1910)
at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1894)
at
org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getContentSummary(FSDirStatAndListingO
p.java:135)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getContentSummary(FSNamesystem.java:3983)
at
org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getContentSummary(NameNodeRpcServer.ja
va:1342)
at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getContentSummary(Cli
entNamenodeProtocolServerSideTranslatorPB.java:925)
at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol
$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:616)
at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:973)
at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2260)
at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2256)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1781)
at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2254)
)'
```

原因分析

1. 根据堆栈信息，可以看出在检查子目录的权限时失败。
   ```
   org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkSubAccess(FSPermissionChecker.java:300)
   ```

2. 检查HDFS上表目录下所有文件目录的权限，发现有一个目录权限为700（只有文件属主能够访问），确认存在异常目录。



解决办法

1. 确认该文件是否为手动异常导入，如不是数据文件或目录，删除该文件或目录。
2. 当无法删除时，建议修改文件或目录权限为770。

15.10.39 Load 数据到 Hive 表失败

问题背景与现象

用户在建表成功后，通过Load命令往此表导入数据，但导入操作中遇到如下问题：

```
.....
> LOAD DATA INPATH '/user/tester1/hive-data/data.txt' INTO TABLE employees_info;
Error: Error while compiling statement: FAILED: SemanticException Unable to load data to destination table.
Error: The file that you are trying to load does not match the file format of the destination table.
(state=42000,code=40000)
.....
```

原因分析

1. 经分析，发现在建表时没有指定存储格式，所以采用了缺省存储格式RCFile。
2. 在导入数据时，被导入数据格式是TEXTFILE格式，最终导致此问题。

解决办法

属于应用侧问题，解决办法有多种。只要保证表所指定存储格式和被导入数据格式是一致的，可以根据实际情况采用合适方法。

- 方法1：

  可以使用具有Hive表操作权限的用户在建表时指定存储格式，例如：

  CREATE TABLE IF NOT EXISTS employees_info(name STRING,age INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

  指定被导入数据格式为TEXTFILE。
- 方法2：

  被导入数据存储格式不能为TEXTFILE，而应为RCFile。

15.10.40 HiveServer 和 HiveHCat 进程故障

用户问题

客户集群HiveServer和WebHCat进程状态均为故障。

问题现象

客户MRS集群Master2节点上的HiveServer和WebHCat进程状态显示为故障，重启之后仍为故障状态。

原因分析

在Manager界面单独启动故障的hiveserver进程，登录后台查找hiveserver.out日志中对应时间点的报错，报错信息为：error parsing conf mapred-site.xml 和 Premature end of file。然后重启webhcat也发现同样报错，原因即为解析mapred-site.xml文件错误。

处理步骤

1. 以root用户登录Master2节点。
2. 执行**find / -name 'mapred-site.xml'**命令获取mapred-site.xml文件所在位置。
   - hiveserver对应路径为/opt/Bigdata/*集群版本*/1_13_HiveServer/etc/mapred-site.xml，
   - webhcat对应路径为/opt/Bigdata/*集群版本*/1_13_WebHCat/etc/mapred-site.xml。
3. 确认mapred-site.xml文件是否有异常，该案例中该配置文件内容为空导致解析失败。
4. 修复mapred-site.xml文件，将Master1节点上对应目录下的配置文件用scp命令拷贝到Master2节点对应目录替换原文件.
5. 执行**chown omm:wheel mapred-site.xml**命令更改所属组和用户。
6. 在Manager界面重启故障的HiveServer和WebHCat进程，恢复正常。

15.10.41 Hive 执行 insert into 语句报错，命令界面报错信息不明

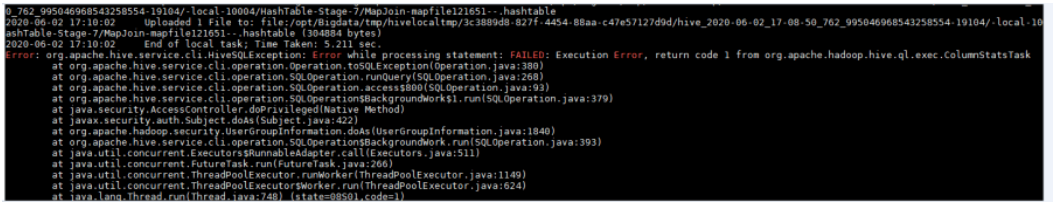
用户问题

客户使用MRS Hive执行一条SQL报错。

问题现象

客户使用MRS Hive执行一条SQL，有如下报错：

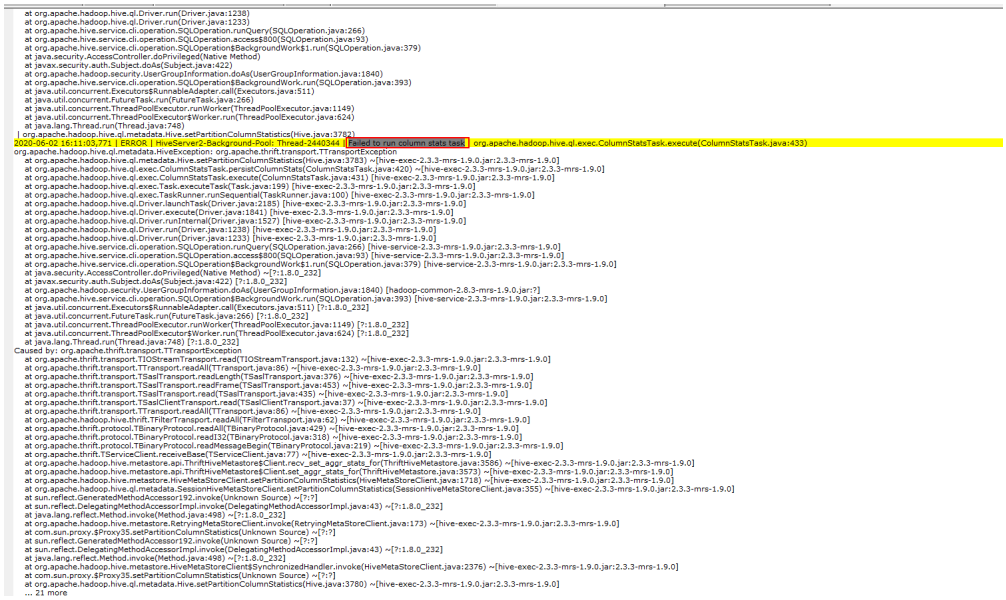
图 15-37 使用 MRS Hive 执行 SQL 报错



原因分析

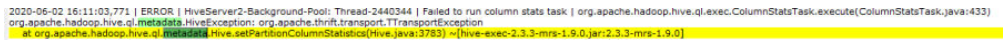
1. 查看Hiveserver日志，在对应时间点，有如下的报错信息。

图 15-38 Hiveserver 日志



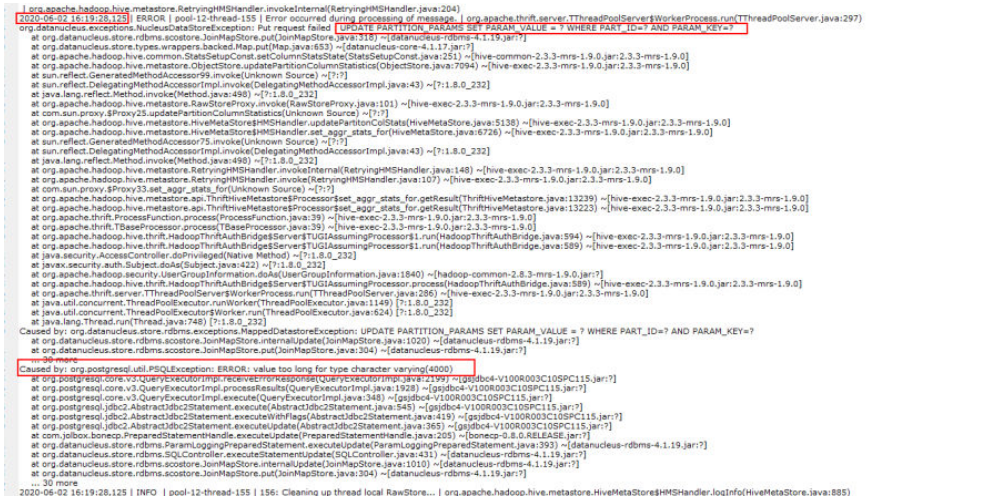
2. 在如上报错信息中未发现重要信息，但从堆栈中发现metadata字样，怀疑报错是和metastore有关。

图 15-39 堆栈中 metadata 字样



3. 查看metastore日志，发现如下报错。

图 15-40 metastore 日志



查看如上错误的上下文，确定是本次执行SQL的报错，在报错信息里面发现如下内容：

```
Caused by: org.postgresql.util.PSQLException: ERROR: value too long for type character varying(4000)
```

确认是客户该条SQL对表的操作，所有列的字节长度超过4000的限制，导致SQL执行失败，需要修改该限制。

处理步骤

步骤1 以root用户登录集群任意一个Master节点，并执行**su - omm**命令切换到omm用户。

步骤2 执行如下命令登录高斯DB。

gsql -p 20051 -d hivemeta -U *username* -W *password*

步骤3 执行如下命令修改限制。

alter table PARTITION_PARAMS alter column PARAM_VALUE type varchar(6000);

----结束

15.10.42 增加 Hive 表字段超时

用户问题

增加Hive表字段报错。

问题现象

Hive对包含10000+分区的表执行**ALTER TABLE table_name ADD COLUMNS(column_name string) CASCADE;**，报错如下：

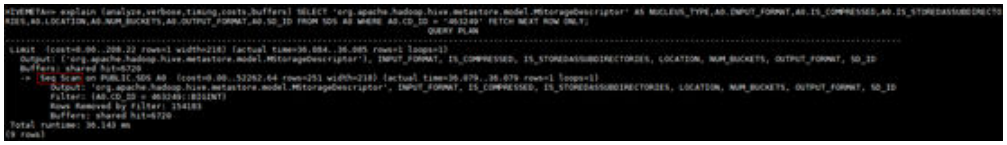
```
Timeout when executing method: alter_table_with_environment_context; 600525ms exceeds 600000ms
```

原因分析

1. MetaStore客户端连接超时，MRS默认MetaStore客户端和服务端连接的超时时间是600s，在Manager页面调大hive.metastore.client.socket.timeout为3600s。
2. 出现另一个报错：
   ```
   Error: org.apache.hive.service.cli.HiveSQLException: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. Unable to alter table. java.net.SocketTimeoutException: Read timed out
   ```
   Metastore元数据JDBC连接超时，默认60ms。
3. 调大javax.jdo.option.ConnectionURL中socketTimeout=60000，仍然产生最初的报错：
   ```
   Timeout when executing method: alter_table_with_environment_context;3600556ms exceeds 3600000ms
   ```
4. 尝试调大hive.metastore.batch.retrieve.max、hive.metastore.batch.retrieve.table.partition.max、dbservice.database.max.connections等参数均未能解决。
5. 怀疑是GaussDB的问题，因为增加字段会遍历每个分区执行getPartitionColumnStatistics和alterPartition。
6. 使用omm用户执行**gsql -p 20051 -U omm -W dbserverAdmin@123 -d hivemeta**登录hive元数据库。
7. 执行**select * from pg_locks;**没有发现锁等待。
8. 执行**select * from pg_stat_activity;**发现进程执行时间较长。
   ```
   SELECT 'org.apache.hadoop.hive.metastore.model.MPartitionColumnStatistics'AS NUCLEUS_TYPE,A0.AVG_COL_LEN,A0."COLUMN_NAME",A0.COLUMN_TYPE,A0.DB_NAME,A0.BIG_DECIMAL_HIGH_VALUE,A0.BIG_DECIMAL_LOW_VALUE,A0.DOUBLE_HIGH_VALUE,A0.DOUBLE_LOW_VALUE,A0.LAST_ANALYZED,A0.LONG_HIGH_VALUE,A0.LONG_LOW_VALUE,A0.MAX_COL_LEN,A0.NUM_DISTINCTS,A0.NUM_FALSES,A0.NUM_NULLS,A0.NUM_TRUES,A0.PARTITION_NAME,A0."TABLE_NAME",A0.CS_
   ```

```
ID,A0.PARTITION_NAMEAS NUCORDER0 FROM PART_COL_STATS A0 WHERE A0."TABLE_NAME" =
'$1' ANDA0.DB_NAME = '$2' AND A0.PARTITION_NAME = '$3' AND((((((A0."COLUMN_NAME" = '$4')
OR (A0."COLUMN_NAME" ='$5')) OR (A0."COLUMN_NAME" = '$6')) OR (A0."COLUMN_NAME"
='$7')) OR (A0."COLUMN_NAME" = '$8')) OR (A0."COLUMN_NAME" ='$9')) ORDER BY NUCORDER0;
```

9. 执行**gs_guc reload -c log_min_duration_statement=100 -D /srv/BigData/dbdata_service/data/**开启sql录制，发现**8.执行select * from pg_sta...**中语句执行时长700ms，而且因为有10000+分区，会触发执行10000+次命令。
10. 在sql前加explain (analyze,verbose,timing,costs,buffers)分析执行计划，发现执行时需要全表扫描。



11. 查看索引，发现不满足最左匹配原则。

```
HIVEMETA=# \d+ PART_COL_STATS
                                    Table "PUBLIC.PART_COL_STATS"
       Column        |          Type          |          Modifiers          | Storage  | Stats target | Description
---------------------+------------------------+-----------------------------+----------+--------------+-------------
 CS_ID               | BIGINT                 | not null                    | plain    |              |
 CAT_NAME            | CHARACTER VARYING(256) | default NULL::CHARACTER VARYING | extended |              |
 DB_NAME             | CHARACTER VARYING(128) | default NULL::CHARACTER VARYING | extended |              |
 TABLE_NAME          | CHARACTER VARYING(256) | default NULL::CHARACTER VARYING | extended |              |
 PARTITION_NAME      | CHARACTER VARYING(767) | default NULL::CHARACTER VARYING | extended |              |
 COLUMN_NAME         | CHARACTER VARYING(767) | default NULL::CHARACTER VARYING | extended |              |
 COLUMN_TYPE         | CHARACTER VARYING(128) | default NULL::CHARACTER VARYING | extended |              |
 PART_ID             | BIGINT                 | not null                    | plain    |              |
 LONG_LOW_VALUE      | BIGINT                 |                             | plain    |              |
 LONG_HIGH_VALUE     | BIGINT                 |                             | plain    |              |
 DOUBLE_LOW_VALUE    | DOUBLE PRECISION       |                             | plain    |              |
 DOUBLE_HIGH_VALUE   | DOUBLE PRECISION       |                             | plain    |              |
 BIG_DECIMAL_LOW_VALUE  | CHARACTER VARYING(4000) | default NULL::CHARACTER VARYING | extended |              |
 BIG_DECIMAL_HIGH_VALUE | CHARACTER VARYING(4000) | default NULL::CHARACTER VARYING | extended |              |
 NUM_NULLS           | BIGINT                 | not null                    | plain    |              |
 NUM_DISTINCTS       | BIGINT                 |                             | plain    |              |
 BIT_VECTOR          | BYTEA                  |                             | extended |              |
 AVG_COL_LEN         | DOUBLE PRECISION       |                             | plain    |              |
 MAX_COL_LEN         | BIGINT                 |                             | plain    |              |
 NUM_TRUES           | BIGINT                 |                             | plain    |              |
 NUM_FALSES          | BIGINT                 |                             | plain    |              |
 LAST_ANALYZED       | BIGINT                 | not null                    | plain    |              |
Indexes:
    "PART_COL_STATS_pkey" PRIMARY KEY, BTREE (CS_ID)
    "PART_COL_STATS_N49" BTREE (PART_ID)
    "PCS_STATS_IDX" BTREE (CAT_NAME, DB_NAME, TABLE_NAME, COLUMN_NAME, PARTITION_NAME)
Foreign-key constraints:
    "PART_COL_STATS_fkey" FOREIGN KEY (PART_ID) REFERENCES PARTITIONS(PART_ID) DEFERRABLE
Has OIDs: no
```

处理步骤

1. 重建索引。

```
su – omm
gsql -p 20051 -U omm -W dbserverAdmin@123 -d hivemeta
DROP INDEX PCS_STATS_IDX;
CREATE INDEX PCS_STATS_IDX ON PART_COL_STATS(DB_NAME, TABLE_NAME, COLUMN_NAME,
PARTITION_NAME);
CREATE INDEX SDS_N50 ON SDS(CD_ID);
```

2. 重新查看执行计划，发现语句已经可以索引查询，且5ms执行完成（原来是700ms）。重新执行hive表字段增加，已经可以添加成功。

15.10.43 Hive 服务重启失败

用户问题

修改Hive服务配置后，保存配置失败，Manager页面Hive服务的配置状态为配置失败。

问题现象

用户A在MRS节点后台上打开了Hive相关配置文件且未关闭，此时用户B在MRS Manager页面的“服务管理”中修改Hive配置项，保存配置并重启Hive服务，此时保存配置失败，并且Hive服务启动失败。

原因分析

由于用户B在MRS Manager页面修改配置时，配置文件被用户A在MRS节点后台打开，导致该配置文件不能被替换，最终导致Hive服务启动失败。

处理步骤

步骤1 用户需要首先手动关闭集群节点后台打开的Hive配置文件。

步骤2 在MRS Manager页面重新修改Hive的配置并保存配置。

步骤3 重启Hive服务。

----结束

15.10.44 hive 执行删除表失败

用户问题

hive表删除失败

问题现象

hive创建的二级分区表有两万多个分区，导致用户在执行**truncate table ${TableName},drop table ${TableName}**时失败。

原因分析

删除文件操作是单线程串行执行的，hive分区数过多导致在元数据数据库会保存大量元数据信息，在执行删表语句时删除元数据就要用很长时间，最终在超时时间内删除不完，就会导致操作失败。

说明

超时时间可通过登录FusionInsight Manager，选择“集群 > 服务 > Hive > 配置 >全部配置 > MetaStore（角色） > 服务初始化”查看，“hive.metastore.client.socket.timeout”对应的值即为超时时间时长，在“描述”列可查看默认值。

处理步骤

步骤1 如果是内部表可以先通过**alter table ${TableName} set TBLPROPERTIES('EXTERNAL'='true')**来将内部表转成外部表，这样hive删除的时候只删除元数据省去了删除hdfs数据的时间。

步骤2 如果要用相同的表名可以先将表结构用**show create table ${TableName}**来导出表结构，再用**ALTER TABLE ${TableName} RENAME TO ${new_table_name};**来将表重命名。这样就可以新建一个和原来一样表。

步骤3 执行**hdfs dfs –rm –r –f ${hdfs_path}**在hdfs上删除表数据。

步骤4 在hive中用**alter table ${Table_Name} drop partition (${PartitionName}<’XXXX’, ${PartitionName}>’XXXX’);**删除分区(具体删除条件可灵活处理),减少文件数。

步骤5 删除分区少于一千个后，直接用**drop table ${TableName}**删掉表即可。

----结束

建议与总结

hive分区虽然可以提高查询效率,但要避免分区不合理导致出现大量小文件的问题,要提前规划好分区策略.

15.10.45 Hive 执行 msck repair table table_name 报错

现象描述

Hive执行msck repair table table_name报错：FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask (state=08S01,code=1)。

可能原因

查看HiveServer日志/var/log/Bigdata/hive/hiveserver/hive.log，发现目录名不符合分区格式。

```
2020-07-15 15:38:10,427 | WARN  | HiveServer2-Background-Pool: Thread-10905216 | Failed to run metacheck: | org.apache.hadoop.hive.ql.exec.DDLTask.msck(DDLTask.java:2023)
org.apache.hadoop.hive.ql.metadata.HiveException: Repair: Cannot add partition ads_marketing.t_marketing_telemarketing_order_list:dtime=2020-04-24 17%3A55%3A00 due to invalid characters in the name
    at org.apache.hadoop.hive.ql.exec.DDLTask.msck(DDLTask.java:1986) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.exec.DDLTask.execute(DDLTask.java:424) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:199) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:100) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:2105) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:1841) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1527) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1238) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1233) [hive-exec-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:266) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hive.service.cli.operation.SQLOperation.access$800(SQLOperation.java:93) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork$1.run(SQLOperation.java:379) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at java.security.AccessController.doPrivileged(Native Method) ~[?:1.8.0_232]
    at javax.security.auth.Subject.doAs(Subject.java:422) [?:1.8.0_232]
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1840) [hadoop-common-2.8.3-mrs-1.9.0.jar:?]
    at org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork.run(SQLOperation.java:393) [hive-service-2.3.3-mrs-1.9.0.jar:2.3.3-mrs-1.9.0]
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511) [?:1.8.0_232]
    at java.util.concurrent.FutureTask.run(FutureTask.java:266) [?:1.8.0_232]
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149) [?:1.8.0_232]
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624) [?:1.8.0_232]
    at java.lang.Thread.run(Thread.java:748) [?:1.8.0_232]
```

处理步骤

- 方法一：删除错误的文件或目录。
- 方法二：执行set hive.msck.path.validation=skip，跳过无效的目录。

15.10.46 在 Hive 中 drop 表后，如何完全释放磁盘空间

用户问题

在Hive命令行执行**drop**表的操作后，通过命令**hdfs dfsadmin -report**查看磁盘空间，发现表没有删除。

原因分析

在Hive命令行执行**drop**表只删除了外部表的表结构，并没有删除该表存储在HDFS上的表数据。

处理步骤

步骤1 使用**root**用户登录安装客户端的节点，并认证用户。

cd *客户端安装目录*

source bigdata_env

kinit *组件业务用户*（未开启Kerberos认证的集群跳过此操作）

步骤2 执行以下命令删除存储在HDFS上的表。

hadoop fs -rm hdfs://hacluster/*表所在的具体路径*

----结束

15.10.47 客户端执行 SQL 报错连接超时

现象描述

客户端执行SQL失败，报错：Timed out waiting for a free available connection。

可能原因

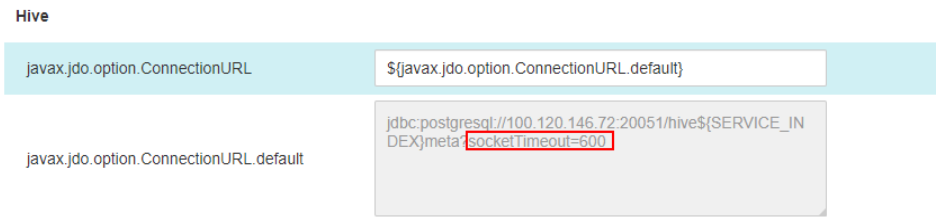
DBService连接较多，获取连接超时。

操作步骤

步骤1 客户端是否使用Spark-SQL客户端执行SQL。

- 是，检查连接的URL中超时参数，将其修改为600，执行**步骤7**。
- 否，执行**步骤2**。

步骤2 登录Manager页面，选择“集群 > 服务 > Hive > 配置 > 全部配置”，搜索“javax.jdo.option.ConnectionURL”，检查超时参数是否小于600。



说明

Hive、HiveServer、MetaStore、WebHCat中均有该参数，请确保它们的参数值一致。

- 是，执行**步骤3**。
- 否，执行**步骤7**。

步骤3 检查参数“javax.jdo.option.ConnectionURL”的值是否为“${javax.jdo.option.ConnectionURL.default}”。

- 是，执行**步骤4**。
- 否，修改URL中超时参数为600，单击“保存”，执行**步骤7**

步骤4 单击“实例”，选择任意HiveServer实例，并使用**root**用户登录实例节点。

步骤5 打开配置文件“${BIGDATA_HOME}/FusionInsight_Current/*HiveServer/etc/hivemetastore-site.xml”，查找配置项“javax.jdo.option.ConnectionURL”，复制配置项值。

```
<property>
<name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:postgresql://100.120.146.72:20051/hivemeta?socketTimeout=60</value>
</property>
<property>
```

步骤6 登录Manager页面，选择“集群 > 服务 > Hive > 配置 > 全部配置”，搜索“javax.jdo.option.ConnectionURL”，修改配置为**步骤5**中复制的URL，并修改超时参数为600，单击“保存”。

说明

Hive、HiveServer、MetaStore、WebHCat中均有该参数，请确保它们的参数值一致。

步骤7 选择“集群 > 服务 > Hive > 配置 > 全部配置”，搜索“maxConnectionsPerPartition”，检查是否小于100。

- 是，修改参数为100，单击“保存”，执行**步骤8**。
- 否，执行**步骤8**。

步骤8 若以上步骤有修改参数，选择“集群 > 服务 > Hive > 概览”，选择“更多 > 滚动重启服务”，未修改则无需执行此步骤。

----结束

15.10.48 WebHCat 健康状态异常导致启动失败

用户问题

WebHCat实例启动失败。

问题现象

在Manager页面上查看到WebHCat实例的健康状态为“故障”，并上报“ALM-12007 进程故障”告警，该告警的服务名称为“Hive”，实例名称为“WebHCat”。且重启Hive服务报错。

查看WebCat实例的日志“/var/log/Bigdata/hive/webhcat/webhcat.log”报错“Service not found in Kerberos database”和“Address already in use”。

处理步骤

步骤1 依次登录WebHCat实例所在节点检查“/etc/hosts”文件中的IP及主机名称映射关系是否正确。且“/etc/hostname”和“/etc/HOSTNAME”文件的WebHCat配置需与“/etc/hosts”保持一致，若不一致则需手动修改。

说明

WebHCat实例的IP地址及主机名称映射关系可登录FusionInsight Manager界面，选择“集群 > 服务 > Hive > 实例”查看。

步骤2 登录WebHCat实例所在节的任一节点，执行以下命令切换到**omm**用户。

su - omm

步骤3 执行以下命令查看是否存在WebHCat进程。

ps -ef|grep webhcat|grep -v grep

若存在，则需执行以下命令结束WebHCat进程：

kill -9 ${webhcat_pid}

步骤4 登录FusionInsight Manager，选择“集群 > 服务 > Hive > 实例”，勾选所有WebHCat实例，选择“更多 > 重启实例”，等待WebHCat重启成功即可。

----结束

15.10.49 mapred-default.xml 文件解析异常导致 WebHCat 启动失败

用户问题

MRS的Hive服务故障，重新启动后，Master2节点上的HiveServer和WebHCat进程启动失败，Master1节点进程正常。

原因分析

登录Master2节点，查看“/var/log/Bigdata/hive/hiveserver/hive.log”日志，发现HiveServer一直加载“/opt/Bigdata/*/*_HiveServer/etc/hive-site.xml”；查看HiveServer退出时的“/var/log/Bigdata/hive/hiveserver/hiveserver.out”日志，发现解析“mapred-default.xml”文件异常。

处理步骤

步骤1 登录Master2节点，使用以下命令查找“mapred-default.xml”所在路径：

find /opt/ -name 'mapred-default.xml'

查询到该配置文件在“/opt/Bigdata/*/*_WebHCat/etc/”目录下面，且该文件内容为空。

步骤2 登录到Master1节点，将“/opt/Bigdata/*/*_WebHCat/etc/mapred-default.xml”文件拷贝到Master2节点，并修改文件的属组为“omm:wheel”。

步骤3 登录Manager，重启异常的HiveServer和WebHCat实例。

----结束

15.11 使用 Hue

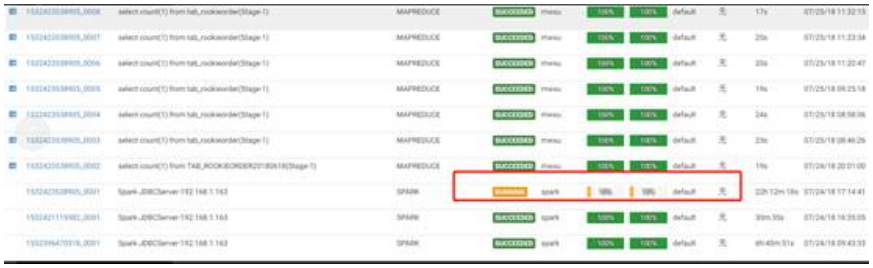
15.11.1 Hue 上有 job 在运行

用户问题

客户查到Hue上有job在运行。

问题现象

客户的MRS装好后， Hue上查到有Job在运行 ，并且目前在运行的job并不是客户操作的。



原因分析

此Job为Spark启动之后，系统自身连接jdbc的一个任务，是常驻的。

处理步骤

非问题，无需处理。

15.11.2 使用 IE 浏览器在 Hue 中执行 HQL 失败

问题背景与现象

使用IE浏览器在Hue中访问Hive Editor并执行所有HQL失败，界面提示“There was an error with your query.”。

原因分析

IE浏览器存在功能问题，不支持在307重定向中处理含有form data的AJAX POST请求，建议更换兼容的浏览器。

解决办法

使用Google Chrome浏览器21及以上版本。

15.11.3 Hue（主）无法打开 web 网页

问题背景与现象

访问Hue（主）的WebUI界面提示如下：

```
Service Unavailable
The server is temporarily unable to service your request due to maintenance downtime or capacity problems. Please try again later.
```

原因分析

- Hue配置过期。
- 单Master节点集群中，Hue服务需要手动修改配置。

解决办法

- Hue配置过期，重启Hue服务即可。
- 单Master节点的集群Hue服务需要手动修改配置。
  a. 登录Master节点。
  b. 执行**hostname -i**获取本机IP。
  c. 执行如下命令获取“HUE_FLOAT_IP”的地址：
     grep "HUE_FLOAT_IP" ${BIGDATA_HOME}/*MRS*_Current/1_*/etc*/ENV_VARS;
     其中*MRS*以实际文件名为准。
  d. 比较本机IP和“HUE_FLOAT_IP”的值是否相同，若不相同，请修改“HUE_FLOAT_IP”的值为本机IP。
  e. 重启Hue服务。

15.11.4 Hue WebUI 访问失败

用户问题

访问Hue WebUI跳转到错误的页面。

问题现象

查看hue web ui报错如下：

```
503 Service Unavailable
The server is temporarily unable to service your requster due to maintenance downtime or capacity problems.Please try again later.
```

原因分析

- hue配置过期。
- 单Master节点集群中，Hue服务需要手动修改配置。

处理步骤

步骤1 登录Master节点。

步骤2 执行**hostname -i**获取本机IP。

步骤3 执行如下命令获取“HUE_FLOAT_IP”的地址：

grep "HUE_FLOAT_IP" ${BIGDATA_HOME}/*MRS*_Current/1_*/etc*/ENV_VARS,其中*MRS*以实际文件名为准。

步骤4 比较本机IP和“HUE_FLOAT_IP”的值是否相同，若不相同，请修改“HUE_FLOAT_IP”的值为本机IP。

步骤5 重启Hue服务。

----结束

15.11.5 Hue 界面无法加载 HBase 表

用户问题

用户在Hue界面将hive数据导入hbase后，报检测不到hbase表的错误。

问题现象

Kerberos集群中，IAM子账户权限不足导致无法加载hbase表。

原因分析

IAM子账户权限不足。

处理步骤

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统管理 > 用户管理”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加supergroup组。

步骤5 单击“确定”完成修改操作。

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行单击“修改”。

步骤4 为用户添加supergroup组。

步骤5 单击“确定”完成修改操作。

----结束

建议与总结

如果是开启Kerberos认证的集群，页面出现 No data available优先排查权限问题。

15.12 使用 Impala

15.12.1 用户连接 impala-shell 失败

用户问题

用户连接impala-shell失败。

问题现象

用户在“组件管理”页面修改任意组件的配置并重启服务后，连接impala-shell，会出现连接失败，报错no such file/directory。

```
[root@node-master1emdj etc]# pwd
/opt/Bigdata/MRS_2.1.0/1_7_KuduMaster/etc
[root@node-master1emdj etc]# impala-shell -i 192.168.0.73
shell-init: error retrieving current directory: getcwd: cannot access parent directories: No such file or directory
chdir: error retrieving current directory: getcwd: cannot access parent directories: No such file or directory
Traceback (most recent call last):
  File "/opt/client/Impala/impala/shell/impala_shell.py", line 38, in <module>
    from impala_client import (ImpalaClient, DisconnectedException, QueryStateException,
  File "/opt/client/Impala/impala/shell/lib/impala_client.py", line 20, in <module>
    import sasl
  File "build/bdist.linux-x86_64/egg/sasl/__init__.py", line 1, in <module>

  File "build/bdist.linux-x86_64/egg/sasl/saslwrapper.py", line 7, in <module>
  File "build/bdist.linux-x86_64/egg/_saslwrapper.py", line 7, in <module>
  File "build/bdist.linux-x86_64/egg/_saslwrapper.py", line 3, in __bootstrap__
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 2594, in <module>
    for comparator, version in req.specs:
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 425, in __init__

  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 440, in add_entry
    `req`.  But, if there is an active distribution for the project and it
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 1688, in find_on_path
    return ()
  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 1835, in _normalize_cached

  File "/usr/lib/python2.7/site-packages/setuptools-0.6c11-py2.7.egg/pkg_resources.py", line 1829, in normalize_path
    register_namespace_handler(object,null_ns_handler)
  File "/usr/lib64/python2.7/posixpath.py", line 368, in realpath
    return abspath(path)
  File "/usr/lib64/python2.7/posixpath.py", line 356, in abspath
    cwd = os.getcwd()
OSError: [Errno 2] No such file or directory
```

原因分析

修改服务配置并重启服务后，部分服务的目录结构会删除并重新创建，如服务的etc目录等。如果重启服务前所在的目录为etc或者其子目录，由于重启后目录重建，仍在原来目录执行impala-shell时会产生某些系统变量或者参数无法找到的情况，所以连接impala-shell连接失败。

处理步骤

任意切换到存在的目录，重新连接impala-shell即可。

15.12.2 创建 Kudu 表报错

用户问题

创建Kudu表报错。

问题现象

新建了集群，在创建表时，报错“[Cloudera]ImpalaJDBCDriver ERROR processing query/statement. Error Code: 0, SQL state: TStatus(statusCode:ERROR_STATUS, sqlState:HY000, errorMessage:AnalysisException: Table property 'kudu.master_addresses' is required when the impalad startup flag -kudu_master_hosts is not used.”

原因分析

客户未在impala sql中指定kudu.master_addresses地址导致报错：Table property 'kudu.master_addresses' is required when the impalad startup flag -kudu_master_hosts is not used.

处理步骤

在创建Kudu表时指定“kudu.master_addresses”地址。

15.12.3 Impala 客户端登录失败

用户问题

运行Impala client会报类似如下错误信息：

```
[root@node-master1avIy ~]# impala-shell -i 192.168.128.49:21000
  File "/opt/client/Impala/impala/shell/impala_shell.py", line 1675
    except Exception, e:
                    ^
SyntaxError: invalid syntax
[root@node-master1avIy ~]#
```

原因分析

由于最新的MRS集群使用的是Euler2.9及以上版本的操作系统，系统自带只python3版本，而Impala client是基于python2实现的，和python3部分语法不兼容，运行Impala client会报错误信息，所以需要手动安装python2以解决Impala client运行问题。

处理步骤

步骤1 使用root用户登录Impala所在节点，执行如下命令，确认当前系统上安装的python版本：

python --version

```
[root@node-master2JgOY ~]# python --version
Python 3.7.4
```

步骤2 执行命令**yum install make**，查看yum是否可用。

- 如果yum install报如下错误，说明yum设置有问题，执行**步骤3**。

```
[root@node-master2JgOY ~]# yum install make
Error: There are no enabled repositories in "/etc/yum.repos.d", "/etc/yum/repos.d", "/etc/distro.repos.d".
```

- 如果没有报错，执行**步骤4**。

步骤3 执行命令**cat /etc/yum.repos.d/EulerOS-base.repo**，查看yum源和系统版本信息不匹配是否匹配，如果不匹配则修改，如下所示：

修改前：

```
[root@node-master1avIy ~]# cat /etc/yum.repos.d/EulerOS-base.repo
[base]
name=EulerOS-2.0SP2 base
baseurl=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.2/os/x86_64/
enabled=1
gpgcheck=1
gpgkey=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.2/os/RPM-GPG-KEY-EulerOS
[root@node-master1avIy ~]# uname -a
Linux node-master1avIy.mrs-mq7v.com 4.18.0-147.5.1.6.h541.eulerosv2r9.x86_64 #1 SMP Wed Aug 4 02:30:13 UTC
 x86_64 GNU/Linux
```

修改后：

```
[base]
name=EulerOS-2.0SP9 base
baseurl=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.9/os/x86_64/
enabled=1
gpgcheck=1
gpgkey=http://mirrors.myhuaweicloud.com/euler/ict/site-euleros/euleros/repo/yum/2.9/os/RPM-GPG-KEY-EulerOS
```

步骤4 执行如下命令，查看yum源上python2开头的软件。

yum list python2*

```
[root@node-master2JgOY ~]# yum list python2*
Last metadata expiration check: 0:02:36 ago on Thu 16 Dec 2021 10:05:52 AM CST.
Available Packages
python2.x86_64                                  2.7.16-16.eulerosv2r9
python2-debug.x86_64                            2.7.16-16.eulerosv2r9
python2-devel.x86_64                            2.7.16-16.eulerosv2r9
python2-help.noarch                             2.7.16-16.eulerosv2r9
python2-pip.noarch                              18.0-13.h2.eulerosv2r9
python2-setuptools.noarch                       40.4.3-4.h1.eulerosv2r9
python2-tkinter.x86_64                          2.7.16-16.eulerosv2r9
python2-tools.x86_64                            2.7.16-16.eulerosv2r9
```

步骤5 执行如下命令，安装python2。

yum install python2

```
[root@node-master2JgOY ~]# yum install python2
Last metadata expiration check: 0:00:48 ago on Thu 16 Dec 2021 10:05:52 AM CST.
Error:
 Problem: problem with installed package python3-unversioned-command-3.7.4-7.h29.eulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h29.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h11.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h13.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h15.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h18.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h33.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - package python3-unversioned-command-3.7.4-7.h38.eulerosv2r9.x86_64 conflicts with python2 provided by python2-2.7.16-16.e
ulerosv2r9.x86_64
  - conflicting requests
(try to add '--allowerasing' to command line to replace conflicting packages or '--skip-broken' to skip uninstallable package
s or '--nobest' to use not only best candidate packages)
```

因为当前系统上已安装python3，所有直接安装python2会有上面的冲突提示。

可以选择--allowerasing或--skip-broken安装，例如：

yum install python2 --skip-broken

```
[root@node-master2JgOY ~]# yum install python2 --skip-broken
Last metadata expiration check: 0:34:08 ago on Thu 16 Dec 2021 10:05:52 AM CST.
Dependencies resolved.
================================================================================================================================
 Package                          Architecture         Version                                 Repository          Size
================================================================================================================================
Installing:
 python2                          x86_64               2.7.16-16.eulerosv2r9                   base               6.4 M
Installing dependencies:
 libXft                           x86_64               2.3.2-13.eulerosv2r9                    base                41 k
```

安装完成后，会自动将python版本修改为python2，如下所示：

```
Installed:
  libXft-2.3.2-13.eulerosv2r9.x86_64                          libXrender-0.9.10-10.eulerosv2r9.x86_64
  python2-2.7.16-16.eulerosv2r9.x86_64                        python2-debug-2.7.16-16.eulerosv2r9.x86_64
  python2-devel-2.7.16-16.eulerosv2r9.x86_64                  python2-help-2.7.16-16.eulerosv2r9.noarch
  python2-setuptools-40.4.3-4.h1.eulerosv2r9.noarch           python2-tkinter-2.7.16-16.eulerosv2r9.x86_64
  python2-tools-2.7.16-16.eulerosv2r9.x86_64                  python3-rpm-generators-9-1.eulerosv2r9.noarch
  tk-1:8.6.8-5.eulerosv2r9.x86_64

Complete!
[root@node-master2JgOY ~]# python --version
Python 2.7.16
```

如果python2安装成功，但是显示的python版本不对，需要执行以下命令手动给“/usr/bin/python2”创建软链接“/usr/bin/python”。

ln -s /usr/bin/python2 /usr/bin/python

步骤6 验证Impala client是否可用。

```
[root@node-master1avIy ~]# impala-shell -i 192.168.128.49:21000
Starting Impala Shell without Kerberos authentication
Opened TCP connection to 192.168.128.49:21000
Connected to 192.168.128.49:21000
Server version: impalad version 3.4.0-RELEASE RELEASE (build eebadd34c1563cbf5825a4e4d361e7b3601f9827)
***********************************************************************************
Welcome to the Impala shell.
(Impala Shell v3.4.0-RELEASE (eebadd3) built on Thu Nov  4 11:29:54 CST 2021)

After running a query, type SUMMARY to see a summary of where time was spent.
***********************************************************************************
[192.168.128.49:21000] default> show databases;
Query: show databases
+------------------+----------------------------------------------+
| name             | comment                                      |
+------------------+----------------------------------------------+
| _impala_builtins | System database for Impala builtin functions |
| default          | Default Hive database                        |
+------------------+----------------------------------------------+
Fetched 2 row(s) in 0.16s
[192.168.128.49:21000] default>
```

----结束

15.13 使用 Kafka

15.13.1 运行 Kafka 获取 topic 报错

用户问题

客户运行Kafka获取topic报错。

问题现象

运行Kafka获取topic时报错，报错内容如下：

```
ERROR org.apache.kafka.common.errors.InvalidReplicationFactorException: Replication factor: 2 larger than available brokers: 0.
```

原因分析

由特殊字符导致获取zookeeper地址的变量错误。

处理步骤

步骤1 登录任意一个Master节点。

步骤2 执行**cat /opt/client/Kafka/kafka/config/server.properties |grep '^zookeeper.connect ='**命令，查看zookeeper地址的变量。

步骤3 重新运行Kafka获取topic，其中从**步骤2**中获取的变量不要添加任何字符。

----结束

15.13.2 Flume 可以正常连接 Kafka，但是发送消息失败。

问题现象

使用MRS版本安装集群，主要安装ZooKeeper、Flume、Kafka。

在使用Flume向Kafka发送数据功能时，发现Flume发送数据到Kafka失败。

可能原因

1. Kafka服务异常。
2. Flume连接Kafka地址错误，导致发送失败。
3. Flume发送超过Kafka大小限制的消息，导致发送失败。

原因分析

Flume发送数据到Kafka失败，可能原因是Flume侧问题或者Kafka侧问题。

1. Manager界面查看当前Kafka状态及监控指标。
   - MRS Manager界面操作：登录MRS Manager，选择 “服务管理 > Kafka ”，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Flume日志，发现打印MessageSizeTooLargeException异常信息，如下所示：

   ```
   2016-02-26 14:55:19,126 | WARN  | [SinkRunner-PollingRunner-DefaultSinkProcessor] |  Produce request with correlation id 349829 failed due to [LOG,7]: kafka.common.MessageSizeTooLargeException  | kafka.utils.Logging$class.warn(Logging.scala:83)
   ```

   通过异常信息，发现当前Flume向Kafka写入的数据超过了Kafka服务端定义的消息的最大值。

3. 通过Manager查看Kafka服务端定义的消息的最大值。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

   进入Kafka配置页面，参数类别选择全部配置，显示所有Kafka相关配置，在“搜索”中输入message.max.bytes进行检索。

   MRS中Kafka服务端默认可以接收的消息最大为1000012 bytes =977KB。

解决办法

与用户确认，当前Flume发送数据确实存在超过1M的消息。因此，为了确保当前这些消息能够写入Kafka，需要调整Kafka服务端相关参数。

步骤1 修改message.max.bytes，使得message.max.bytes的值大于当前业务中消息最大值，使得Kafka服务端可以接收全部消息。

步骤2 修改replica.fetch.max.bytes，使得**replica.fetch.max.bytes >= message.max.bytes**，使得不同Broker上的Partition的Replica可以同步到全部消息。

- MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
- FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

进入Kafka配置页面，参数类别选择全部配置， 显示所有Kafka相关配置，在“搜索”中输入replica.fetch.max.bytes进行检索。

步骤3 单击“保存”，并重启Kafka服务，使得Kafka相关配置生效。

步骤4 修改消费者业务应用中fetch.message.max.bytes，使得fetch.message.max.bytes >= message.max.bytes，确保消费者可以消费到全部消息。

----结束

15.13.3 Producer 发送数据失败，抛出 NullPointerException

问题现象

使用MRS安装集群，主要安装ZooKeeper、Kafka。

在使用Producer向Kafka发送数据功能时，发现客户端抛出NullPointerException。

可能原因

1. Kafka服务异常。
2. 客户端Producer侧配置jaas和keytab文件错误。

原因分析

Producer发送数据到Kafka失败，可能原因客户端Producer侧问题或者Kafka侧问题。

1. 通过Manager页面查看kafka服务状态及监控指标。
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。

- FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

  查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。

2. 查看Producer客户端日志，发现打印NullPointerException异常信息，如图15-41所示。

图 15-41 Producer 客户端日志

```
[2016-12-06 02:04:05,906]-[schedule-C50D0717-4643-4D4E-9D6E-B940E4BD7159]-[kafka-producer-network-thread |
SZX1000161910-10.21.219.222-bigdata-producer-5]-[1005]-[org.apache.kafka.clients.producer.internals.Sender.run
thread:
java.lang.NullPointerException
    at org.apache.kafka.common.network.Selector.pollSelectionKeys(Selector.java:302)
    at org.apache.kafka.common.network.Selector.poll(Selector.java:283)
    at org.apache.kafka.clients.NetworkClient.poll(NetworkClient.java:260)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:229)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:134)
    at java.lang.Thread.run(Thread.java:745)
[2016-12-06 02:04:05,921]-[schedule-C50D0717-4643-4D4E-9D6E-B940E4BD7159]-[kafka-producer-network-thread |
SZX1000161910-10.21.219.222-bigdata-producer-3]-[1005]-[org.apache.kafka.clients.producer.internals.Sender.run
thread:
java.lang.NullPointerException
    at org.apache.kafka.common.network.Selector.pollSelectionKeys(Selector.java:302)
    at org.apache.kafka.common.network.Selector.poll(Selector.java:283)
    at org.apache.kafka.clients.NetworkClient.poll(NetworkClient.java:260)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:229)
    at org.apache.kafka.clients.producer.internals.Sender.run(Sender.java:134)
    at java.lang.Thread.run(Thread.java:745)
```

或者日志中只有异常信息没有堆栈信息（只有NullPointerException无堆栈信息，出现这个问题是jdk的自我保护，相同堆栈打印太多，就会触发这个保护开关，后续不再打印堆栈），如图15-42所示。

图 15-42 异常信息

```
[2016-11-23 04:06:53,973] [kafka-producer-network-thread | producer-1] [ERROR] [org.apache.kafka.clients.producer.internals.Sender] (run:130)- Uncaught error in kafka producer I/O thread:
java.lang.NullPointerException
[2016-11-23 04:06:53,973] [kafka-producer-network-thread | producer-1] [ERROR] [org.apache.kafka.clients.producer.internals.Sender] (run:130)- Uncaught error in kafka producer I/O thread:
java.lang.NullPointerException
[2016-11-23 04:06:53,973] [kafka-producer-network-thread | producer-1] [ERROR] [org.apache.kafka.clients.producer.internals.Sender] (run:130)- Uncaught error in kafka producer I/O thread:
java.lang.NullPointerException
[2016-11-23 04:06:53,973] [kafka-producer-network-thread | producer-1] [ERROR] [org.apache.kafka.clients.producer.internals.Sender] (run:130)- Uncaught error in kafka producer I/O thread:
java.lang.NullPointerException
```

3. 查看Producer客户端日志，发现打印Failed to configure SaslClientAuthenticator异常信息，如图15-43所示。

图 15-43 异常日志信息

```
Caused by: org.apache.kafka.common.KafkaException: Failed to configure SaslClientAuthenticator
at org.apache.kafka.common.security.authenticator.SaslClientAuthenticator.configure(SaslClientAuthenticator.java:96)
at org.apache.kafka.common.network.SaslChannelBuilder.buildChannel(SaslChannelBuilder.java:89)
... 9 more
Caused by: org.apache.kafka.common.KafkaException: Failed to create SaslClient
at org.apache.kafka.common.security.authenticator.SaslClientAuthenticator.createSaslClient(SaslClientAuthenticator.java:112)
at org.apache.kafka.common.security.authenticator.SaslClientAuthenticator.configure(SaslClientAuthenticator.java:94)
... 10 more
Caused by: javax.security.sasl.SaslException: PLAIN: authorization ID and password must be specified
at com.sun.security.sasl.PlainClient.<init>(PlainClient.java:58)
at com.sun.security.sasl.ClientFactoryImpl.createSaslClient(ClientFactoryImpl.java:97)
at javax.security.sasl.Sasl.createSaslClient(Sasl.java:384)
at com.ibm.messagehub.login.MessageHubSaslClientFactory.createSaslClient(MessageHubSaslClientFactory.java:77)
at javax.security.sasl.Sasl.createSaslClient(Sasl.java:384)
at org.apache.kafka.common.security.authenticator.SaslClientAuthenticator$1.run(SaslClientAuthenticator.java:107)
at org.apache.kafka.common.security.authenticator.SaslClientAuthenticator$1.run(SaslClientAuthenticator.java:102)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.kafka.common.security.authenticator.SaslClientAuthenticator.createSaslClient(SaslClientAuthenticator.java:102)
... 11 more
```

4. 认证失败导致创建KafkaChannel失败，导致通过channel(key)方法获取的KafkaChannel为空，以至于疯狂打印NullPointerException，上述日志可以发现，认证失败的原因是用户密码不正确，密码不正确的原因可能是用户名不匹配导致。

5. 检查Jaas文件和Keytab文件，发现Jaas文件中配置使用的pricipal为stream。

图 15-44 检查 Jaas 文件

```
KafkaClient {
com.sun.security.auth.module.Krb5LoginModule required
debug=false
keyTab="/opt/client/user.keytab"
useTicketCache=false
storeKey=true
principal="stream@HADOOP.COM"
useKeyTab=true;
};
```

查看user.keytab文件，发现principal为zmk_kafka。

图 15-45 查看 user.keytab 文件

```
[root@8-5-148-6 client]# klist -kt user.keytab
Keytab name: FILE:user.keytab
KVNO Timestamp         Principal
---- ----------------- --------------------------------------------------------
   1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
   1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
```

发现jaas文件和user.keytab文件中principal不对应。

该情况是由于应用程序自动定时更新Jaas文件，但是有两个不同的进程在进行更新，一个进程写入正确的Principal而另一个却写入了错误的Principal，以至于程序时而正常，时而异常。

解决办法

步骤1 修改Jaas文件，确保使用的Principal在Keytab文件中存在。

----结束

15.13.4 Producer 发送数据失败，抛出 TOPIC_AUTHORIZATION_FAILED

问题现象

使用MRS安装集群，主要安装ZooKeeper、Kafka。

在使用Producer向Kafka发送数据功能时，发现客户端抛出TOPIC_AUTHORIZATION_FAILED。

可能原因

1. Kafka服务异常。
2. 客户端Producer侧采用非安全访问，服务端配置禁止访问。
3. 客户端Producer侧采用非安全访问，Kafka Topic设置ACL。

原因分析

Producer发送数据到Kafka失败，可能原因客户端Producer侧问题或者Kafka侧问题。

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Producer客户端日志，发现打印TOPIC_AUTHORIZATION_FAILED异常信息。

   ```
   [root@10-10-144-2 client]# kafka-console-producer.sh --broker-list 10.5.144.2:9092 --topic test
   1
   [2017-01-24 16:58:36,671] WARN Error while fetching metadata with correlation id 0 :
   {test=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
   [2017-01-24 16:58:36,672] ERROR Error when sending message to topic test with key: null, value: 1
   bytes with error: Not authorized to access topics: [test]
   (org.apache.kafka.clients.producer.internals.ErrorLoggingCallback)
   ```

   Producer采用9092端口来访问Kafka，9092为非安全端口。
3. 通过Manager页面，查看当前Kafka集群配置，发现未设置自定义配置“allow.everyone.if.no.acl.found”=“false”。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。
4. 当acl设置为false不允许采用9092来进行访问。
5. 查看Producer客户端日志，发现打印TOPIC_AUTHORIZATION_FAILED异常信息。

   ```
   [root@10-10-144-2 client]# kafka-console-producer.sh --broker-list 10.5.144.2:21005 --topic test_acl
   1
   [2017-01-25 11:09:40,012] WARN Error while fetching metadata with correlation id 0 :
   {test_acl=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
   [2017-01-25 11:09:40,013] ERROR Error when sending message to topic test_acl with key: null, value:
   1 bytes with error: Not authorized to access topics: [test_acl]
   (org.apache.kafka.clients.producer.internals.ErrorLoggingCallback)
   [2017-01-25 11:14:40,010] WARN Error while fetching metadata with correlation id 1 :
   {test_acl=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
   ```

   Producer采用21005端口来访问Kafka，21005为非安全端口。
6. 通过客户端命令查看topic的acl权限设置信息。

   ```
   [root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:24002/
   kafka --list --topic topic_acl
   Current ACLs for resource `Topic:topic_acl`:
    User:test_user has Allow permission for operations: Describe from hosts: *
   User:test_user has Allow permission for operations: Write from hosts: *
   ```

   Topic设置acl，则不允许采用9092来访问。
7. 查看Producer客户端日志，发现打印TOPIC_AUTHORIZATION_FAILED异常信息。

   ```
   [root@10-10-144-2 client]# kafka-console-producer.sh --broker-list 10.5.144.2:21007 --topic topic_acl
   --producer.config /opt/client/Kafka/kafka/config/producer.properties
   1
   [2017-01-25 12:43:58,506] WARN Error while fetching metadata with correlation id 0 :
   {topic_acl=TOPIC_AUTHORIZATION_FAILED} (org.apache.kafka.clients.NetworkClient)
   [2017-01-25 12:43:58,507] ERROR Error when sending message to topic topic_acl with key: null,
   value: 1 bytes with error: Not authorized to access topics: [topic_acl]
   (org.apache.kafka.clients.producer.internals.ErrorLoggingCallback)
   ```

Producer采用21007端口来访问Kafka。

8. 通过客户端命令klist查询当前认证用户。

```
[root@10-10-144-2 client]# klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: test@HADOOP.COM

Valid starting     Expires            Service principal
01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
```

当前认证用户为**test**。

9. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
User:test_user has Allow permission for operations: Write from hosts: *
```

Topic设置acl，用户test_user具有producer权限。test无权限进行producer操作。

解决方法参考步骤2。

10. 通过SSH登录Kafka Broker：

通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

查看kafka-authorizer.log发现如下日志提示用户不属于kafka或者kafkaadmin组。

```
2017-01-25 13:26:33,648 | INFO  | [kafka-request-handler-0] | The principal is test, belongs to Group : [hadoop, ficommon] | kafka.authorizer.logger (SimpleAclAuthorizer.scala:169)
2017-01-25 13:26:33,648 | WARN  | [kafka-request-handler-0] | The user is not belongs to kafka or kafkaadmin group, authorize failed! | kafka.authorizer.logger (SimpleAclAuthorizer.scala:170)
```

解决方法参考步骤3。

解决办法

步骤1 配置自定义配置“allow.everyone.if.no.acl.found”参数为“true”，重启Kafka服务。

步骤2 采用具有权限用户登录。

例如：

kinit test_user

或者赋予用户相关权限。

须知

需要使用Kafka管理员用户（属于kafkaadmin组）操作。

例如：

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --topic topic_acl --producer --add --allow-principal User:test

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=8.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
 User:test_user has Allow permission for operations: Write from hosts: *
  User:test has Allow permission for operations: Describe from hosts: *
  User:test has Allow permission for operations: Write from hosts: *
```

步骤3 用户加入**Kafka**组或者**Kafkaadmin**组。

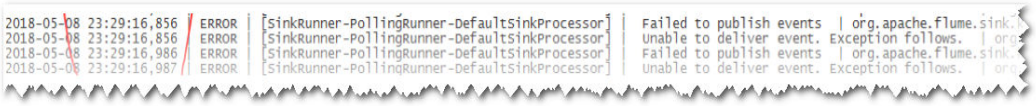
----结束

15.13.5 Producer 偶现发送数据失败，日志提示 Too many open files in system

问题背景与现象

在使用Producer向Kafka发送数据功能时，发现客户端发送失败。

图 15-46 Producer 发送数据失败



可能原因

1. Kafka服务异常。
2. 网络异常。
3. Kafka Topic异常。

原因分析

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

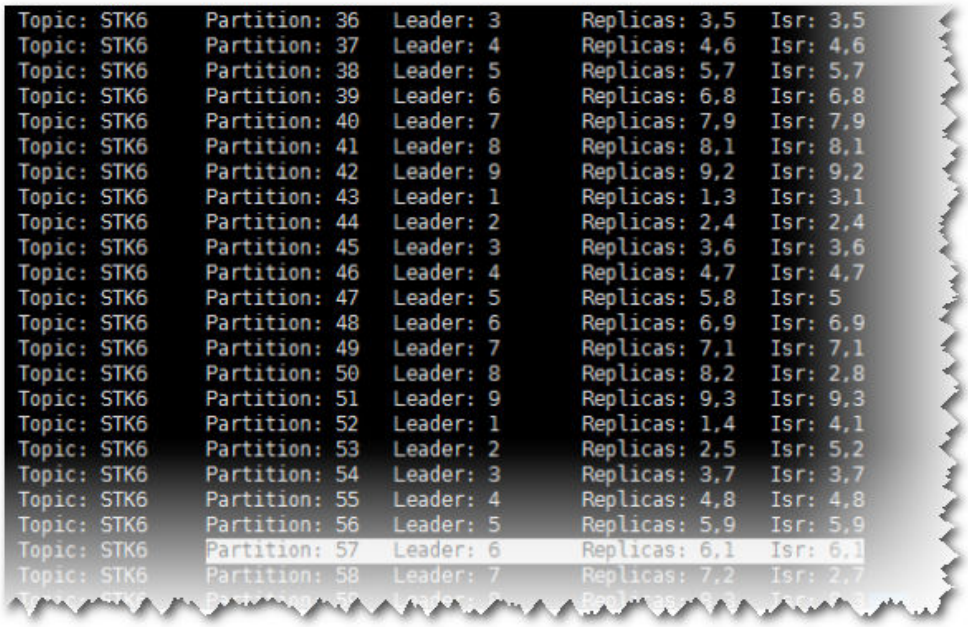
     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看SparkStreaming日志中提示错误的Topic信息。

   执行Kafka相关命令，获取Topic分布信息和副本同步信息，观察返回结果。

   kafka-topics.sh --describe --zookeeper <zk_host:port/chroot>

   如图15-47所示，发现对应Topic状态正常。所有Partition均存在正常Leader信息。

图 15-47 Topic 状态



3. 通过telnet命令，查看是否可以连接Kafka。

   telnet kafka业务ip kafka业务port

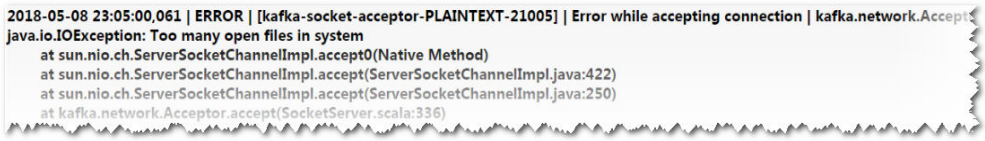
   如果无法telnet成功，请检查网络安全组与ACL。

4. 通过SSH登录Kafka Broker。

   通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

   查看server.log发现如下日志抛出java.io.IOException: Too many open files in system。

图 15-48 日志异常



5. 通过lsof命令查看当前节点Kafka进程句柄使用情况，发现占用的句柄数达到了47万。

图 15-49 句柄数

6. 排查业务代码，不停地new新的producer对象，未正常关闭。

解决办法

步骤1 停止当前应用，保证服务端句柄不再疯狂增加影响服务正常运行。

步骤2 优化应用代码，解决句柄泄露问题。

建议：全局尽量使用一个Producer对象，在使用完成之后主动调用close接口进行句柄关闭。

----结束

15.13.6 Consumer 初始化成功，但是无法从 Kafka 中获取指定 Topic 消息

问题背景与现象

使用MRS安装集群，主要安装ZooKeeper、Flume、Kafka、Storm、Spark。

使用Storm、Spark、Flume或者自己编写consumer代码来消费Kafka中指定Topic的消息时，发现消费不到任何数据。

可能原因

1. Kafka服务异常。
2. Consumer中ZooKeeper相关连接地址配置错误，导致无法消费。
3. Consumer发生ConsumerRebalanceFailedException异常，导致无法消费。
4. Consumer发生网络导致的ClosedChannelException异常，导致无法消费。

原因分析

Storm、Spark、Flume或者自定义Consumer代码可以都称为Consumer。

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 通过Kafka Client，判断是否可以正常消费数据。

   假设客户端已经安装在/opt/client目录，test为需要消费的Topic名称，192.168.234.231为ZooKeeper的IP地址。

   ```
   cd /opt/client
   source bigdata_env
   kinit admin
   kafka-topics.sh --zookeeper 192.168.234.231:2181/kafka --describe --topic testkafka-console-consumer.sh --topic test --zookeeper 192.168.234.231:2181/kafka --from-beginning
   ```

   当可以消费到数据时，表示集群服务正常。
3. 查看Consumer相关配置，发现ZooKeeper连接地址错误。
   - Flume

     ```
     server.sources.Source02.type=org.apache.flume.source.kafka.KafkaSource

     server.sources.Source02.zookeeperConnect=192.168.234.231:2181
     ```

```
server.sources.Source02.topic = test
server.sources.Source02.groupId = test_01
```

- Spark

```
val zkQuorum = "192.168.234.231:2181"
```

- Storm

```
BrokerHosts brokerHosts = new ZKHosts("192.168.234.231:2181");
```

- Consumer API

```
zookeeper.connect="192.168.234.231:2181"
```

MRS中Kafka在ZooKeeper存储的ZNode是以/kafka为根路径，有别于开源。Kafka对应的ZooKeeper连接配置为192.168.234.231:2181/kafka。

Consumer中配置为ZooKeeper连接配置为192.168.234.231:2181，导致无法正确获取Kafka中Topic相关信息。

解决方法参考**步骤1**。

4. 查看Consumer相关日志，发现打印ConsumerRebalanceFailedException异常。

```
2016-02-03 15:55:32,557 | ERROR | [ZkClient-EventThread-75- 192.168.234.231:2181/kafka] |  Error handling event ZkEvent[New session event sent to kafka.consumer.ZookeeperConsumerConnector$ZKSessionExpireListener@34b41dfe]  | org.I0Itec.zkclient.ZkEventThread.run(ZkEventThread.java:77)
kafka.common.ConsumerRebalanceFailedException: pc-zjqbetl86-1454482884879-2ec95ed3 can't rebalance after 4 retries
at kafka.consumer.ZookeeperConsumerConnector$ZKRebalancerListener.syncedRebalance(ZookeeperConsumerConnector.scala:633)
at kafka.consumer.ZookeeperConsumerConnector$ZKSessionExpireListener.handleNewSession(ZookeeperConsumerConnector.scala:487)
at org.I0Itec.zkclient.ZkClient$4.run(ZkClient.java:472)
at org.I0Itec.zkclient.ZkEventThread.run(ZkEventThread.java:71)
```

通过异常信息，发现当前Consumer没有在指定的重试次数内完成Rebalance，使得Consumer没有被分配Kafka Topic-Partition，则无法消费消息。

解决方法参考**步骤3**。

5. 查看Consumer相关日志，发现打印Fetching topic metadata with correlation id 0 for topics [Set(test)] from broker [id:26,host:192-168-234-231,port:9092] failed错误信息和ClosedChannelException异常。

```
[2016-03-04 03:33:53,047] INFO Fetching metadata from broker id:26,host:192-168-234-231,port:9092 with correlation id 0 for 1 topic(s) Set(test) (kafka.client.ClientUtils$)
[2016-03-04 03:33:55,614] INFO Connected to 192-168-234-231:21005 for producing (kafka.producer.SyncProducer)
[2016-03-04 03:33:55,614] INFO Disconnecting from 192-168-234-231:21005 (kafka.producer.SyncProducer)
[2016-03-04 03:33:55,615] WARN Fetching topic metadata with correlation id 0 for topics [Set(test)] from broker [id:26,host: 192-168-234-231,port:21005] failed (kafka.client.ClientUtils$)
java.nio.channels.ClosedChannelException
at kafka.network.BlockingChannel.send(BlockingChannel.scala:100)
at kafka.producer.SyncProducer.liftedTree1$1(SyncProducer.scala:73)
at kafka.producer.SyncProducer.kafka$producer$SyncProducer$$doSend(SyncProducer.scala:72)
at kafka.producer.SyncProducer.send(SyncProducer.scala:113)
at kafka.client.ClientUtils$.fetchTopicMetadata(ClientUtils.scala:58)
at kafka.client.ClientUtils$.fetchTopicMetadata(ClientUtils.scala:93)
at kafka.consumer.ConsumerFetcherManager$LeaderFinderThread.doWork(ConsumerFetcherManager.scala:66)
at kafka.utils.ShutdownableThread.run(ShutdownableThread.scala:60)
[2016-03-04 03:33:55,615] INFO Disconnecting from 192-168-234-231:21005 (kafka.producer.SyncProducer)
```

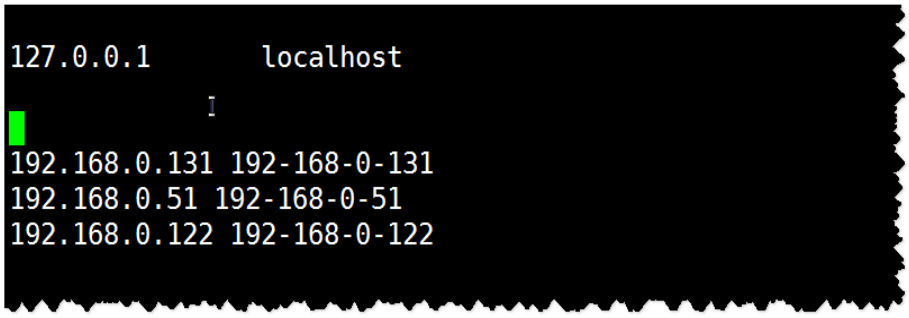
通过异常信息，发现当前Consumer无法从Kafka Broker 192-168-234-231节点获取元数据，导致无法连接正确的Broker获取消息。

6. 检查网络是否存在问题，如果网络没有问题，检查是否配置主机和IP的对应关系

- Linux

  执行**cat /etc/hosts**命令。

图 15-50 示例 1



- Windows
  打开“C:\Windows\System32\drivers\etc\hosts”。

图 15-51 示例 2

```
# For example:
#
#      192.168.94.97     rhino.acme.com          # source server
#      192.168.63.10     x.acme.com              # x client host

# localhost name resolution is handled within DNS itself.
#       127.0.0.1       localhost
#       ::1             localhost
10.82.129.120 rms.huawei.com # modified by IrmTool at 2015-01-18 17:55:13
```

解决方法参考**步骤4**。

解决办法

步骤1 ZooKeeper连接地址配置错误。

步骤2 修改Consumer配置中的ZooKeeper连接地址信息，保证和MRS相一致。

- Flume
  ```
  server.sources.Source02.type=org.apache.flume.source.kafka.KafkaSource
  server.sources.Source02.zookeeperConnect=192.168.234.231:2181/kafka
  server.sources.Source02.topic = test
  server.sources.Source02.groupId = test_01
  ```
- Spark
  ```
  val zkQuorum = "192.168.234.231:2181/kafka"
  ```
- Storm
  ```
  BrokerHosts brokerHosts = new ZKHosts("192.168.234.231:2181/kafka");
  ```
- Consumer API
  ```
  zookeeper.connect="192.168.234.231:2181/kafka"
  ```

步骤3 Rebalance异常。

同一个消费者组(consumer group)有多个consumer先后启动，就是一个消费者组内有多个consumer同时消费多个partition数据，consumer端也会有负载均衡（consumer个数小于partitions数量时）。

consumer实际上是靠存储在zk中的临时节点来表明针对哪个topic的partition拥有读权限的。所在路径为：/consumers/consumer-group-xxx/owners/topic-xxx/x。

当触发负载均衡后，原来的consumer会重新计算并释放已占用的partitions，此过程需要一定的处理时间，新来的consumer抢占该partitions时很有可能会失败。

表 15-3 参数说明

名称	作用	默认值
rebalance.max.retries	Rebalance最大重试次数	4
rebalance.backoff.ms	Rebalance每次重试间隔	2000
zookeeper.session.timeout.ms	Zookeeper连接会话超时时间	15000

可以适当调大上述三个参数，可以参考如下数值：

```
zookeeper.session.timeout.ms = 45000
rebalance.max.retries = 10
rebalance.backoff.ms = 5000
```

参数设置应遵循：

rebalance.max.retries * rebalance.backoff.ms > zookeeper.session.timeout.ms

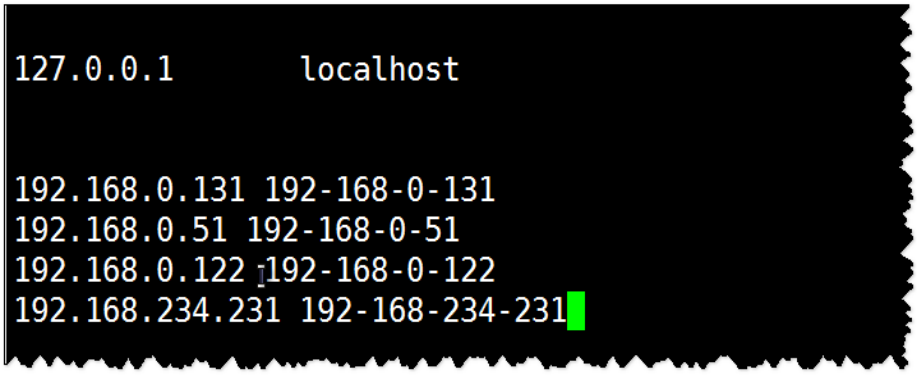
步骤4 网络异常。

在hosts文件中没有配置主机名和IP的对应关系，导致使用主机名进行访问时，无法获取信息。

步骤5 在hosts文件中添加对应的主机名和IP的对应关系。

- Linux

  图 15-52 示例 3

- Windows

图 15-53 示例 4

```
# For example:
#
#      192.168.94.97     rhino.acme.com          # source server
#      192.168.63.10     x.acme.com              # x client host

# localhost name resolution is handled within DNS itself.
#       127.0.0.1       localhost
#       ::1             localhost
10.82.129.120 rms.huawei.com # modified by IrmTool at 2015-01-18 17:55:13
192.168.234.231 192-168-234-231
```

----结束

15.13.7 Consumer 消费数据失败，Consumer 一直处于等待状态

问题现象

使用MRS服务安装集群，主要安装ZooKeeper、Kafka。

在使用Consumer从Kafka消费数据时，发现客户端一直处于等待状态。

可能原因

1. Kafka服务异常。
2. 客户端Consumer侧采用非安全访问，服务端配置禁止访问。
3. 客户端Consumer侧采用非安全访问，Kafka Topic设置ACL。

原因分析

Consumer向Kafka消费数据失败，可能原因客户端Consumer侧问题或者Kafka侧问题。

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 查看Consumer客户端日志发现频繁连接Broker节点和断链打印信息，如下所示。

   ```
   [root@10-10-144-2 client]# kafka-console-consumer.sh --topic test --zookeeper 10.5.144.2:2181/kafka --from-beginning


   [2017-03-07 09:22:00,658] INFO Fetching metadata from broker BrokerEndPoint(1,10.5.144.2,9092) with correlation id 26 for 1 topic(s) Set(test) (kafka.client.ClientUtils$)
   [2017-03-07 09:22:00,659] INFO Connected to 10.5.144.2:9092 for producing (kafka.producer.SyncProducer)
   [2017-03-07 09:22:00,659] INFO Disconnecting from 10.5.144.2:9092 (kafka.producer.SyncProducer)
   ```

   Consumer采用9092端口来访问Kafka，9092为非安全端口。
3. 通过Manager页面查看当前Kafka集群配置，发现未配置自定义参数“allow.everyone.if.no.acl.found”=“false”。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。

- FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Kafka > 配置"。

4. 当acl设置为false不允许采用9092来进行访问。
5. 查看Consumer客户端日志发现频繁连接Broker节点和断链打印信息，如下所示。

```
[root@10-10-144-2 client]# kafka-console-consumer.sh --topic test_acl --zookeeper 10.5.144.2:2181/kafka --from-beginning


[2017-03-07 09:49:16,992] INFO Fetching metadata from broker BrokerEndPoint(2,10.5.144.3,9092) with correlation id 16 for 1 topic(s) Set(topic_acl) (kafka.client.ClientUtils$)
[2017-03-07 09:49:16,993] INFO Connected to 10.5.144.3:9092 for producing (kafka.producer.SyncProducer)
[2017-03-07 09:49:16,994] INFO Disconnecting from 10.5.144.3:9092 (kafka.producer.SyncProducer)
```

   Consumer采用21005端口来访问Kafka，21005为非安全端口。

6. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
User:test_user has Allow permission for operations: Write from hosts: *
```

   Topic设置acl，则不允许采用9092来访问。

7. 查看Consumer客户端日志发现打印信息，如下所示。

```
[root@10-10-144-2 client]# kafka-console-consumer.sh --topic topic_acl --bootstrap-server 10.5.144.2:21007 --consumer.config /opt/client/Kafka/kafka/config/consumer.properties --from-beginning --new-consumer

[2017-03-07 10:19:18,478] INFO Kafka version : 0.9.0.0 (org.apache.kafka.common.utils.AppInfoParser)
[2017-03-07 10:19:18,478] INFO Kafka commitId : unknown (org.apache.kafka.common.utils.AppInfoParser)
```

   Consumer采用21007端口来访问Kafka。

8. 通过客户端命令klist查询当前认证用户：

```
[root@10-10-144-2 client]# klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: test@HADOOP.COM

Valid starting     Expires            Service principal
01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
```

   当前认证用户为**test**。

9. 通过客户端命令查看topic的acl权限设置信息。

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:24002/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
 User:test_user has Allow permission for operations: Write from hosts: *
 User:ttest_user has Allow permission for operations: Read from hosts: *
```

   Topic设置acl，test无权限进行consumer操作。

   解决方法参考**步骤2**。

10. 通过SSH登录Kafka Broker：

    通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

    查看kafka-authorizer.log发现如下日志提示用户不属于kafka或者kafkaadmin组。

```
2017-01-25 13:26:33,648 | INFO  | [kafka-request-handler-0] | The principal is test, belongs to Group : [hadoop, ficommon] | kafka.authorizer.logger (SimpleAclAuthorizer.scala:169)
2017-01-25 13:26:33,648 | WARN  | [kafka-request-handler-0] | The user is not belongs to kafka or kafkaadmin group, authorize failed! | kafka.authorizer.logger (SimpleAclAuthorizer.scala:170)
```

    解决方法参考**步骤3**。

解决办法

步骤1 配置自定义参数“allow.everyone.if.no.acl.found”参数为“true”，重启Kafka服务。

步骤2 采用具有权限用户登录。

例如：

kinit test_user

或者赋予用户相关权限。

须知

需要使用Kafka管理员用户（属于kafkaadmin组）操作。

例如：

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --topic topic_acl --consumer --add --allow-principal User:test --group test

```
[root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=8.5.144.2:2181/kafka --list --topic topic_acl
Current ACLs for resource `Topic:topic_acl`:
 User:test_user has Allow permission for operations: Describe from hosts: *
 User:test_user has Allow permission for operations: Write from hosts: *
  User:test has Allow permission for operations: Describe from hosts: *
  User:test has Allow permission for operations: Write from hosts: *
  User:test has Allow permission for operations: Read from hosts: *
```

步骤3 用户加入**Kafka**组或者**Kafkaadmin**组。

----结束

15.13.8 SparkStreaming 消费 Kafka 消息失败，提示 Error getting partition metadata

问题现象

使用SparkStreaming来消费Kafka中指定Topic的消息时，发现无法从Kafka中获取到数据。提示如下错误： Error getting partition metadata。

```
Exception in thread "main" org.apache.spark.SparkException:  Error getting partition metadata for 'testtopic'. Does the topic exist?
org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
scala.util.Either.fold(Either.scala:97)
org.apache.spark.streaming.kafka.KafkaCluster$.checkErrors(KafkaCluster.scala:365)
org.apache.spark.streaming.kafka.KafkaUtils$.createDirectStream(KafkaUtils.scala:422)
com.XXXXXX.bigdata.spark.examples.FemaleInfoCollectionPrint$.main(FemaleInfoCollectionPrint.scala:45)
com.XXXXXX.bigdata.spark.examples.FemaleInfoCollectionPrint.main(FemaleInfoCollectionPrint.scala)
sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
java.lang.reflect.Method.invoke(Method.java:498)
org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$
$runMain(SparkSubmit.scala:762)
org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:183)
org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:208)
```

```
org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:123)
org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
```

可能原因

1. Kafka服务异常。
2. 客户端Consumer侧采用非安全访问，服务端配置禁止访问。
3. 客户端Consumer侧采用非安全访问，Kafka Topic设置ACL。

原因分析

1. 查看kafka服务状态：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka，

     查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 通过Manager页面，查看当前Kafka集群配置，发现未配置“allow.everyone.if.no.acl.found”或配置为“false”。
   - MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Kafka> 配置”。
   - FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka > 配置"。
3. 当acl设置为false不允许采用Kafka非安全端口21005来进行访问。
4. 通过客户端命令查看topic的acl权限设置信息：

   ```
   [root@10-10-144-2 client]# kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --list --topic topic_acl
   Current ACLs for resource `Topic:topic_acl`:
    User:test_user has Allow permission for operations: Describe from hosts: *
   User:test_user has Allow permission for operations: Write from hosts: *
   ```

   Topic设置acl，则不允许采用Kafka非安全端口21005来访问。

解决办法

步骤1 修改或者添加自定义配置“allow.everyone.if.no.acl.found”参数为“true”，重启Kafka服务。

步骤2 删除Topic设置的ACL。

例如：

kinit test_user

须知

需要使用Kafka管理员用户（属于kafkaadmin组）操作。

例如：

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --remove --allow-principal User:test_user --producer --topic topic_acl

kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka --remove --allow-principal User:test_user --consumer --topic topic_acl --group test

----结束

15.13.9 新建集群 Consumer 消费数据失败，提示 GROUP_COORDINATOR_NOT_AVAILABLE

问题背景与现象

新建Kafka集群，部署Broker节点数为2，使用Kafka客户端可以正常生产，但是无法正常消费。Consumer消费数据失败，提示GROUP_COORDINATOR_NOT_AVAILABLE，关键日志如下：

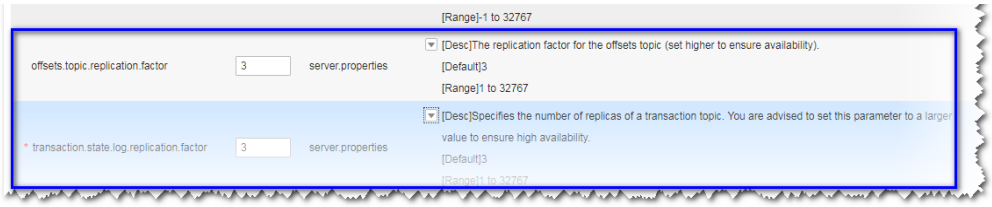
```
2018-05-12 10:58:42,561 | INFO | [kafka-request-handler-3] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 118 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,562 | INFO | [executor-Heartbeat] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
```

可能原因

__consumer_offsets无法创建。

原因分析

1. 查看日志，发现大量__consumer_offset创建异常。
2. 查看集群发现当前Broker数量为2。
3. 查看__cosumer_offset topic要求副本为3，因此创建失败。



解决办法

可以将扩容至至少3个流式core节点，或参考如下步骤修改服务配置参数。

步骤1 进入服务参数配置界面。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > Kafka > 服务配置”，“参数类别”设置为“全部配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager。选择“集群 > 服务 > Kafka”，单击“配置”，选择“全部配置”。

步骤2 搜索并修改offsets.topic.replication.factor和transaction.state.log.replication.factor的值为2。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

15.13.10 SparkStreaming 消费 Kafka 消息失败，提示 Couldn't find leader offsets

问题背景与现象

使用SparkStreaming来消费Kafka中指定Topic的消息时，发现无法从Kafka中获取到数据。提示如下错误： Couldn't find leader offsets。

```
2018-05-30 12:01:17,816 | INFO  | [Driver] | Reconnect due to socket error: java.net.SocketTimeoutException | kafka.utils.Logging$class.info(Logging.scala:68)
2018-05-30 12:01:47,859 | ERROR | [Driver] | User class threw exception: org.apache.spark.SparkException: java.net.SocketTimeoutException
org.apache.spark.SparkException: Couldn't find leader offsets for Set([STK6,57], [STK6,21]) | org.apache.spark.Logging$class.logError(Logging.scala:96)
org.apache.spark.SparkException: java.net.SocketTimeoutException
org.apache.spark.SparkException: Couldn't find leader offsets for Set([STK6,57], [STK6,21])
        at org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
        at org.apache.spark.streaming.kafka.KafkaCluster$$anonfun$checkErrors$1.apply(KafkaCluster.scala:366)
        at scala.util.Either.fold(Either.scala:97)
        at org.apache.spark.streaming.kafka.KafkaCluster$.checkErrors(KafkaCluster.scala:365)
        at org.apache.spark.streaming.kafka.KafkaUtils$.createDirectStream(KafkaUtils.scala:422)
        at org.apache.spark.streaming.kafka.KafkaUtils$.createDirectStream(KafkaUtils.scala:532)
        at org.apache.spark.streaming.kafka.KafkaUtils.createDirectStream(KafkaUtils.scala)
        at com.stk.bigdata.sparkstreaming.notify.SparkAlarmControlwz.main(SparkAlarmControlwz.java:194)
        at com.stk.bigdata.sparkstreaming.submit.SparkNotifyA.main(SparkNotifyA.java:14)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.spark.deploy.yarn.ApplicationMaster$$anon$2.run(ApplicationMaster.scala:540)
2018-05-30 12:01:47,863 | INFO  | [Driver] | Final app status: FAILED, exitCode: 15, (reason: User class threw exception: org.apache.spark.SparkException: java.
org.apache.spark.SparkException: Couldn't find leader offsets for Set([STK6,57], [STK6,21])) | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
2018-05-30 12:01:47,866 | INFO  | [pool-1-thread-1] | Invoking stop() from shutdown hook | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
```

可能原因

- Kafka服务异常。
- 网络异常。
- Kafka Topic异常。

原因分析

步骤1 通过Manager页面，查看Kafka集群当前状态，发现状态为“良好”，且监控指标内容显示正确。

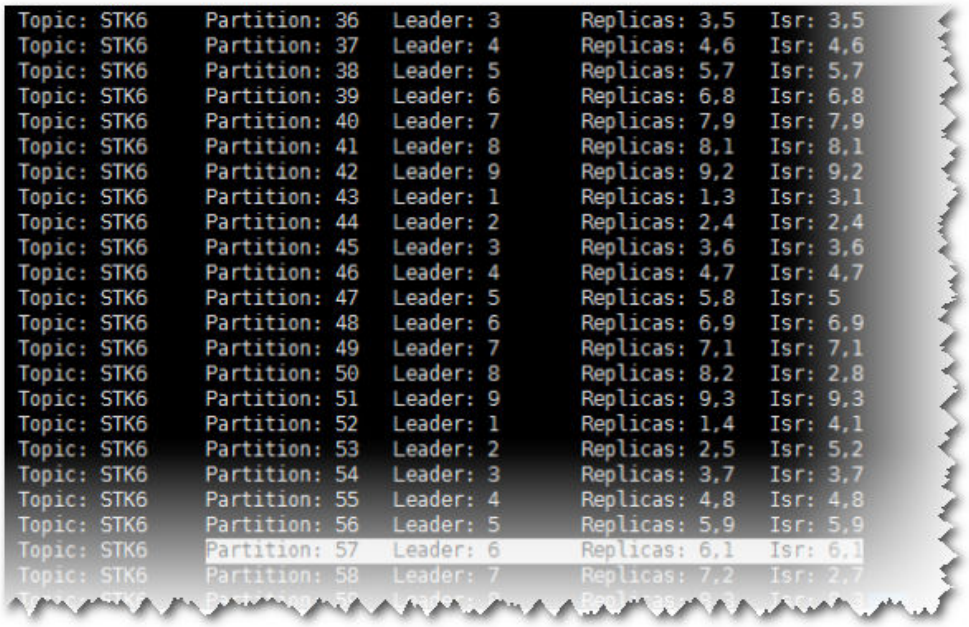
步骤2 查看SparkStreaming日志中提示错误的Topic信息。

执行Kafka相关命令，获取Topic分布信息和副本同步信息，观察返回结果。

kafka-topics.sh --describe --zookeeper <zk_host:port/chroot> --topic <topic name>

如下所示，发现对应Topic状态正常。所有Partition均存在正常Leader信息。

图 15-54 Topic 分布信息和副本同步信息



步骤3 检查客户端与Kafka集群网络是否连通，若网络不通协调网络组进行处理。

步骤4 通过SSH登录Kafka Broker。

通过**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

查看server.log发现如下日志抛出java.lang.OutOfMemoryError: Direct buffer memory。

```
2018-05-30 12:02:00,246 | ERROR | [kafka-network-thread-6-PLAINTEXT-3] | Processor got uncaught exception. | kafka.network.Processor (Logging.scala:103)

java.lang.OutOfMemoryError: Direct buffer memory
 at java.nio.Bits.reserveMemory(Bits.java:694)
 at java.nio.DirectByteBuffer.<init>(DirectByteBuffer.java:123)
 at java.nio.ByteBuffer.allocateDirect(ByteBuffer.java:311)
 at sun.nio.ch.Util.getTemporaryDirectBuffer(Util.java:241)
 at sun.nio.ch.IOUtil.read(IOUtil.java:195)
 at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:380)
```

at org.apache.kafka.common.network.PlaintextTransportLayer.read(PlaintextTransportLayer.java:110)

步骤5 通过Manager页面，查看当前Kafka集群配置。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > Kafka > 服务配置”，“参数类别”设置为“全部配置”，发现“KAFKA_JVM_PERFORMANCE_OPTS”的中“-XX:MaxDirectMemorySize”值为“1G”。
- FusionInsight Manager界面操作：登录FusionInsight Manager。选择“集群 > 服务 > Kafka”，单击“配置”，选择“全部配置”，发现“KAFKA_JVM_PERFORMANCE_OPTS”的中“-XX:MaxDirectMemorySize”值为“1G”。

步骤6 直接内存配置过小导致报错，而且一旦直接内存溢出，该节点将无法处理新请求，会导致其他节点或者客户端访问超时失败。

----结束

解决办法

步骤1 登录到Manager，进入 Kafka 配置页面。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > Kafka > 服务配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager。选择“集群 > 服务 > Kafka”，单击“配置”。

步骤2 选择“全部配置”，搜索并修改KAFKA_JVM_PERFORMANCE_OPTS的值。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

15.13.11 Consumer 消费数据失败，提示 SchemaException: Error reading field 'brokers'

问题背景与现象

Consumer来消费Kafka中指定Topic的消息时，发现无法从Kafka中获取到数据。提示如下错误： org.apache.kafka.common.protocol.types.SchemaException: Error reading field 'brokers': Error reading field 'host': Error reading string of length 28271, only 593 bytes available。

```
Exception in thread "Thread-0" org.apache.kafka.common.protocol.types.SchemaException: Error reading field 'brokers': Error reading field 'host': Error reading string of length 28271, only 593 bytes available
at org.apache.kafka.common.protocol.types.Schema.read(Schema.java:73)
at org.apache.kafka.clients.NetworkClient.parseResponse(NetworkClient.java:380)
at org.apache.kafka.clients.NetworkClient.handleCompletedReceives(NetworkClient.java:449)
at org.apache.kafka.clients.NetworkClient.poll(NetworkClient.java:269)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.clientPoll(ConsumerNetworkClient.java:360)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.poll(ConsumerNetworkClient.java:224)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.poll(ConsumerNetworkClient.java:192)
at
org.apache.kafka.clients.consumer.internals.ConsumerNetworkClient.poll(ConsumerNetworkClient.java:163)
atorg.apache.kafka.clients.consumer.internals.AbstractCoordinator.ensureCoordinatorReady(AbstractCoordinator.java:179)
at org.apache.kafka.clients.consumer.KafkaConsumer.pollOnce(KafkaConsumer.java:973)
at org.apache.kafka.clients.consumer.KafkaConsumer.poll(KafkaConsumer.java:937)
at KafkaNew.Consumer$ConsumerThread.run(Consumer.java:40)
```

可能原因

客户端和服务端Jar版本不一致。

解决办法

修改Consumer应用程序中kafka jar，确保和服务端保持一致。

15.13.12 Consumer 消费数据是否丢失排查

问题背景与现象

客户将消费完的数据存入数据库，发现数据与生产数据不一致，怀疑Kafka消费丢数据

可能原因

- 客户代码原因
- Kafka生产数据写入异常
- Kafka消费数据异常

解决办法

Kafka排查：

步骤1 通过consumer-groups.sh来观察写入和消费的offerset的变化情况（生产一定数量的消息，客户端进行消费，观察offerset的变化）。

步骤2 新建一个消费组，用客户端进行消费，然后查看消费的消息。

new-consumer:

kafka-console-consumer.sh --topic <topic name> --bootstrap-server <IP1:PORT, IP2:PORT,...> --new-consumer --consumer.config <config file>

----结束

客户代码排查：

步骤1 查看客户端里有没有提交offerset的报错。

步骤2 如果没有报错把消费的API里加上打印消息，打印少量数据（只打印key即可），查看丢失的数据。

----结束

15.13.13 账号锁定导致启动组件失败

问题背景与现象

新安装集群，启动Kafka失败。显示认证失败，导致启动失败。

```
/home/omm/kerberos/bin/kinit -k -t /opt/XXXXXX/Bigdata/etc/2_15_ Broker /kafka.keytab kafka/
hadoop.hadoop.com -c /opt/XXXXXX/Bigdata/etc/2_15_ Broker /11846 failed.
export key tab file for kafka/hadoop.hadoop.com failed.export and check keytab file failed, errMsg=]}] for
Broker #192.168.1.92@192-168-1-92.
[2015-07-11 02:34:33] RoleInstance started failure for ROLE[name: Broker].
[2015-07-11 02:34:34] Failed to complete the instances start operation. Current operation entities: [Broker
#192.168.1.92@192-168-1-92], Failure entites : [Broker #192.168.1.92@192-168-1-92].Operation
```

```
Failed.Failed to complete the instances start operation. Current operation entities:
[Broker#192.168.1.92@192-168-1-92], Failure entites: [Broker #192.168.1.92@192-168-1-92].
```

原因分析

查看Kerberos日志，/var/log/Bigdata/kerberos/krb5kdc.log，发现有集群外的IP使用kafka用户连接，导致多次认证失败，最终导致Kafka账号被锁定。

```
Jul 11 02:49:16 192-168-1-91 krb5kdc[1863](info): AS_REQ (2 etypes {18 17}) 192.168.1.93:
NEEDED_PREAUTH: kafka/hadoop.hadoop.com@HADOOP.COM for krbtgt/HADOOP.COM@HADOOP.COM,
Additional pre-authentication required
Jul 11 02:49:16 192-168-1-91 krb5kdc[1863](info): preauth (encrypted_timestamp) verify failure: Decrypt
integrity check failed
Jul 11 02:49:16 192-168-1-91 krb5kdc[1863](info): AS_REQ (2 etypes {18 17}) 192.168.1.93:
PREAUTH_FAILED: kafka/hadoop.hadoop.com@HADOOP.COM for krbtgt/HADOOP.COM@HADOOP.COM,
Decrypt integrity check failed
```

解决办法

进入集群外的节点（如原因分析示例中的192.168.1.93），断开其对Kafka的认证。等待5分钟，此账号就会被解锁。

15.13.14 Kafka Broker 上报进程异常，日志提示 IllegalArgumentException

问题背景与现象

使用Manager提示进程故障告警，查看告警进程为Kafka Broker。

可能原因

Broker配置异常。

原因分析

1. 在Manager页面，在告警页面得到主机信息。
2. 通过SSH登录Kafka Broker，执行**cd /var/log/Bigdata/kafka/broker**命令进入日志目录。

   查看server.log发现如下日志抛出IllegalArgumentException异常，提示java.lang.IllegalArgumentException: requirement failed: replica.fetch.max.bytes should be equal or greater than message.max.bytes。

   ```
   2017-01-25 09:09:14,930 | FATAL | [main] | | kafka.Kafka$ (Logging.scala:113)
   java.lang.IllegalArgumentException: requirement failed: replica.fetch.max.bytes should be equal or
   greater than message.max.bytes
       at scala.Predef$.require(Predef.scala:233)
       at kafka.server.KafkaConfig.validateValues(KafkaConfig.scala:959)
       at kafka.server.KafkaConfig.<init>(KafkaConfig.scala:944)
       at kafka.server.KafkaConfig$.fromProps(KafkaConfig.scala:701)
       at kafka.server.KafkaConfig$.fromProps(KafkaConfig.scala:698)
       at kafka.server.KafkaServerStartable$.fromProps(KafkaServerStartable.scala:28)
       at kafka.Kafka$.main(Kafka.scala:60)
       at kafka.Kafka.main(Kafka.scala)
   ```

   Kafka要求replica.fetch.max.bytes 需要大于等于message.max.bytes。
3. 进入Kafka配置页面，选择“全部配置”， 显示所有Kafka相关配置，分别搜索message.max.bytes、

   replica.fetch.max.bytes进行检索，发现replica.fetch.max.bytes小于message.max.bytes。

解决办法

步骤1 登录Manager界面，进入Kafka配置页面。

- MRS 3.x之前的版本：登录MRS Manager，选择“服务管理 > Kafka > 配置 > 全部配置”。
- MRS 3.x及后续版本，登录FusionInsight Manager，选择“集群 > 服务 > Kafka > 配置 > 全部配置”。

步骤2 搜索并修改replica.fetch.max.bytes参数，使得replica.fetch.max.bytes的值大于等于message.max.bytes，使得不同Broker上的Partition的Replica可以同步到全部消息。

步骤3 保存配置，查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。

步骤4 修改消费者业务应用中fetch.message.max.bytes，使得fetch.message.max.bytes的值大于等于message.max.bytes，确保消费者可以消费到全部消息。

----结束

15.13.15 执行 Kafka Topic 删除操作，发现无法删除

问题背景与现象

在使用Kafka客户端命令删除Topic时，发现Topic无法被删除。

```
kafka-topics.sh --delete --topic test --zookeeper 192.168.234.231:2181/kafka
```

可能原因

- 客户端命令连接ZooKeeper地址错误。
- Kafka服务异常Kafka部分节点处于停止状态。
- Kafka服务端配置禁止删除。
- Kafka配置自动创建，且Producer未停止。

原因分析

1. 客户端命令，打印ZkTimeoutException异常。

   ```
   [2016-03-09 10:41:45,773] WARN Can not get the principle name from server 192.168.234.231 (org.apache.zookeeper.ClientCnxn)
   Exception in thread "main" org.I0Itec.zkclient.exception.ZkTimeoutException: Unable to connect to zookeeper server within timeout: 30000
   at org.I0Itec.zkclient.ZkClient.connect(ZkClient.java:880)
   at org.I0Itec.zkclient.ZkClient.<init>(ZkClient.java:98)
   at org.I0Itec.zkclient.ZkClient.<init>(ZkClient.java:84)
   at kafka.admin.TopicCommand$.main(TopicCommand.scala:51)
   at kafka.admin.TopicCommand.main(TopicCommand.scala)
   ```

   解决方法参考**步骤1**。

2. 客户端查询命令。

   kafka-topics.sh --list --zookeeper 192.168.0.122:2181/kafka

   ```
   test - marked for deletion
   ```

   通过Manager查看Kafka Broker实例的运行状态。

   通过**cd /var/log/Bigdata/kafka/broker**命令进入RunningAsController节点日志目录，在controller.log发现ineligible for deletion: test。

```
2016-03-09 11:11:26,228 | INFO  | [main] | [Controller 1]: List of topics to be deleted:  |
kafka.controller.KafkaController (Logging.scala:68)
2016-03-09 11:11:26,229 | INFO  | [main] | [Controller 1]: List of topics ineligible for deletion: test |
kafka.controller.KafkaController (Logging.scala:68)
```

3. 通过Manager查询Broker删除Topic相关配置。

   解决方法参考**步骤2**

4. 客户端查询命令：

   ```
   kafka-topics.sh --describe –topic test --zookeeper 192.168.0.122:2181/kafka
   ```

   ```
   Topic:test      PartitionCount:10       ReplicationFactor:2     Configs:
           Topic: test     Partition: 0    Leader: -1      Replicas: 1,2   Isr:
           Topic: test     Partition: 1    Leader: -1      Replicas: 2,3   Isr:
           Topic: test     Partition: 2    Leader: -1      Replicas: 3,1   Isr:
           Topic: test     Partition: 3    Leader: -1      Replicas: 1,3   Isr:
           Topic: test     Partition: 4    Leader: -1      Replicas: 2,1   Isr:
           Topic: test     Partition: 5    Leader: -1      Replicas: 3,2   Isr:
           Topic: test     Partition: 6    Leader: -1      Replicas: 1,2   Isr:
           Topic: test     Partition: 7    Leader: -1      Replicas: 2,3   Isr:
           Topic: test     Partition: 8    Leader: -1      Replicas: 3,1   Isr:
           Topic: test     Partition: 9    Leader: -1      Replicas: 1,3   Isr:
   ```

   进入RunningAsController节点日志目录，在controller.log发现marked ineligible for deletion。

   ```
   2016-03-10 11:11:17,989 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Handling
   deletion for topics test | kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
   2016-03-10 11:11:17,990 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Not retrying
   deletion of topic test at this time since it is marked ineligible for deletion |
   kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
   ```

5. 通过Manager查询Broker状态。

   其中一个Broker处于停止或者故障状态。Topic进行删除必须保证该Topic的所有Partition所在的Broker必须处于正常状态。

   解决方法参考**步骤3**。

6. 进入RunningAsController节点日志目录，在controller.log发现Deletion successfully，然后又出现New topics: [Set(test)]，表明被再次创建。

   ```
   2016-03-10 11:33:35,208 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Deletion of topic
   test successfully completed | kafka.controller.TopicDeletionManager$DeleteTopicsThread
   (Logging.scala:68)

   2016-03-10 11:33:38,501 | INFO  | [ZkClient-
   EventThread-19-192.168.0.122:2181,160.172.0.52:2181,160.172.0.51:2181/kafka] |
   [TopicChangeListener on Controller 3]: New topics: [Set(test)], deleted topics: [Set()], new partition
   replica assignment
   ```

7. 通过Manager查询Broker创建Topic相关配置。

   经确认，对该Topic操作的应用没有停止。

   解决方法参考**步骤4**。

解决办法

步骤1 ZooKeeper连接失败导致。

Kafka客户端连接ZooKeeper服务超时。检查客户端到ZooKeeper的网络连通性。

网络连接失败，通过Manager界面查看Zookeeper服务信息。

配置错误，修改客户端命令中ZooKeeper地址。

步骤2 Kafka服务端配置禁止删除。

通过Manager界面修改delete.topic.enable为true。保存配置并重启服务。

客户端查询命令，无Topic:test。

```
kafka-topics.sh --list --zookeeper 192.168.0.122:24002/kafka
```

进入RunningAsController节点日志目录，在controller.log发现Deletion of topic test successfully。

```
2016-03-10 10:39:40,665 | INFO  | [delete-topics-thread-3] | [Partition state machine on Controller 3]: Invoking state change to OfflinePartition for partitions [test,2],[test,15],[test,6],[test,16],[test,12],[test,7],[test,10],[test,13],[test,9],[test,19],[test,3],[test,5],[test,1],[test,0],[test,17],[test,8],[test,4],[test,11],[test,14],[test,18] | kafka.controller.PartitionStateMachine (Logging.scala:68)
2016-03-10 10:39:40,668 | INFO  | [delete-topics-thread-3] | [Partition state machine on Controller 3]: Invoking state change to NonExistentPartition for partitions [test,2],[test,15],[test,6],[test,16],[test,12],[test,7],[test,10],[test,13],[test,9],[test,19],[test,3],[test,5],[test,1],[test,0],[test,17],[test,8],[test,4],[test,11],[test,14],[test,18] | kafka.controller.PartitionStateMachine (Logging.scala:68)
2016-03-10 10:39:40,977 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Deletion of topic test successfully completed | kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
```

步骤3 Kafka部分节点处于停止或者故障状态。

启动停止的Broker实例。

客户端查询命令，无Topic:test。

```
kafka-topics.sh --list --zookeeper 192.168.0.122:24002/kafka
```

进入RunningAsController节点日志目录，在controller.log发现Deletion of topic test successfully。

```
2016-03-10 11:17:56,463 | INFO  | [delete-topics-thread-3] | [Partition state machine on Controller 3]: Invoking state change to NonExistentPartition for partitions [test,4],[test,1],[test,8],[test,2],[test,5],[test,9],[test,7],[test,6],[test,0],[test,3] | kafka.controller.PartitionStateMachine (Logging.scala:68)
2016-03-10 11:17:56,726 | INFO  | [delete-topics-thread-3] | [delete-topics-thread-3], Deletion of topic test successfully completed | kafka.controller.TopicDeletionManager$DeleteTopicsThread (Logging.scala:68)
```

步骤4 Kafka配置自动创建，且Producer未停止。

停止相关应用，通过Manager界面修改“auto.create.topics.enable”为“false”，保存配置并重启服务。

步骤5 再次执行delete操作。

----结束

15.13.16 执行 Kafka Topic 删除操作，提示 AdminOperationException

问题背景与现象

在使用Kafka客户端命令设置Topic ACL权限时，发现Topic无法被设置。

```
kafka-topics.sh --delete --topic test4 --zookeeper 10.5.144.2:2181/kafka
```

提示错误ERROR kafka.admin.AdminOperationException: Error while deleting topic test4。

具体如下：

```
Error while executing topic command : Error while deleting topic test4
[2017-01-25 14:00:20,750] ERROR kafka.admin.AdminOperationException: Error while deleting topic test4
at kafka.admin.TopicCommand$$anonfun$deleteTopic$1.apply(TopicCommand.scala:177)
at kafka.admin.TopicCommand$$anonfun$deleteTopic$1.apply(TopicCommand.scala:162)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
```

```
at kafka.admin.TopicCommand$.deleteTopic(TopicCommand.scala:162)
at kafka.admin.TopicCommand$.main(TopicCommand.scala:68)
at kafka.admin.TopicCommand.main(TopicCommand.scala)
 (kafka.admin.TopicCommand$)
```

可能原因

用户不属于**kafkaadmin**组，Kafka提供安全访问接口，kafkaamdin组用户才可以进行topic删除操作。

原因分析

1. 使用客户端命令，打印AdminOperationException异常。
2. 通过客户端命令**klist**查询当前认证用户：

   ```
   [root@10-10-144-2 client]# klist
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: test@HADOOP.COM

   Valid starting     Expires            Service principal
   01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   如上例中当前认证用户为test。
3. 通过命令**id**查询用户组信息

   ```
   [root@10-10-144-2 client]# id test
   uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10003(kafka)
   ```

解决办法

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 用户管理”。

步骤3 在操作用户对应的“操作”列，单击“修改”。

步骤4 为用户加入**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

15.13.17 执行 Kafka Topic 创建操作，发现无法创建提示 NoAuthException

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现Topic无法被创建。

```
kafka-topics.sh --create --zookeeper 192.168.234.231:2181/kafka --replication-factor 1 --partitions 2 --topic test
```

提示错误NoAuthException，KeeperErrorCode = NoAuth for /config/topics。

具体如下：

```
Error while executing topic command org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /config/topics
org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /config/topics
 at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
 at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:685)
 at org.I0Itec.zkclient.ZkClient.create(ZkClient.java:304)
 at org.I0Itec.zkclient.ZkClient.createPersistent(ZkClient.java:213)
 at kafka.utils.ZkUtils$.createParentPath(ZkUtils.scala:215)
 at kafka.utils.ZkUtils$.updatePersistentPath(ZkUtils.scala:338)
 at kafka.admin.AdminUtils$.writeTopicConfig(AdminUtils.scala:247)
```

可能原因

用户不属于**kafkaadmin**组，Kafka提供安全访问接口，kafkaamdin组用户才可以进行topic删除操作。

原因分析

1. 使用客户端命令，打印NoAuthException异常。

   ```
   Error while executing topic command org.apache.zookeeper.KeeperException$NoAuthException:
   KeeperErrorCode = NoAuth for /config/topics
   org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException:
   KeeperErrorCode = NoAuth for /config/topics
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:685)
    at org.I0Itec.zkclient.ZkClient.create(ZkClient.java:304)
    at org.I0Itec.zkclient.ZkClient.createPersistent(ZkClient.java:213)
    at kafka.utils.ZkUtils$.createParentPath(ZkUtils.scala:215)
    at kafka.utils.ZkUtils$.updatePersistentPath(ZkUtils.scala:338)
    at kafka.admin.AdminUtils$.writeTopicConfig(AdminUtils.scala:247)
   ```

2. 通过客户端命令**klist**查询当前认证用户：

   ```
   [root@10-10-144-2 client]# klist
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: test@HADOOP.COM

   Valid starting     Expires            Service principal
   01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   如上例中当前认证用户为**test**。

3. 通过命令**id**查询用户组信息。

   ```
   [root@10-10-144-2 client]# id test
   uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10003(kafka)
   ```

解决办法

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 用户管理”。

步骤3 在操作用户对应的“操作”列，单击“修改”。

步骤4 为用户加入**kafkaadmin**组。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

15.13.18 执行 Kafka Topic 设置 ACL 操作失败，提示 NoAuthException

问题背景与现象

在使用Kafka客户端命令设置Topic ACL权限时，发现Topic无法被设置。

```
kafka-acls.sh --authorizer-properties zookeeper.connect=10.5.144.2:2181/kafka  --topic topic_acl --producer
--add --allow-principal User:test_acl
```

提示错误NoAuthException: KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002。

具体如下：

```
Error while executing ACL command: org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002
org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException:
KeeperErrorCode = NoAuth for /kafka-acl-changes/acl_changes_0000000002
at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:995)
at org.I0Itec.zkclient.ZkClient.delete(ZkClient.java:1038)
at kafka.utils.ZkUtils.deletePath(ZkUtils.scala:499)
at kafka.common.ZkNodeChangeNotificationListener$$anonfun$purgeObsoleteNotifications
$1.apply(ZkNodeChangeNotificationListener.scala:118)
at kafka.common.ZkNodeChangeNotificationListener$$anonfun$purgeObsoleteNotifications
$1.apply(ZkNodeChangeNotificationListener.scala:112)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
at
kafka.common.ZkNodeChangeNotificationListener.purgeObsoleteNotifications(ZkNodeChangeNotificationLis
```

```
tener.scala:112)
at kafka.common.ZkNodeChangeNotificationListener.kafka$common$ZkNodeChangeNotificationListener$
$processNotifications(ZkNodeChangeNotificationListener.scala:97)
at
kafka.common.ZkNodeChangeNotificationListener.processAllNotifications(ZkNodeChangeNotificationListener
.scala:77)
at kafka.common.ZkNodeChangeNotificationListener.init(ZkNodeChangeNotificationListener.scala:65)
at kafka.security.auth.SimpleAclAuthorizer.configure(SimpleAclAuthorizer.scala:136)
at kafka.admin.AclCommand$.withAuthorizer(AclCommand.scala:73)
at kafka.admin.AclCommand$.addAcl(AclCommand.scala:80)
at kafka.admin.AclCommand$.main(AclCommand.scala:48)
at kafka.admin.AclCommand.main(AclCommand.scala)
Caused by: org.apache.zookeeper.KeeperException$NoAuthException: KeeperErrorCode = NoAuth for /kafka-
acl-changes/acl_changes_0000000002
at org.apache.zookeeper.KeeperException.create(KeeperException.java:117)
at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
at org.apache.zookeeper.ZooKeeper.delete(ZooKeeper.java:1416)
at org.I0Itec.zkclient.ZkConnection.delete(ZkConnection.java:104)
at org.I0Itec.zkclient.ZkClient$11.call(ZkClient.java:1042)
at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:985)
```

可能原因

用户不属于**kafkaadmin**组，Kafka提供安全访问接口，**kafkaamdin**组用户才可以进行设置操作。

原因分析

1. 使用客户端命令，打印NoAuthException异常。
2. 通过客户端命令**klist**查询当前认证用户：

   ```
   [root@10-10-144-2 client]# klist
   Ticket cache: FILE:/tmp/krb5cc_0
   Default principal: test@HADOOP.COM

   Valid starting     Expires            Service principal
   01/25/17 11:06:48  01/26/17 11:06:45  krbtgt/HADOOP.COM@HADOOP.COM
   ```

   如上例中当前认证用户为**test**。
3. 通过命令**id**查询用户组信息。

   ```
   [root@10-10-144-2 client]# id test
   uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10003(kafka)
   ```

解决办法

MRS Manager界面操作：

步骤1 登录MRS Manager。

步骤2 选择“系统设置 > 用户管理”。

步骤3 在操作用户对应的“操作”列，单击“修改”。

步骤4 为用户加入**kafkaadmin**组。

步骤5 通过命令**id**查询用户组信息。

```
[root@host1 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager。

步骤2 选择“系统 > 权限 > 用户”。

步骤3 在使用的用户所在行的单击“修改”。

步骤4 为用户添加**kafkaadmin**组。单击“确定”完成修改操作。

步骤5 通过命令**id**查询用户组信息。

```
[root@10-10-144-2 client]# id test
uid=20032(test) gid=10001(hadoop) groups=10001(hadoop),9998(ficommon),10002(kafkaadmin),
10003(kafka)
```

----结束

15.13.19 执行 Kafka Topic 创建操作，发现无法创建提示 NoNode for /brokers/ids

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现Topic无法被创建。

```
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper
192.168.234.231:2181
```

提示错误NoNodeException: KeeperErrorCode = NoNode for /brokers/ids。

具体如下：

```
Error while executing topic command : org.apache.zookeeper.KeeperException$NoNodeException:
KeeperErrorCode = NoNode for /brokers/ids
[2017-09-17 16:35:28,520] ERROR org.I0Itec.zkclient.exception.ZkNoNodeException:
org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /brokers/ids
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:47)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:995)
    at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:675)
    at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:671)
    at kafka.utils.ZkUtils.getChildren(ZkUtils.scala:541)
    at kafka.utils.ZkUtils.getSortedBrokerList(ZkUtils.scala:176)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:235)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:105)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:60)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
Caused by: org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /
brokers/ids
    at org.apache.zookeeper.KeeperException.create(KeeperException.java:115)
    at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
    at org.apache.zookeeper.ZooKeeper.getChildren(ZooKeeper.java:2256)
    at org.apache.zookeeper.ZooKeeper.getChildren(ZooKeeper.java:2284)
    at org.I0Itec.zkclient.ZkConnection.getChildren(ZkConnection.java:114)
    at org.I0Itec.zkclient.ZkClient$4.call(ZkClient.java:678)
    at org.I0Itec.zkclient.ZkClient$4.call(ZkClient.java:675)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:985)
    ... 8 more
 (kafka.admin.TopicCommand$)
```

可能原因

- Kafka服务处于停止状态。
- 客户端命令中zookeeper地址参数配置错误。

原因分析

1. 使用客户端命令，打印NoNodeException异常。

```
Error while executing topic command : org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /brokers/ids
[2017-09-17 16:35:28,520] ERROR org.I0Itec.zkclient.exception.ZkNoNodeException: org.apache.zookeeper.KeeperException$NoNodeException: KeeperErrorCode = NoNode for /brokers/ids
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:47)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:995)
    at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:675)
    at org.I0Itec.zkclient.ZkClient.getChildren(ZkClient.java:671)
    at kafka.utils.ZkUtils.getChildren(ZkUtils.scala:541)
    at kafka.utils.ZkUtils.getSortedBrokerList(ZkUtils.scala:176)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:235)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:105)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:60)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
```

2. 通过Manager查看Kafka服务是否处于正常状态。
3. 检查客户端命令中ZooKeeper地址是否正确，访问ZooKeeper上所存放的Kafka信息，其路径（Znode）应该加上/kafka，发现配置中缺少/kafka：

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper 192.168.234.231:2181
```

解决办法

步骤1 保证Kafka服务处于正常状态。

步骤2 创建命令中ZooKeeper地址信息需要添加/kafka。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper 192.168.234.231:2181/kafka
```

----结束

15.13.20 执行 Kakfa Topic 创建操作，发现无法创建提示 replication factor larger than available brokers

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现Topic无法被创建。

```
kafka-topics.sh --create  --replication-factor 2 --partitions 2 --topic test --zookeeper 192.168.234.231:2181
```

提示错误replication factor larger than available brokers。

具体如下：

```
Error while executing topic command : replication factor: 2 larger than available brokers: 0
[2017-09-17 16:44:12,396] ERROR kafka.admin.AdminOperationException: replication factor: 2 larger than available brokers: 0
    at kafka.admin.AdminUtils$.assignReplicasToBrokers(AdminUtils.scala:117)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:403)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:110)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:61)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
 (kafka.admin.TopicCommand$)
```

可能原因

- Kafka服务处于停止状态。
- Kafka服务当前可用Broker小于设置的replication-factor。
- 客户端命令中Zookeeper地址参数配置错误。

原因分析

1. 使用客户端命令，打印replication factor larger than available brokers异常。

```
Error while executing topic command : replication factor: 2 larger than available brokers: 0
[2017-09-17 16:44:12,396] ERROR kafka.admin.AdminOperationException: replication factor: 2 larger than available brokers: 0
    at kafka.admin.AdminUtils$.assignReplicasToBrokers(AdminUtils.scala:117)
    at kafka.admin.AdminUtils$.createTopic(AdminUtils.scala:403)
    at kafka.admin.TopicCommand$.createTopic(TopicCommand.scala:110)
    at kafka.admin.TopicCommand$.main(TopicCommand.scala:61)
    at kafka.admin.TopicCommand.main(TopicCommand.scala)
 (kafka.admin.TopicCommand$)
```

2. 通过Manager参看Kafka服务是否处于正常状态，当前可用Broker是否小于设置的replication-factor。
3. 检查客户端命令中ZooKeeper地址是否正确，访问ZooKeeper上所存放的Kafka信息，其路径（Znode）应该加上/kafka，发现配置中缺少/kafka。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 2 --partitions 2 --topic test --zookeeper 192.168.234.231:2181
```

解决办法

步骤1 保证Kafka服务处于正常状态，且可用Broker不小于设置的replication-factor。

步骤2 创建命令中ZooKeeper地址信息需要添加/kafka。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper 192.168.234.231:2181/kafka
```

----结束

15.13.21 Consumer 消费数据存在重复消费现象

问题背景与现象

当数据量较大时会频繁的发生rebalance导致出现重复消费的情况，关键日志如下：

```
2018-05-12 10:58:42,561 | INFO | [kafka-request-handler-3] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 118 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:58:43,245 | INFO | [kafka-request-handler-5] | [GroupCoordinator 2]: Stabilized group DemoConsumer generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:58:43,560 | INFO | [kafka-request-handler-7] | [GroupCoordinator 2]: Assignment received from leader for group DemoConsumer for generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,562 | INFO | [executor-Heartbeat] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 119 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,790 | INFO | [kafka-request-handler-3] | [GroupCoordinator 2]: Stabilized group DemoConsumer generation 120 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:13,791 | INFO | [kafka-request-handler-0] | [GroupCoordinator 2]: Assignment received from leader for group DemoConsumer for generation 120 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 10:59:43,802 | INFO | [kafka-request-handler-2] | Rolled new log segment for '__consumer_offsets-17' in 2 ms. | kafka.log.Log (Logging.scala:68)
2018-05-12 10:59:52,456 | INFO | [group-metadata-manager-0] | [Group Metadata Manager on Broker 2]:
```

```
Removed 0 expired offsets in 0 milliseconds. | kafka.coordinator.GroupMetadataManager (Logging.scala:68)
2018-05-12 11:00:49,772 | INFO | [kafka-scheduler-6] | Deleting segment 0 from log __consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:49,773 | INFO | [kafka-scheduler-6] | Deleting index /srv/BigData/kafka/data4/kafka-logs/__consumer_offsets-17/00000000000000000000.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:00:49,773 | INFO | [kafka-scheduler-2] | Deleting segment 2147948547 from log __consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:49,773 | INFO | [kafka-scheduler-4] | Deleting segment 4282404355 from log __consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:49,775 | INFO | [kafka-scheduler-2] | Deleting index /srv/BigData/kafka/data4/kafka-logs/__consumer_offsets-17/00000000002147948547.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:00:49,775 | INFO | [kafka-scheduler-4] | Deleting index /srv/BigData/kafka/data4/kafka-logs/__consumer_offsets-17/00000000004282404355.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:00:50,533 | INFO | [kafka-scheduler-6] | Deleting segment 4283544095 from log __consumer_offsets-17. | kafka.log.Log (Logging.scala:68)
2018-05-12 11:00:50,569 | INFO | [kafka-scheduler-6] | Deleting index /srv/BigData/kafka/data4/kafka-logs/__consumer_offsets-17/00000000004283544095.index.deleted | kafka.log.OffsetIndex (Logging.scala:68)
2018-05-12 11:02:21,178 | INFO | [kafka-request-handler-2] | [GroupCoordinator 2]: Preparing to restabilize group DemoConsumer with old generation 120 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 11:02:22,839 | INFO | [kafka-request-handler-4] | [GroupCoordinator 2]: Stabilized group DemoConsumer generation 121 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 11:02:23,169 | INFO | [kafka-request-handler-1] | [GroupCoordinator 2]: Assignment received from leader for group DemoConsumer for generation 121 | kafka.coordinator.GroupCoordinator (Logging.scala:68)
2018-05-12 11:02:49,913 | INFO | [kafka-request-handler-6] | Rolled new log segment for '__consumer_offsets-17' in 2 ms. | kafka.log.Log (Logging.scala:68)
```

其中Preparing to restabilize group DemoConsumer with old generation表示正在发生rebalance。

可能原因

参数设置不合理。

原因分析

原因：由于参数设置不当，数据量大时数据处理时间过长，导致频繁发生balance，此时offset无法正常提交，导致重复消费数据。

原理：每次poll的数据处理完后才提交offset，如果poll数据后的处理时长超出了session.timeout.ms的设置时长，此时发生rebalance导致本次消费失败，已经消费数据的offset无法正常提交，所以下次重新消费时还是在旧的offset消费数据，从而导致消费数据重复。

解决办法

建议用户在Manager页面调整以下服务参数：

request.timeout.ms=100000

session.timeout.ms=90000

max.poll.records=50

heartbeat.interval.ms=3000

其中：

request.timeout.ms要比session.timeout.ms大10s。

session.timeout.ms的大小设置要在服务端参数group.min.session.timeout.ms和group.max.session.timeout.ms之间。

以上参数可以根据实际情况进行适当的调整，特别是max.poll.records，这个参数是为了控制每次poll数据的records量，保证每次的处理时长尽量保持稳定。目的是为了保证poll数据以后的处理时间不要超过session.timeout.ms的时间。

参考信息

- poll之后的数据处理效率要高，不要阻塞下一次poll
- poll方法和数据处理建议异步处理

15.13.22 执行 Kafka Topic 创建操作，发现 Partition 的 Leader 显示为 none

问题背景与现象

在使用Kafka客户端命令创建Topic时，发现创建Topic Partition的Leader显示为none。

```
[root@10-10-144-2 client]#
kafka-topics.sh --create  --replication-factor 1 --partitions 2 --topic test --zookeeper 10.6.92.36:2181/kafka

Created topic "test".
```

```
[root@10-10-144-2 client]#
kafka-topics.sh --describe --zookeeper 10.6.92.36:2181/kafka

Topic:test      PartitionCount:2        ReplicationFactor:2     Configs:
        Topic: test     Partition: 0    Leader: none        Replicas: 2,3   Isr:
        Topic: test     Partition: 1    Leader: none        Replicas: 3,1   Isr:
```

可能原因

- Kafka服务处于停止状态。
- 找不到用户组信息。

原因分析

1. 查看kafka服务状态及监控指标。
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Kafka ，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Kafka，查看当前Kafka状态，发现状态为良好，且监控指标内容显示正确。
2. 在Kafka概览页面获取Controller节点信息。
3. 登录Controller所在节点，通过**cd /var/log/Bigdata/kafka/broker**命令进入节点日志目录，在state-change.log发现存在ZooKeeper权限异常，提示NoAuthException。

   ```
   2018-05-31 09:20:42,436 | ERROR | [ZkClient-EventThread-34-10.6.92.36:24002,10.6.92.37:24002,10.6.92.38:24002/kafka] | Controller 4 epoch 6 initiated state change for partition [test,1] from NewPartition to OnlinePartition failed | state.change.logger (Logging.scala:103)

   org.I0Itec.zkclient.exception.ZkException: org.apache.zookeeper.KeeperException$NoAuthException: KeeperErrorCode = NoAuth for /brokers/topics/test/partitions
    at org.I0Itec.zkclient.exception.ZkException.create(ZkException.java:68)
    at org.I0Itec.zkclient.ZkClient.retryUntilConnected(ZkClient.java:1000)
   ```

```
at org.I0Itec.zkclient.ZkClient.create(ZkClient.java:527)
at org.I0Itec.zkclient.ZkClient.createPersistent(ZkClient.java:293)
```

4. 查看对应时间段ZooKeeper审计日志，权限异常。

```
2018-05-31 09:20:42,421 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/0/state    result=failure
2018-05-31 09:20:42,423 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/0    result=failure
2018-05-31 09:20:42,435 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions    result=failure
2018-05-31 09:20:42,439 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/1/state    result=failure
2018-05-31 09:20:42,441 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions/1    result=failure
2018-05-31 09:20:42,453 | ERROR | CommitProcWorkThread-1 | session=0xc3000007015d5a18
user=10.6.92.39,kafka/hadoop.hadoop.com@HADOOP.COM,kafka/
hadoop.hadoop.com@HADOOP.COM    ip=10.6.92.39    operation=create znode
target=ZooKeeperServer    znode=/kafka/brokers/topics/test/partitions    result=failure
```

5. 在ZooKeeper各个实例节点上执行**id -Gn kafka**命令，发现有一个节点无法查询用户组信息。

```
[root @bdpsit3ap03 ~]# id -Gn kafka
id:  kafka: No such user
[root @bdpsit3ap03 ~]#
```

6. MRS集群中的用户管理由LDAP服务管理提供，又依赖于操作系统的sssd（red hat），nscd（suse）服务，用户的建立到同步到sssd服务需要一定时间，如果此时用户没有生效，或者sssd版本存在bug的情况下，某些情况下在ZooKeeper节点会出现用户无效的情况，导致创建Topic异常。

解决办法

步骤1 重启sssd/nscd服务。

- Red Hat

```
service sssd restart
```

- SUSE

```
sevice nscd restart
```

步骤2 重启相关服务后，在节点通过**id username**命令查看相应用户信息是否已有效。

----结束

15.13.23 Kafka 安全使用说明

Kafka API 简单说明

- 新Producer API

  指org.apache.kafka.clients.producer.KafkaProducer中定义的接口，在使用“kafka-console-producer.sh”时，默认使用此API。

- 旧Producer API

  指kafka.producer.Producer中定义的接口，在使用“kafka-console-producer.sh”时，加“--old-producer”参数会调用此API。

- 新Consumer API

  指org.apache.kafka.clients.consumer.KafkaConsumer中定义的接口，在使用“kafka-console-consumer.sh”时，加“--new-consumer”参数会调用此API。

- 旧Consumer API

  指kafka.consumer.ConsumerConnector中定义的接口，在使用“kafka-console-consumer.sh”时，默认使用此API。

说明

新Producer API和新Consumer API，在下文中统称为新API。

Kafka 访问协议说明

Kafka当前支持四种协议类型的访问：PLAINTEXT、SSL、SASL_PLAINTEXT、SASL_SSL。

Kafka服务启动时，默认会启动PLAINTEXT和SASL_PLAINTEXT两种协议类型的访问监听。可通过设置Kafka服务配置参数“ssl.mode.enable”为“true”，来启动SSL和SASL_SSL两种协议类型的访问监听。

下表是四中协议类型的简单说明：

协议类型	说明	支持的API	默认端口
PLAINTEXT	支持无认证的明文访问	新API和旧API	9092
SASL_PLAINTEXT	支持Kerberos认证的明文访问	新API	21007
SSL	支持无认证的SSL加密访问	新API	9093
SASL_SSL	支持Kerberos认证的SSL加密访问	新API	21009

Topic 的 ACL 设置

Kafka支持安全访问，因此可以针对Topic进行ACL设置，从而控制不同的用户可以访问不同的Topic。Topic的权限信息，需要在Linux客户端上，使用“kafka-acls.sh”脚本进行查看和设置。

- 操作场景

  该任务指导Kafka管理员根据业务需求，为其他使用Kafka的系统用户授予相关Topic的特定权限。

  Kafka默认用户组信息表所示。

用户组名	描述
kafkaadmin	Kafka管理员用户组。添加入本组的用户，拥有所有Topic的创建，删除，授权及读写权限。
kafkasuperuser	添加入本组的用户，拥有所有Topic的读写权限。

用户组名	描述
kafka	Kafka普通用户组。添加入本组的用户，需要被kafkaadmin组用户授予特定Topic的读写权限，才能访问对应Topic。

- 前提条件
  a. 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组）。
  b. 已安装Kafka客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。
     cd /opt/kafkaclient
  c. 执行以下命令，配置环境变量。
     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通集群跳过此步骤）
     kinit *组件业务用户*
  e. 执行以下命令，切换到Kafka客户端安装目录。
     cd Kafka/kafka/bin
  f. 使用“kafka-acl.sh”进行用户授权常用命令如下：
     - 查看某Topic权限控制列表：
       ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --list --topic <Topic名称>
     - 添加给某用户Producer权限：
       ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --add --allow-principal User:<用户名> --producer --topic <Topic名称>
     - 删除某用户Producer权限：
       ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --remove --allow-principal User:<用户名> --producer --topic <Topic名称>
     - 添加给某用户Consumer权限：
       ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --add --allow-principal User:<用户名> --consumer --topic <Topic名称> --group <消费者组名称>
     - 删除某用户Consumer权限：
       ./kafka-acls.sh --authorizer-properties zookeeper.connect=<ZooKeeper集群业务IP:2181/kafka > --remove

--allow-principal User:<用户名> --consumer --topic <Topic名称> --group <消费者组名称>

针对不同的 Topic 访问场景，Kafka 新旧 API 使用说明

- 场景一：访问设置了ACL的Topic

使用的API	用户属组	客户端参数	服务端参数	访问的端口
新API	用户需满足以下条件之一即可： • 属于系统管理员组 • 属于kafkaadmin组 • 属于kafkasuperuser组 • 被授权的kafka组的用户	security.protocol=SASL_PLAINTEXT sasl.kerberos.service.name = kafka	-	sasl.port（默认21007）
		security.protocol=SASL_SSL sasl.kerberos.service.name = kafka	ssl.mode.enable配置为true	sasl-ssl.port（默认21009）
旧API	不涉及	不涉及	不涉及	不涉及

- 场景二：访问未设置ACL的Topic

使用的API	用户属组	客户端参数	服务端参数	访问的端口
新API	用户需满足以下条件之一： • 属于系统管理员组 • 属于kafkaadmin组 • 属于kafkasuperuser组	security.protocol=SASL_PLAINTEXT sasl.kerberos.service.name = kafka	-	sasl.port（默认21007）
	用户属于kafka组		allow.everyone.if.no.acl.found配置为true	sasl.port（默认21007）

使用的API	用户属组	客户端参数	服务端参数	访问的端口
	用户需满足以下条件之一： • 属于系统管理员组 • 属于kafkaadmin组 • kafkasuperuser组用户	security.protocol=SASL_SSLsasl.kerberos.service.name = kafka	ssl-enable配置为“true”	sasl-ssl.port（默认21009）
	用户属于kafka组		allow.everyone.if.no.acl.found配置为“true” ssl-enable配置为“true”	sasl-ssl.port（默认21009）
	-	security.protocol=PLAINTEXT	allow.everyone.if.no.acl.found配置为“true”	port（默认21005）
	-	security.protocol=SSL	allow.everyone.if.no.acl.found配置为“true” ssl-enable配置为“true”	ssl.port（默认21008）
旧Producer	-	-	allow.everyone.if.no.acl.found配置为“true”	port（默认21005）
旧Consumer	-	-	allow.everyone.if.no.acl.found配置为“true”	ZooKeeper服务端口：clientPort（默认24002）

15.13.24 如何获取 Kafka Consumer Offset 信息

问题背景与现象

使用Kafka Consumer消费数据时，如何获取Kafka Consumer Offset相关信息?

Kafka API 简单说明

- 新Producer API

  指org.apache.kafka.clients.producer.KafkaProducer中定义的接口，在使用“kafka-console-producer.sh”时，默认使用此API。

- 旧Producer API

  指kafka.producer.Producer中定义的接口，在使用“kafka-console-producer.sh”时，加“--old-producer”参数会调用此API。

- 新Consumer API

  指org.apache.kafka.clients.consumer.KafkaConsumer中定义的接口，在使用“kafka-console-consumer.sh”时，加“--new-consumer”参数会调用此API。

- 旧Consumer API

  指kafka.consumer.ConsumerConnector中定义的接口，在使用“kafka-console-consumer.sh”时，默认使用此API。

说明

新Producer API和新Consumer API，在下文中统称为新API。

处理步骤

旧Consumer API

- 前提条件
  a. 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组）。
  b. 已安装Kafka客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/kafkaclient”。

     cd /opt/kafkaclient
  c. 执行以下命令，配置环境变量。

     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通模式跳过此步骤）

     kinit *组件业务用户*
  e. 执行以下命令，切换到Kafka客户端安装目录。

     cd Kafka/kafka/bin
  f. 执行以下命令，获取consumer offset metric信息。

     ```
     bin/kafka-consumer-groups.sh --zookeeper <zookeeper_host:port>/kafka --list
     ```

     ```
     bin/kafka-consumer-groups.sh --zookeeper <zookeeper_host:port>/kafka --describe --group test-consumer-group
     ```

     例如：

     ```
     kafka-consumer-groups.sh --zookeeper 192.168.100.100:2181/kafka --list
     kafka-consumer-groups.sh --zookeeper 192.168.100.100:2181/kafka --describe --group test-consumer-group
     ```

新Consumer API

- 前提条件
  a. 系统管理员已明确业务需求，并准备一个Kafka管理员用户（属于kafkaadmin组）。
  b. 已安装Kafka客户端。

- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/client”。
     cd /opt/client
  c. 执行以下命令，配置环境变量。
     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通模式跳过此步骤）
     kinit *组件业务用户*
  e. 执行以下命令，切换到Kafka客户端安装目录。
     cd Kafka/kafka/bin
  f. 执行以下命令，获取consumer offset metric信息。
     kafka-consumer-groups.sh --bootstrap-server <broker_host:port> --describe --group my-group
     例如：
     kafka-consumer-groups.sh --bootstrap-server 192.168.100.100:9092 --describe --group my-group

15.13.25 如何针对 Topic 进行配置增加和删除

问题背景与现象

使用Kafka过程中常常需要对特定Topic进行配置或者修改。

Topic级别可以修改参数列表：

```
cleanup.policy
compression.type
delete.retention.ms
file.delete.delay.ms
flush.messages
flush.ms
index.interval.bytes
max.message.bytes
min.cleanable.dirty.ratio
min.insync.replicas
preallocate
retention.bytes
retention.ms
segment.bytes
segment.index.bytes
segment.jitter.ms
segment.ms
unclean.leader.election.enable
```

处理步骤

- 前提条件
  已安装Kafka客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/client”。
     cd /opt/client

c. 执行以下命令，配置环境变量。

source bigdata_env

d. 执行以下命令，进行用户认证。（普通模式跳过此步骤）

kinit *组件业务用户*

e. 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

f. 执行以下命令，针对Topic修改配置和删除配置。

kafka-topics.sh --alter --topic <topic_name> --zookeeper <zookeeper_host:port>/kafka --config <name=value>

kafka-topics.sh --alter --topic <topic_name> --zookeeper <zookeeper_host:port>/kafka --delete-config <name>

例如：

kafka-topics.sh --alter --topic test1 --zookeeper 192.168.100.100:2181/kafka --config retention.ms=86400000

kafka-topics.sh --alter --topic test1 --zookeeper 192.168.100.100:2181/kafka --delete-config retention.ms

g. 执行以下命令，查询topic信息。

kafka-topics.sh --describe -topic <topic_name> --zookeeper <zookeeper_host:port>/kafka

15.13.26 如何读取“__consumer_offsets”内部 topic 的内容

用户问题

kafka如何将consumer 消费的offset保存在内部topic“ __consumer_offsets”中?

处理步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤3 执行以下命令，配置环境变量。

source bigdata_env

步骤4 执行以下命令，进行用户认证。（普通集群跳过此步骤）

kinit *组件业务用户*

步骤5 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤6 执行以下命令，获取consumer offset metric信息。

kafka-console-consumer.sh --topic __consumer_offsets --zookeeper <zk_host:port>/kafka --formatter "kafka.coordinator.group.GroupMetadataManager\$OffsetsMessageFormatter" --consumer.config <property file> --from-beginning

其中<property file>配置文件中需要增加如下内容。

```
exclude.internal.topics = false
```

例如：

kafka-console-consumer.sh --topic __consumer_offsets --zookeeper 10.5.144.2:2181/kafka --formatter "kafka.coordinator.group.GroupMetadataManager\$OffsetsMessageFormatter" --consumer.config ../config/consumer.properties --from-beginning

```
[example-group1,test2,0]::[OffsetMetadata[0,NO_METADATA],CommitTime 1487121209218,ExpirationTime 148720760
9218]
[example-group1,test2,1]::[OffsetMetadata[0,NO_METADATA],CommitTime 1487121209218,ExpirationTime 148720760
9218]
[example-group1,test2,0]::[OffsetMetadata[2,NO_METADATA],CommitTime 1487121269208,ExpirationTime 148720760
9208]
[example-group1,test2,1]::[OffsetMetadata[1,NO_METADATA],CommitTime 1487121269208,ExpirationTime 148720760
```

----结束

15.13.27 如何配置客户端 shell 命令的日志

用户问题

如何设置客户端shell命令的日志输出级别?

处理步骤

步骤1 以客户端安装用户，登录安装Kafka客户端的节点。

步骤2 切换到Kafka客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤3 切换到Kafka客户端配置目录。

cd Kafka/kafka/config

步骤4 编辑tools-log4j.properties文件，将WARN修改为INFO，并保存。

```
log4j.rootLogger=WARN, stderr

log4j.appender.stderr=org.apache.log4j.ConsoleAppender
log4j.appender.stderr.layout=org.apache.log4j.PatternLayout
log4j.appender.stderr.layout.ConversionPattern=[%d] %p %m (%c)%n
log4j.appender.stderr.Target=System.err
```

```
log4j.rootLogger=INFO, stderr

log4j.appender.stderr=org.apache.log4j.ConsoleAppender
log4j.appender.stderr.layout=org.apache.log4j.PatternLayout
log4j.appender.stderr.layout.ConversionPattern=[%d] %p %m (%c)%n
log4j.appender.stderr.Target=System.err
```

步骤5 切换到Kafka客户端安装目录，例如“/opt/client”。

cd /opt/client

步骤6 执行以下命令，配置环境变量。

source bigdata_env

步骤7 执行以下命令，进行用户认证。（普通集群跳过此步骤）

kinit *组件业务用户*

步骤8 执行以下命令，切换到Kafka客户端安装目录。

cd Kafka/kafka/bin

步骤9 执行以下命令，获取topic信息，在控制台可见日志打印。

```
kafka-topics.sh --list --zookeeper 10.5.144.2:2181/kafka
[2017-02-17 14:34:27,005] INFO JAAS File name: /opt/client/Kafka/./kafka/config/jaas.conf
(org.I0Itec.zkclient.ZkClient)
[2017-02-17 14:34:27,007] INFO Starting ZkClient event thread. (org.I0Itec.zkclient.ZkEventThread)
[2017-02-17 14:34:27,013] INFO Client environment:zookeeper.version=V100R002C10, built on 05/12/2016
08:56 GMT (org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:host.name=10-10-144-2
(org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:java.version=1.8.0_72
(org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:java.vendor=Oracle Corporation
(org.apache.zookeeper.ZooKeeper)
[2017-02-17 14:34:27,013] INFO Client environment:java.home=/opt/client/JDK/jdk/jre
(org.apache.zookeeper.ZooKeeper)
Test
__consumer_offsets
counter
test
test2
test3
test4
```

----结束

15.13.28 如何获取 Topic 的分布信息

用户问题

如何获取Topic在Broker实例的分布信息?

前置操作

- 前提条件
  已安装Kafka、ZooKeeper客户端。
- 操作步骤
  a. 以客户端安装用户，登录安装Kafka客户端的节点。
  b. 切换到Kafka客户端安装目录，例如“/opt/client”。
     cd /opt/client
  c. 执行以下命令，配置环境变量。
     source bigdata_env
  d. 执行以下命令，进行用户认证。（普通集群跳过此步骤）
     kinit ***组件业务用户***
  e. 执行以下命令，切换到Kafka客户端安装目录。
     cd Kafka/kafka/bin

f. 执行Kafka相关命令，获取Topic分布信息和副本同步信息，观察返回结果。

kafka-topics.sh --describe --zookeeper <zk_host:port/chroot>

例如：

```
[root@mgtdat-sh-3-01-3 client]#kafka-topics.sh --describe --zookeeper 10.149.0.90:2181/kafka
Topic:topic1   PartitionCount:2  ReplicationFactor:2     Configs:
Topic: topic1  Partition: 0 Leader: 26 Replicas: 23,25 Isr: 26
Topic: topic1  Partition: 1 Leader: 24 Replicas: 24,23 Isr: 24,23
```

其中，Replicas对应副本分布信息，Isr对应副本同步信息。

处理方法 1

1. 在ZooKeeper中查询Broker ID的对应关系。

   sh zkCli.sh -server <zk_host:port>

2. 在ZooKeeper客户端执行如下命令。

   ls /kafka/brokers/ids

   get /kafka/brokers/ids/<查询出的Broker id>

   例如：

```
[root@node-master1gAMQ kafka]# zkCli.sh -server node-master1gAMQ:2181
Connecting to node-master1gAMQ:2181
Welcome to ZooKeeper!
JLine support is enabled

WATCHER::

WatchedEvent state:SyncConnected type:None path:null
[zk: node-master1gAMQ:2181(CONNECTED) 0] ls /kafka/brokers/
ids     seqid    topics
[zk: node-master1gAMQ:2181(CONNECTED) 0] ls /kafka/brokers/ids
[1]
[zk: node-master1gAMQ:2181(CONNECTED) 1] get /kafka/brokers/ids/1
{"listener_security_protocol_map":{"PLAINTEXT":"PLAINTEXT","SSL":"SSL"},"endpoints":["PLAINTEXT://192.168.2.242:9092","SSL://192.168.2.242:9093"],"rack":"/default/rack0","jmx_port":21006,"host":"192.168.2.242","timestamp":"1580886124398","port":9092,"version":4}
[zk: node-master1gAMQ:2181(CONNECTED) 2]
```

处理方法 2

获取节点和broker ID的对应关系

kafka-broker-info.sh --zookeeper <zk_host:port/chroot>

例如：

```
[root@node-master1gAMQ kafka]# bin/kafka-broker-info.sh --zookeeper 192.168.2.70:2181/kafka
Broker_ID     IP_Address
---------------------------
 1          192.168.2.242
```

15.13.29 Kafka 高可靠使用说明

Kafka 高可靠、高可用说明

Kafka消息传输保障机制，可以通过配置不同的参数来保障消息传输，进而满足不同的性能和可靠性要求的应用场景。

- **Kafka高可用、高性能**

如果业务需要保证高可用和高性能，可以采用参数：

参数	默认值	说明
unclean.leader.election.enable	true	是否允许不在ISR中的副本被选举为Leader，若设置为true，可能会造成数据丢失。
auto.leader.rebalance.enable	true	是否使用Leader自动均衡功能。 如果设为true，Controller会周期性的为所有节点的每个分区均衡Leader，将Leader分配给更优先的副本。
acks	1	需要Leader确认消息是否已经接收并认为已经处理完成。该参数会影响消息的可靠性和性能。 ● acks=0 ： 如果设置为0，Producer将不会等待服务端任何响应。消息将会被认为成功。 ● acks=1 ： 如果设置为1，当副本所在Leader确认数据已写入，但是其不会等待所有的副本完全写入即返回响应。在这种情况下，如果Leader确认后但是副本未同步完成时Leader异常，那么数据就会丢失。 ● acks=-1 ： 如果设置为-1（all），意味着等待所有的同步副本确认后才认为成功，配合min.insync.replicas可以确保多副本写入成功，只要有一个副本保持活跃状态，记录将不会丢失。 **说明** 该参数在kafka客户端配置文件中配置。
min.insync.replicas	1	当Producer设置acks为-1时，指定需要写入成功的副本的最小数目。

配置高可用、高性能的影响：

须知

配置高可用、高性能模式后，数据可靠性会降低。在磁盘故障、节点故障等场景下存在数据丢失风险。

- **Kafka高可靠性配置说明**

  如果业务需要保证数据高可靠性，可以采用相关参数：

参数	建议值	说明
unclean.leader.election.enable	false	是否允许不在ISR中的副本被选举为Leader。

acks	-1	Producer需要Leader确认消息是否已经接收并认为已经处理完成。 acks=-1需要表示等待在ISR列表的副本都确认接收到消息并处理完成才表示消息成功。配合min.insync.replicas可以确保多副本写入成功，只要有一个副本保持活跃状态，记录将不会丢失。 **说明** 该参数在kafka客户端配置文件中配置。
min.insync.replicas	2	当Producer设置acks为-1时，指定需要写入成功的副本的最小数目。 需要满足min.insync.replicas <= replication.factor。

配置高可靠性的影响：

- 性能降低：

  需要所有的ISR列表副本，且满足最小成功的副本数确认写入成功。这样会导致单条消息时延增加，客户端处理能力下降，具体性能以现场实际测试数据为准。

- 可用性降低：

  不允许不在ISR中的副本被选举为Leader。如果Leader下线时，其他副本均不在ISR列表中，那么该分区将保持不可用，直到Leader节点恢复。

  需要所有的ISR列表副本，且满足最小成功的副本数确认写入成功。当分区的一个副本所在节点故障时，无法满足最小成功的副本数，那么将会导致业务写入失败。

配置影响

请根据业务场景对可靠性和性能要求进行评估，采用合理参数配置。

说明

- 对于价值数据，两种场景下建议Kafka数据目录磁盘配置raid1或者raid5，从而提高单个磁盘故障情况下数据可靠性。
- 不同Producer API对应的acks参数名称不同
  - 新Producer API

    指**org.apache.kafka.clients.producer.KafkaProducer**中定义的接口，acks配置为acks。
  - 旧Producer API

    指**kafka.producer.Producer**中定义的接口，acks配置名称为request.required.acks。
- 参数配置项均为Topic级别可修改的参数，默认采用服务级配置。可针对不同Topic可靠性要求对Topic进行单独配置。

  例如，配置Topic名称为test的可靠性参数：

  kafka-topics.sh --zookeeper 192.168.1.205:2181/kafka --alter --topic test --config unclean.leader.election.enable=false --config min.insync.replicas=2

  其中192.168.1.205为ZooKeeper业务IP地址。
- 如果修改服务级配置需要重启Kafka，建议在变更窗口做服务级配置修改。

15.13.30 Kafka 生产者写入单条记录过长问题

问题背景与现象

用户在开发一个Kafka应用，作为一个生产者调用新接口（org.apache.kafka.clients.producer.*）往Kafka写数据，单条记录大小为1100055，超过了kafka配置文件server.properties中message.max.bytes=1000012。用户修改了Kafka服务配置中message.max.bytes大小为5242880，同时也将replica.fetch.max.bytes大小修改为5242880后，仍然无法成功。报异常大致如下：

```
.....
14749 [Thread-0] INFO  com.XXXXXX.bigdata.kafka.example.NewProducer  - The ExecutionException occured : {}.
java.util.concurrent.ExecutionException: org.apache.kafka.common.errors.RecordTooLargeException: The message is 1100093 bytes when serialized which is larger than the maximum request size you have configured with the max.request.size configuration.
at org.apache.kafka.clients.producer.KafkaProducer$FutureFailure.<init>(KafkaProducer.java:739)
at org.apache.kafka.clients.producer.KafkaProducer.doSend(KafkaProducer.java:483)
at org.apache.kafka.clients.producer.KafkaProducer.send(KafkaProducer.java:430)
at org.apache.kafka.clients.producer.KafkaProducer.send(KafkaProducer.java:353)
at com.XXXXXX.bigdata.kafka.example.NewProducer.run(NewProducer.java:150)
Caused by: org.apache.kafka.common.errors.RecordTooLargeException: The message is **** bytes when serialized which is larger than the maximum request size you have configured with the max.request.size configuration.
.....
```

原因分析

经分析因为在写数据到Kafka时，Kafka客户端会先比较配置项“max.request.size ”值和本次写入数据大小，若写入数据大小超过此配置项“max.request.size ”的缺省值，则抛出上述异常。

解决办法

步骤1 在初始化Kafka生产者实例时，设置此配置项“max.request.size ”的值。

例如，参考本例，可以将此配置项设置为“5252880”：

```
// 协议类型:当前支持配置为SASL_PLAINTEXT或者PLAINTEXT
props.put(securityProtocol, kafkaProc.getValues(securityProtocol, "SASL_PLAINTEXT"));
// 服务名
props.put(saslKerberosServiceName, "kafka");
props.put("max.request.size", "5252880");
.....
```

----结束

15.13.31 Kakfa 消费者读取单条记录过长问题

问题背景与现象

和“Kafka生产者写入单条记录过长问题”相对应的，在写入数据后，用户开发一个应用，以消费者调用新接口（org.apache.kafka.clients.consumer.*）到Kafka上读取数据，但读取失败，报异常大致如下：

```
.....
1687 [KafkaConsumerExample] INFO  org.apache.kafka.clients.consumer.internals.AbstractCoordinator  - Successfully joined group DemoConsumer with generation 1
1688 [KafkaConsumerExample] INFO  org.apache.kafka.clients.consumer.internals.ConsumerCoordinator  - Setting newly assigned partitions [default-0, default-1, default-2] for group DemoConsumer
2053 [KafkaConsumerExample] ERROR com.XXXXXX.bigdata.kafka.example.NewConsumer  -
```

```
[KafkaConsumerExample], Error due to
org.apache.kafka.common.errors.RecordTooLargeException: There are some messages at [Partition=Offset]:
{default-0=177} whose size is larger than the fetch size 1048576 and hence cannot be ever returned.
Increase the fetch size on the client (using max.partition.fetch.bytes), or decrease the maximum message
size the broker will allow (using message.max.bytes).
2059 [KafkaConsumerExample] INFO  com.XXXXXX.bigdata.kafka.example.NewConsumer  -
[KafkaConsumerExample], Stopped
.......
```

原因分析

经分析因为在读取数据时Kafka客户端会比较待读取数据大小和配置项“max.partition.fetch.bytes”值，若超过此配置项值，则抛出上述异常。

解决办法

步骤1 在初始化建立Kafka消费者实例时，设置此配置项“max.partition.fetch.bytes”的值。

例如，参考本例，可以将此配置项设置为“5252880”：

```
......
// 安全协议类型
props.put(securityProtocol, kafkaProc.getValues(securityProtocol, "SASL_PLAINTEXT"));
// 服务名
props.put(saslKerberosServiceName, "kafka");

props.put("max.partition.fetch.bytes","5252880");
......
```

----结束

15.13.32 Kafka 集群节点内多磁盘数据量占用高处理办法

用户问题

Kafka流式集群节点内有多块磁盘的使用量很高。当达到100%时就会造成kafka不可用如何处理?

问题现象

客户创建的MRS Kafka流式集群节点内有多块磁盘，由于分区不合理及业务原因导致某几个磁盘的使用量很高。当达到100%时就会造成kafka不可用。

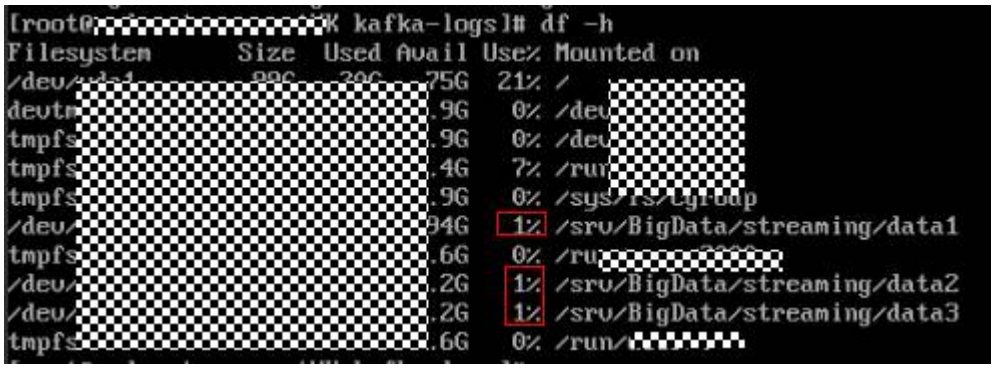
原因分析

需要提前干预处理磁盘数据，全局的log.retention.hours修改需要重启服务。为了不断服，可以将数据量大的单个topic老化时间根据需要改短。

处理步骤

步骤1 登录Kafka集群的l流式Core节点。

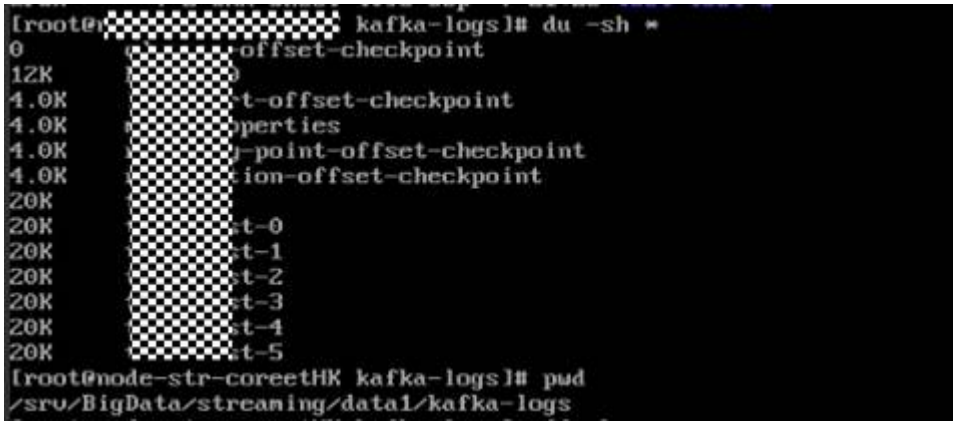
步骤2 执行**df –h**命令查看磁盘使用率。

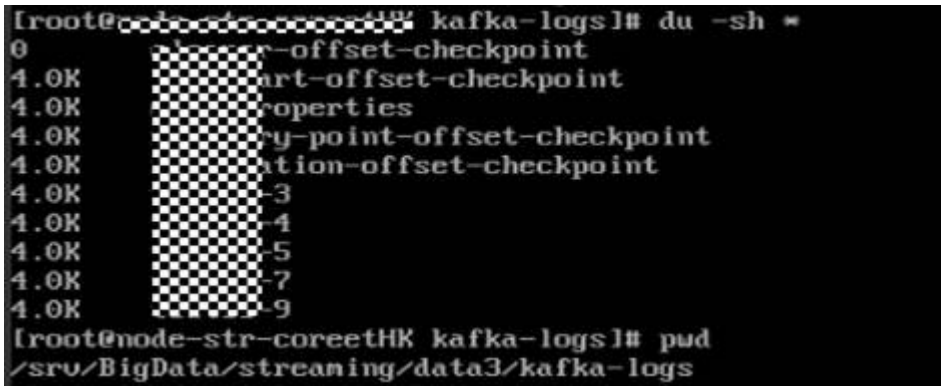


步骤3 通过kafka配置文件**opt/Bigdata/MRS_2.1.0/1_11_Broker/etc/server.properties**中的配置项**log.dirs** 获得数据存储目录。其中配置文件路径请根据时间环境的集群版本修改，当磁盘有多块时，该配置项有多个，逗号间隔。

步骤4 使用**cd**命令进入使用率较高的磁盘对应的**步骤3**中获取的数据存储目录下。

步骤5 使用**du -sh ***命令打印出当前topic的名称及大小。





步骤6 由于kafka的全局的数据保留时间默认是7天。部分topic由于业务写入量大，而这些topic的分区正好在上面使用量高的磁盘上，因此导致磁盘使用率较高。

- 可以通过修改全局数据的保留期为较短时间来释放磁盘空间，该方式需要重启Kafka服务才能生效，可能会影响业务运行。具体请参见**步骤7**。
- 可以单独将topic的数据保留期改为较短时间来释放磁盘空间，该方式无需重启Kafka服务即可生效。具体请参见**步骤8**。

步骤7 登录Manager页面，在Kafka的服务配置页面，切换为“全部配置”并搜索“log.retention.hours”配置项，该值默认为7天，请根据需要进行修改。

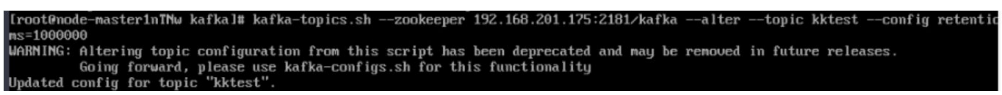
步骤8 可以单独将这些磁盘上的topic的数据老化时间修改为较短时间来解决该问题。

1. 查看topic数据过期时间。

   bin/kafka-topics.sh --describe --zookeeper *<ZooKeeper集群业务IP>*:2181/kafka --topic kktest

2. 设置topic数据过期时间，其中**--topic**表示具体topic名称，retention.ms=具体的数据过期时间，单位是毫秒。

   kafka-topics.sh --zookeeper *<ZooKeeper集群业务IP>*:2181/kafka --alter --topic kktest --config retention.ms=1000000

   设置数据过期时间之后会可能不会立刻执行，删除操作在参数**log.retention.check.interval.ms**所规定时间之后开始执行删，可以通过查看kafka的server.log检索是否有delete字段有判断删除操作是否生效，有delete字段

则表示已经 生效，也可以通过执行**df –h**命令查看磁盘的数据量占用情况判断设置是否生效。

```
log.retention.check.interval.ms = 300000
```

----结束

15.14 使用 Oozie

15.14.1 当并发提交大量 oozie 任务时，任务一直没有运行

用户问题

并发提交大量oozie任务的时候，任务一直没有运行。

问题现象

并发提交大量oozie任务的时候，任务一直没有运行。

原因分析

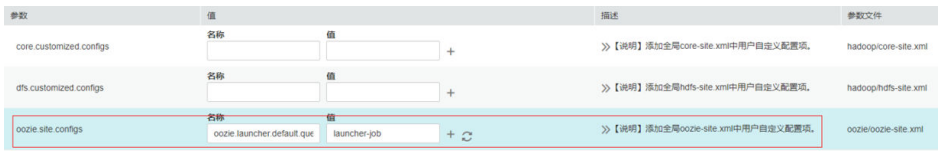
Oozie提交任务会先启动一个oozie-launcher，然后由oozie-launcher提交真正的作业运行。默认情况下launcher和真实作业会在同一个队列中。

当并发提交大量oozie任务的时候就有可能出现启动了一堆oozie-launcher，将队列的资源耗完，而没有更多资源启动真实作业，最终导致任务一直没有运行。

处理步骤

步骤1 参考“用户指南 > 管理现有集群 > 租户管理 > 添加租户”章节新建一个队列给oozie使用，也可以直接使用创建MRS集群时生成的launcher-job队列。

步骤2 在Manager页面选择“集群 > 服务 > Oozie > 配置”，搜索参数“oozie.site.configs”，在值列添加名称“oozie.launcher.default.queue”，值为“launcher-job”。



----结束

15.15 使用 Presto

15.15.1 配置 sql-standard-with-group 创建 schema 失败报 Access Denied

用户问题

配置sql-standard-with-group创建schema失败，报Access Denied的错误，如何处理?

问题现象

```
CREATE SCHEMA hive.sf2 WITH (location = 'obs://obs-zy1234/sf2');Query 20200224_031203_00002_g6gzy failed: Access Denied: Cannot create schema sf2
```

原因分析

presto创建schema需要hive的管理者权限。

处理步骤

MRS Manager界面操作：

- 方法一：
  a. 登录MRS Manager页面，选择“系统设置 > 用户管理”。
  b. 在对应用户所在行的“操作”列，单击“修改”。
  c. 单击“选择并绑定角色”，为用户添加System_administrator的权限。
  d. 单击“确定”完成修改。
- 方法二：
  a. 登录MRS Manager页面，选择“系统设置 > 角色管理”。
  b. 单击“添加角色”，并配置如下参数。
     - 角色名称：配置角色名称，例如hive_admin。
     - 权限：选择“Hive”，并勾选Hive Admin Privilege。
  c. 单击“确定”保存角色。
  d. 选择“系统设置 > 用户管理”。
  e. 在对应用户所在行的“操作”列，单击“修改”。
  f. 单击“选择并绑定角色”，为用户添加新创建的hive_admin的权限。
  g. 单击“确定”完成修改。

FusionInsight Manager界面操作：

- 方法一：
  a. 登录FusionInsight Manager页面，选择“系统 > 权限 > 用户”。
  b. 在对应用户所在行的“操作”列，单击“修改”。
  c. 单击角色后的“添加”，为用户添加System_administrator的权限。
  d. 单击“确定”完成修改。
- 方法二：

a. 登录FusionInsight Manager页面，选择“系统 > 权限 > 角色”。
b. 单击“添加角色”，并配置如下参数。
   - 角色名称：配置角色名称，例如hive_admin。
   - 配置资源权限：选择“Hive”，并勾选“Hive管理员权限”。
c. 单击“确定”保存角色。
d. 选择“系统 > 权限 > 用户”。
e. 在对应用户所在行的“操作”列，单击“修改”。
f. 单击角色后的“添加”，为用户添加新创建的hive_admin的权限。
g. 单击“确定”完成修改。

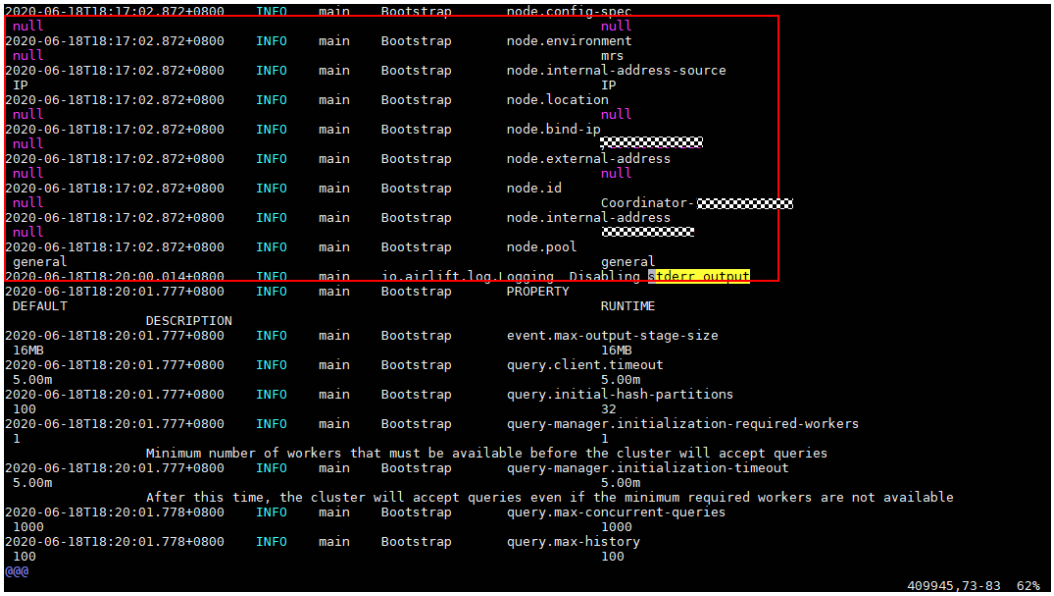
15.15.2 Presto 的 coordinator 无法正常启动

用户问题

Presto的coordinator未知原因被kill，或者presto的coordinator进程无法正常启动。

问题现象

Presto的coordinator无法正常启动，Manager页面上显示presto coordinator进程正常启动且状态正常，但查看后台日志coordinator进程未真正启动，只有如下日志：



presto的coordinator未真正启动即被Kill了，不再打印其他日志，查看presto的其他日志也未发现为何被kill。

原因分析

presto的健康检查脚本的端口检查逻辑中未做好端口的区分。

处理步骤

步骤1 使用工具分别登录集群的Master节点执行如下操作。

步骤2 执行如下命令编辑文件。

vim /opt/Bigdata/MRS_xxx/install/FusionInsight-Presto-*/ha/module/harm/plugin/script/pcd.sh

该文件中的第31行修改为“http_port_exists=$(netstat -apn | awk '{print $4, $6}' | grep :${HTTP_PORT} | grep LISTEN | wc -l)”。

```
25
26 check_status()
27 {
28         proc_exists=$(ps -ef | grep com.facebook.presto.server.PrestoServer | grep -v grep | wc -l)
29         param="-u $PRESTO_SERVER/v1/cluster"
30         if [[ ${proc_exists} == 1 ]]; then
31             http_port_exists=$(netstat -apn | awk '{print $4, $6}' | grep :${HTTP_PORT} | grep LISTEN | wc -l)
32
33             if [[ ${http_port_exists} == 1 ]] ;then
34                 log ${PCD_LOG_FILE} "INFO" "return [ normal ]"
35                 return 0
36             else
37                 log ${PCD_LOG_FILE} "ERROR" "HTTP PORT does not exist, return [ abnormal ]"
38                 return 2
39             fi
40         else
41             log ${PCD_LOG_FILE} "INFO" " coordinator process not exists, return [ abnormal ]"
42             return 2
43         fi
44 }
45
```

步骤3 保存如上修改，再在manager页面上选择“服务管理 > Presto > 实例”，重启Coordinator进程。

----结束

15.15.3 Presto 查询 Kudu 表报错

用户问题

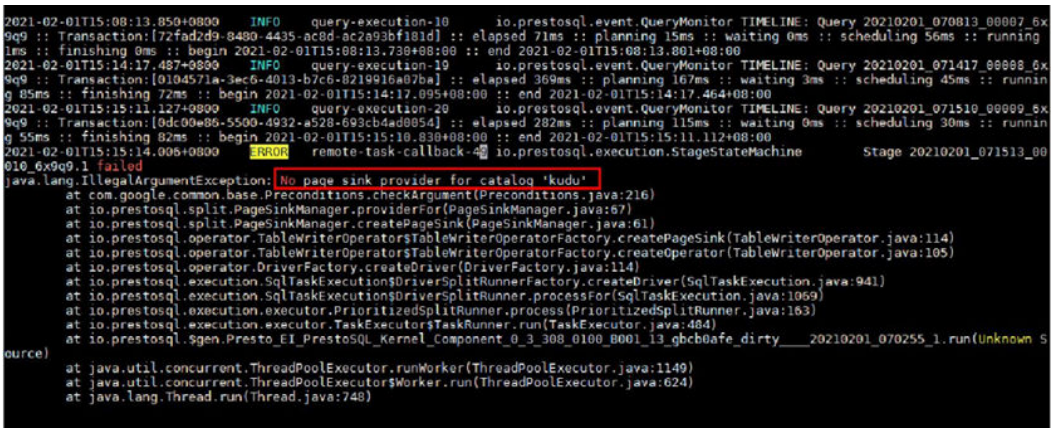
使用presto查询Kudu表报错。

问题现象

使用presto查询Kudu表，报表找不到的错误：

```
presto:default> show tables;
          Table

 impala::default.kudu_taobao
 impala::default.kudu_tt
 impala::default.kudutest
(3 rows)

Query 20210201_030636_00026_95mzd, FINISHED, 4 nodes
Splits: 53 total, 53 done (100.00%)
0:00 [3 rows, 125B] [18 rows/s, 766B/s]

presto:default> select count(*) from kudu.default.kudu_taobao;
Query 20210201_030653_00027_95mzd failed: line 1:22: Table kudu.default.kudu_taobao does not exist
select count(*) from kudu.default.kudu_taobao

presto:default> select count(*) from kudu_taobao;
Query 20210201_030939_00028_95mzd failed: line 1:22: Table kudu.default.kudu_taobao does not exist
select count(*) from kudu_taobao

presto:default>
```

后台报错：



原因分析

在实际的运行节点（worker实例所在节点）没有kudu相关配置。

处理步骤

步骤1 在集群presto所有的worker实例节点添加配置文件kudu.properties。

配置文件保存路径：/opt/Bigdata/MRS_xxx/1_x_Worker/etc/catalog/ (请根据集群实际版本修改路径)

配置文件内容：

```
connector.name=kudu
kudu.client.master-addresses=KuduMasterIP1:port,KuduMasterIP2:port,KuduMasterIP3:port
```

说明

- KuduMaster节点IP和端口请根据实际情况填写。
- 为配置文件添加和文件保存路径下其他文件一致的文件权限、属组。

步骤2 修改完成之后，请在集群详情页面选择“组件管理 > Kudu”，单击“更多 > 重启服务”。

----结束

15.15.4 Presto 查询 Hive 表无数据

用户问题

使用presto查询Hive表无数据。

问题现象

通过Tez引擎执行union相关语句写入的数据，Presto无法查询。

原因分析

由于Hive使用Tez引擎在执行union语句时，生成的输出文件会保存在HIVE_UNION_SUBDIR目录中，而Presto默认不读取子目录下的文件，所以没有读取到HIVE_UNION_SUBDIR目录下的数据。

处理步骤

步骤1 在集群详情页面选择“组件管理 > Presto > 服务配置”。

步骤2 切换“基础配置”为全部配置“。

步骤3 在左侧导航处选择“Presto > Hive”，在catalog/hive.properties文件中增加hive.recursive-directories参数，值为true。

步骤4 单击“保存配置”并勾选“重新启动受影响的服务或实例。”。

----结束

15.16 使用 Spark

15.16.1 Spark 应用下修改 split 值时报错

用户问题

在Spark应用下修改split值时报错。

问题现象

客户需要通过修改一个split最大值来实现多个mapper ，从而达到提速的目的, 但是目前执行**set $参数**命令修改Hive的配置时报错。

```
0: jdbc:hive2://192.168.1.18:21066/> set mapred.max.split.size=1000000;
Error: Error while processing statement: Cannot nodify mapred.max.split.size at runtime. It is not in list of params that are allowed to be modified at runtime( state=42000,code=1)
```

原因分析

- 在安全模式下配置白名单启停参数**hive.security.whitelist.switch**时，需要运行的参数必须在**hive.security.authorization.sqlstd.confwhitelist** 中配置。
- 默认白名单中没有包含**mapred.max.split.size**参数，所以运行的时候会提示不允许。

处理步骤

步骤1 进入Hive服务配置页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 > Hive > 服务配置”，单击“基础配置”下拉菜单，选择“全部配置”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 > Hive > 配置 > 全部配置”。

步骤2 搜索**hive.security.authorization.sqlstd.confwhitelist.append**，把**mapred.max.split.size**加进**hive.security.authorization.sqlstd.confwhitelist.append**中，详细信息可参考“组件操作指南 > > 使用Hive > 从零开始使用Hive”。

步骤3 修改完成后，保存配置，重启Hive组件。

步骤4 执行**set mapred.max.split.size=1000000;**，系统不再报错，则表示修改成功。

----结束

15.16.2 使用 Spark 时报错

用户问题

在使用spark时，集群运行失败。

问题现象

客户在使用spark组件时，集群运行失败。

```
[omm@node-master1-qxvMQ spark]$
[omm@node-master1-qxvMQ spark]$
[omm@node-master1-qxvMQ spark]$
[omm@node-master1-qxvMQ spark]$ ./bin/spark-submit --class cn.interf.Test  --master yarn-client /opt/client/Spark/spark1-1.0-SNAPSHOT.jar;
Error: Unrecognized option: --class cn.interf.Test  --master

Java HotSpot(TM) 64-Bit Server VM warning: Cannot open file <LOG_DIR>/gc.log due to No such file or directory

Usage: spark-submit [options] <app jar | python file> [app arguments]
Usage: spark-submit --kill [submission ID] --master [spark://...]
Usage: spark-submit --status [submission ID] --master [spark://...]
Usage: spark-submit run-example [options] example-class [example args]

Options:
  --master MASTER_URL         spark://host:port, mesos://host:port, yarn, or local.
  --deploy-mode DEPLOY_MODE   Whether to launch the driver program locally ("client") or
                              on one of the worker machines inside the cluster ("cluster")
                              (Default: client).
  --class CLASS_NAME          Your application's main class (for Java / Scala apps).
  --name NAME                 A name of your application.
  --jars JARS                 Comma-separated list of local jars to include on the driver
```

原因分析

- 执行命令时，引入非法字符
- 上传的jar包属主属组有问题

处理步骤

步骤1 检查用户命令**./bin/spark-submit --class cn.interf.Test --master yarn-client /opt/client/Spark/spark1-1.0-SNAPSHOT.jar;**，排查是否引入非法字符。

步骤2 如果是，修改非法字符，重新执行命令。

步骤3 重新执行命令后，发生其他错误，查看该jar包的属主属组，发现全为root。

步骤4 修改jar包的属主属组为 omm:wheel，重新执行成功。

----结束

15.16.3 引入 jar 包不正确，导致 Spark 任务无法运行

用户问题

执行Spark任务，任务无法运行。

问题现象

执行Spark任务，任务无法运行。

原因分析

执行Spark任务时，引入的jar包不正确，导致Spark任务运行失败。

处理步骤

步骤1 登录任意Master节点。

步骤2 执行**cd /opt/Bigdata/MRS_*/install/FusionInsight-Spark-*/spark/examples/jars**命令， 查看样例程序的jar包。

说明

jar包名最多为1023字符，不能包含;|&>,<'$特殊字符，且不可为空或全空格。

步骤3 检查OBS桶上的执行程序，执行程序可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。

说明

- OBS存储路径：以“obs://”开头。示例：obs://wordcount/program/hadoop-mapreduce-examples-2.7.x.jar
- HDFS存储路径：以“/user”开头。Spark Script需要以“.sql”结尾，MR和Spark需要以“.jar”结尾。sql、jar不区分大小写。

----结束

15.16.4 Spark 任务由于内存不够，作业卡住

用户问题

Spark提交作业内存不足导致任务长时间处于pending状态或者运行中内存溢出。

问题现象

使用Spark提交作业后，长期卡住不动。反复运行作业后报错，内容如下：

```
Exception in thread "main" org.apache.spark.SparkException: Job aborted due to stage failure:
Aborting TaskSet 3.0 because task 0 (partition 0) cannot run anywhere due to node and executor blacklist.
Blacklisting behavior can be configured via spark.blacklist.*.
```

原因分析

内存不足导致Spark提交的作业任务长时间处于pending状态。

处理步骤

步骤1 登录MRS Console页面，在现有集群中，选择集群名称，在“节点信息”页面，查看当前集群的节点规格。

步骤2 提高nodemanager进程所持有的集群资源。

MRS Manager界面操作：

1. 登录MRS Manager页面，选择“服务管理 > Yarn > 服务配置”。
2. 在“参数类别”中选择“全部配置”，然后在搜索框中搜索**yarn.nodemanager.resource.memory-mb**，查看该参数值。建议配置成节点物理内存总量的75%-90%。

FusionInsight Manager界面操作：

1. 登录FusionInsight Manager。选择“集群 > 服务 > Yarn”。
2. 单击“配置”，选择“全部配置”。然后在搜索框中搜索**yarn.nodemanager.resource.memory-mb**，查看该参数值。建议配置成节点物理内存总量的75%-90%。

步骤3 修改Spark的服务配置。

MRS Manager界面操作：

1. 登录MRS Manager页面，选择“服务管理”>“Spark” >“服务配置”。
2. 在“参数类别”中选择“全部配置”，然后在搜索框中搜索**spark.driver.memory**和**spark.executor.memory**

   根据作业的需要调大或者调小该值，具体以提交的Spark作业的复杂度和内存需要为参考（一般调大）。

FusionInsight Manager界面操作：

1. 登录FusionInsight Manager。选择“集群 > 服务 > Spark”。
2. 单击“配置”，选择“全部配置”。然后在搜索框中搜索**spark.driver.memory**和**spark.executor.memory**，根据作业的需要调大或者调小该值，具体以提交的Spark作业的复杂度和内存需要为参考（一般调大）。

说明

- 如果使用到SparkJDBC作业，搜索并修改**SPARK_EXECUTOR_MEMORY**和**SPARK_DRIVER_MEMORY**两个参数取值，具体以提交的Spark作业的复杂度和内存需要为参考（一般调大）。
- 如果对核数有要求，可以搜索并修改**spark.driver.cores**和**spark.executor.cores**的核数取值。

步骤4 Spark依赖内存做计算，如果以上还是不能满足任务的提交需要，建议扩容集群。

----结束

15.16.5 运行 Spark 报错

用户问题

运行Spark作业报找不到指定的类。

问题现象

运行Spark作业报找不到指定的类。报错内容如下：

```
Exception encountered | org.apache.spark.internal.Logging$class.logError(Logging.scala:91)
org.apache.hadoop.hbase.DoNotRetryIOException: java.lang.ClassNotFoundException:
org.apache.phoenix.filter.SingleCQKeyValueComparisonFilter
```

原因分析

用户配置的默认路径不正确。

处理步骤

步骤1 登录任意Master节点。

步骤2 修改Spark客户端目录下的配置文件。

执行**vim /opt/client/Spark/spark/conf/spark-defaults.conf**命令，打开spark-defaults.conf文件，设置“spark.executor.extraClassPath” 取值为“${PWD}/*”。

----结束

15.16.6 Driver 端提示 executor memory 超限

问题背景与现象

内存超限导致提交Spark任务失败。

原因分析

在Driver日志中直接打印申请的executor memory超过集群限制。

```
16/02/06 14:11:25 INFO Client: Verifying our application has not requested more than the maximum
memory capability of the cluster (6144 MB per container)
16/02/06 14:11:29 ERROR SparkContext: Error initializing SparkContext.
java.lang.IllegalArgumentException: Required executor memory (10240+1024 MB) is above the max
threshold (6144 MB) of this cluster!
```

Spark任务提交至Yarn上面，运行task的executor使用的资源受yarn的管理。从报错信息可看出，用户申请启动executor时，指定10G的内存，超出了Yarn设置的每个container的最大内存的限制，导致任务无法启动。

解决办法

修改Yarn的配置，提高对container的限制。如可通过调整“yarn.scheduler.maximum-allocation-mb”参数的大小，可控制启动的executor的资源，修改之后要重启Yarn服务。

配置修改方法：

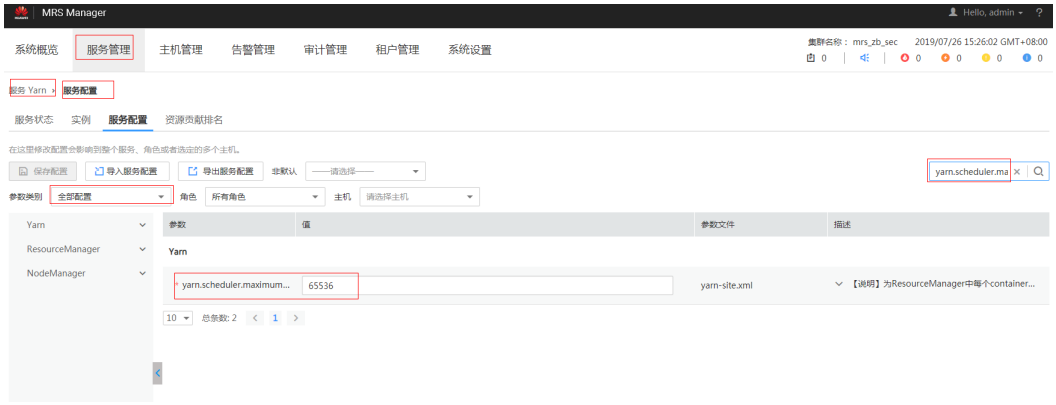
MRS Manager界面操作：

步骤1 登录MRS Manager页面。

步骤2 选择“服务管理 > Yarn > 服务配置”将“参数类别”修改为“全部配置”。

步骤3 在“搜索”栏输入“yarn.scheduler.maximum-allocation-mb”修改参数并保存重启服务。如下图所示：

图 15-55 修改 Yarn 服务参数



----结束

FusionInsight Manager界面操作：

步骤1 登录FusionInsight Manager页面。

步骤2 选择“集群 > 服务 > Yarn”，单击“配置”，选择“全部配置”。

步骤3 在“搜索”栏输入“yarn.scheduler.maximum-allocation-mb”修改参数并保存重启服务。

----结束

15.16.7 Yarn-cluster 模式下，Can't get the Kerberos realm 异常

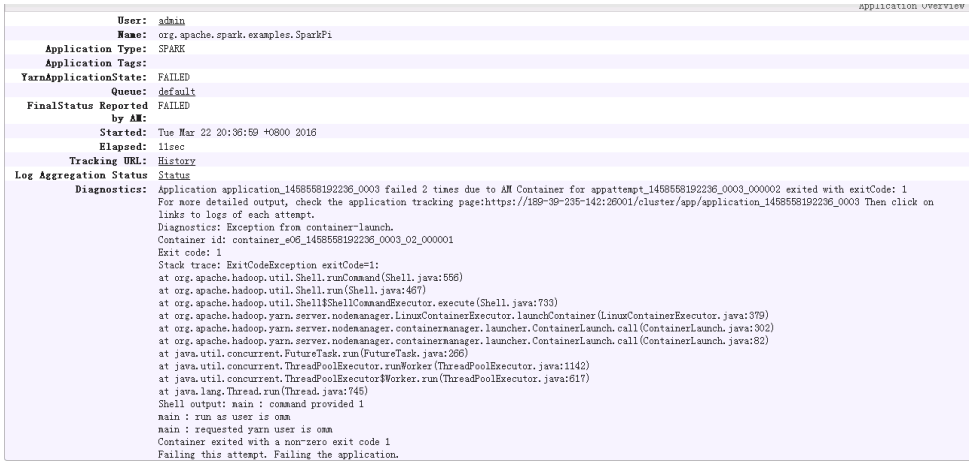
问题背景与现象

认证异常导致提交Spark任务失败。

原因分析

1. 在driver端打印异常找不到连接hdfs的token，报错如下：
   ```
   16/03/22 20:37:10 WARN Client: Exception encountered while connecting to the server :
   org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.token.SecretManager
   $InvalidToken): token (HDFS_DELEGATION_TOKEN token 192 for admin) can't be found in cache
   16/03/22 20:37:10 WARN Client: Failed to cleanup staging dir .sparkStaging/
   application_1458558192236_0003
   org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.token.SecretManager
   $InvalidToken): token (HDFS_DELEGATION_TOKEN token 192 for admin) can't be found in cache
   ```
2. 在Yarn原生页面显示am启动两次均失败，任务退出，如图15-56信息：

图 15-56 am 启动失败



3. 查看ApplicationMaster日志看到如下异常信息：

   ```
   Exception in thread "main" java.lang.ExceptionInInitializerError
   Caused by: org.apache.spark.SparkException: Unable to load YARN support
   Caused by: java.lang.IllegalArgumentException: Can't get Kerberos realm
   Caused by: java.lang.reflect.InvocationTargetException
   Caused by: KrbException: Cannot locate default realm
   Caused by: KrbException: Generic error (description in e-text) (60) - Unable to locate Kerberos realm
   org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1410)
   ... 86 more
   Caused by: javax.jdo.JDOFatalInternalException: Unexpected exception caught.
   NestedThrowables:java.lang.reflect.InvocationTargetException
   ... 110 more
   ```

4. 执行**./spark-submit --class yourclassname --master yarn-cluster /yourdependencyjars**任务以yarn-cluster模式提交任务，driver端会在集群中启用，由于加载的是客户端的spark.driver.extraJavaOptions，在集群节点上对应路径下找不到对应的kdc.conf文件，无法获取kerberos认证所需信息，导致am启动失败。

解决办法

在客户端提交任务时，在命令行中配置自定义的spark.driver.extraJavaOptions参数这样任务运行时就不会自动加载客户端路径下spark-defaults.conf中的spark.driver.extraJavaOptions；或者在启动spark任务时，通过--conf来指定driver的配置，如下（此处spark.driver.extraJavaOptions“=”号后面的引号部分不能缺少）。

./spark-submit -class yourclassname --master yarn-cluster --conf spark.driver.extraJavaOptions="

-Dlog4j.configuration=file:/opt/client/Spark/spark/conf/log4j.properties -Djetty.version=x.y.z -Dzookeeper.server.principal=zookeeper/hadoop.794bbab6_9505_44cc_8515_b4eddc84e6c1.com -Djava.security.krb5.conf=/opt/client/KrbClient/kerberos/var/krb5kdc/krb5.conf -Djava.security.auth.login.config=/opt/client/Spark/spark/conf/jaas.conf -Dorg.xerial.snappy.tempdir=/opt/client/Spark/tmp -Dcarbon.properties.filepath=/opt/client/Spark/spark/conf/carbon.properties" ../yourdependencyjars

15.16.8 JDK 版本不匹配启动 spark-sql，spark-shell 失败

问题背景与现象

JDK版本不匹配导致客户端启动spark-sql，spark-shell失败。

原因分析

1. 在Driver端打印异常如下：
   ```
   Exception Occurs: BadPadding 16/02/22 14:25:38 ERROR Schema: Failed initialising database. Unable to open a test connection to the given database. JDBC url = jdbc:postgresql://ip:port/sparkhivemeta, username = spark. Terminating connection pool (set lazyInit to true if you expect to start your database after your app).
   ```
2. Sparksql任务使用时，需要访问DBService以获取元数据信息，在客户端需要解密密文来访问，在使用过程中，用户没有按照流程操作，没有执行配置环境变量操作，且在其客户端环境变量中存在默认的jdk版本，导致在执行解密过程中调用的解密程序执行解密异常，会引起用户被锁。

解决办法

步骤1 使用**which java**命令查看默认的java命令是否是客户端的java。

步骤2 如果不是，请按正常的客户端执行流程。

source ${client_path}/bigdata_env

kinit *用户名*，然后输入用户名对应的密码，启动任务即可。

----结束

15.16.9 Yarn-client 模式提交 ApplicationMaster 尝试启动两次失败

问题背景与现象

Yarn-client模式提交任务AppMaster尝试启动两次失败。

原因分析

1. Driver端异常:
   ```
   16/05/11 18:10:56 INFO Client:
   client token: N/A
   diagnostics: Application application_1462441251516_0024 failed 2 times due to AM Container for appattempt_1462441251516_0024_000002 exited with  exitCode: 10
   For more detailed output, check the application tracking page:https://hdnode5:26001/cluster/app/application_1462441251516_0024 Then click on links to logs of each attempt.
   Diagnostics: Exception from container-launch.
   Container id: container_1462441251516_0024_02_000001
   ```
2. 在ApplicationMaster日志中，异常如下:
   ```
   2016-05-12 10:21:23,715 | ERROR | [main] | Failed to connect to driver at 192.168.30.57:23867, retrying ... | org.apache.spark.Logging$class.logError(Logging.scala:75)
   2016-05-12 10:21:24,817 | ERROR | [main] | Failed to connect to driver at 192.168.30.57:23867, retrying ... | org.apache.spark.Logging$class.logError(Logging.scala:75)
   2016-05-12 10:21:24,918 | ERROR | [main] | Uncaught exception:  | org.apache.spark.Logging$class.logError(Logging.scala:96)
   org.apache.spark.SparkException: Failed to connect to driver!
   at org.apache.spark.deploy.yarn.ApplicationMaster.waitForSparkDriver(ApplicationMaster.scala:426)
   at org.apache.spark.deploy.yarn.ApplicationMaster.runExecutorLauncher(ApplicationMaster.scala:292)
   ```

```
...
2016-05-12 10:21:24,925 | INFO  | [Thread-1] | Unregistering ApplicationMaster with FAILED (diag message: Uncaught exception: org.apache.spark.SparkException: Failed to connect to driver!) | org.apache.spark.Logging$class.logInfo(Logging.scala:59)
```

Spark-client模式任务Driver运行在客户端节点上(通常是集群外的某个节点)，启动时先在集群中启动AppMaster进程，进程启动后要向Driver进程注册信息，注册成功后，任务才能继续。从AppMaster日志中可以看出，无法连接至Driver，所以任务失败。

解决办法

步骤1 请测试一下Driver进程所在的IP是否可以**ping**通。

步骤2 启动一个sparkpi任务，在console端会有类似如下打印信息。

```
16/05/11 18:07:20 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.1.100:23662]
16/05/11 18:07:20 INFO Utils: Successfully started service 'sparkDriver' on port 23662.
```

步骤3 在该节点，也就是**步骤2**中示例的192.168.1.100上执行**netstat - anp | grep 23662**看下此端口是否打开，如下打印标明，相关端口是打开的。

```
tcp        0      0  ip:port    :::*                    LISTEN      107274/java
tcp        0      0  ip:port    ip:port                 ESTABLISHED 107274/java
```

步骤4 在AppMaster启动的节点执行**telnet 192.168.1.100 23662**看下是否可以联通该端口，请使用**root**用户和**omm**用户都执行一遍。 如果出现**Escape character is '^]'**类似打印则说明可以联通，如果出现**connection refused**则表示失败，无法连接到相关端口。

如果相关端口打开，但是从别的节点无法联通到该端口，则需要排查下相关网络配置。

说明

23662这个端口每次都是随机的，所以要根据自己启动任务打开的端口来测试。

----结束

15.16.10 提交 Spark 任务时，连接 ResourceManager 异常

问题背景与现象

连接ResourceManager异常，导致Spark任务提交失败。

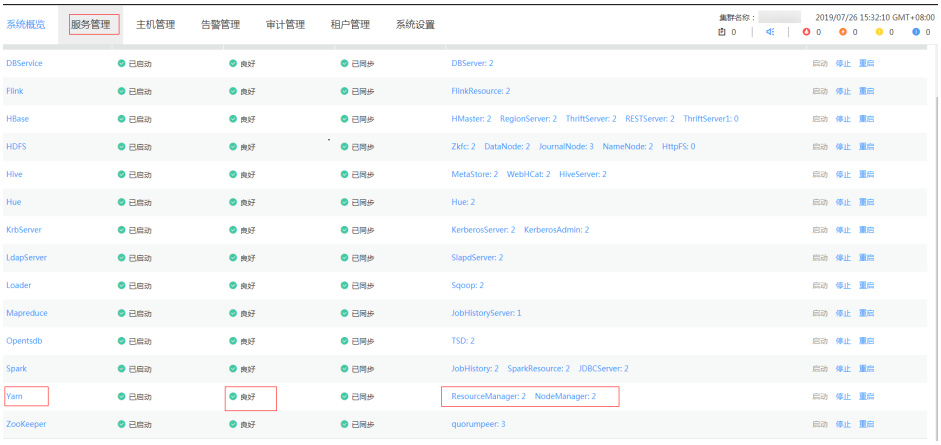
原因分析

1. 在driver端打印异常如下，打印连接两个ResourceManager主备节点的26004端口均被拒绝：

```
15/08/19 18:36:16 INFO RetryInvocationHandler: Exception while invoking getClusterMetrics of class ApplicationClientProtocolPBClientImpl over 33 after 1 fail over attempts. Trying to fail over after sleeping for 17448ms.
 java.net.ConnectException: Call From ip0 to ip1:26004 failed on connection exception: java.net.ConnectException: Connection refused.
INFO RetryInvocationHandler: Exception while invoking getClusterMetrics of class ApplicationClientProtocolPBClientImpl over 32 after 2 fail over attempts. Trying to fail over after sleeping for 16233ms.
 java.net.ConnectException: Call From ip0 to ip2:26004 failed on connection exception: java.net.ConnectException: Connection refused;
```

2. 在MRS Manager页面查看ResourceManager此时是否功能正常，如图15-57所示，如果Yarn状态故障或某个yarn服务的实例出现未知之类的异常说明此时集群的RM可能异常。

图 15-57 服务状态



3. 排查使用的客户端是否是集群最新的客户端。

   排查集群是否做过实例RM迁移相关操作（先卸载某个RM实例，然后在其他节点添加回来）。

4. 在MRS Manager页面单击“审计管理”，查看审计日志，是否有相关操作的记录。

   使用**ping**命令，查看IP是否可连通。

解决办法

- 如果RM出现异常，可参考Yarn相关章节查看解决方法。
- 如果客户端不是最新，请重新下载客户端。
- 若使用**ping**命令查看IP不通，需要协调网络管理相关人员协助排查网络。

15.16.11 DataArts Studio 调度 spark 作业失败

用户问题

DataArts Studio作业调度失败，显示读取/thriftserver/active_thriftserver路径下的数据失败。

问题现象

DataArts Studio作业调度失败，显示读取/thriftserver/active_thriftserver路径下的数据失败，

报错信息为：Can not get JDBC Connection, due to KeeperErrorCode = NoNode for /thriftserver/active_thriftserver。

原因分析

DataArts Studio提交spark作业时调用spark的JDBC方式，而Spark会启动一个名为thriftserver的进程以供客户端提供JDBC连接，JDBCServer在启动时会在zk的/

thriftserver目录下创建子目录active_thriftserver，并且注册相关连接信息。如果读不到该连接信息就会JDBC连接异常。

处理步骤

检查zookeeper下面是否有目标目录和注册的信息

步骤1 以root用户登录任意一个Master节点并初始化环境变量。

source /opt/client/bigdata_env

步骤2 执行**zkCli.sh -server 'ZookeeperIp:2181'**命令登录zk。

步骤3 执行**ls /thriftserver**查看是否有active_thriftserver目录。

- 如果有active_thriftserver目录，执行**get /thriftserver/active_thriftserver**查看该目录下是否有注册的配置信息。
  - 如果有注册的配置信息，联系支持人员处理。
  - 如果没有注册的配置信息，执行**步骤4**
- 如果没有active_thriftserver目录，执行**步骤4**。

步骤4 登录Manager界面，查看Spark的JDBCServer实例的主备状态是否未知。

- 是，执行**步骤5**。
- 否，联系运维人员处理。

步骤5 重启两个JDBCServer实例，查看主备实例状态恢复正常且zk下面有了目标目录和数据，作业即可恢复正常。若实例状态没有恢复请联系支持人员处理。

----结束

15.16.12 Spark 作业 api 提交状态为 error

用户问题

使用API提交spark作业后，作业状态显示为error。

问题类型

作业管理类。

问题现象

修改/opt/client/Spark/spark/conf/log4j.properties中的日志级别，使用API V1.1接口作业提交后，状态显示为error。

原因分析

executor会监控作业日志回显，确定作业执行结果，改为error后，检测不到输出结果，因此过期后判断作业状态为异常。

处理步骤

将/opt/client/Spark/spark/conf/log4j.properties中的日志级别修改为**info**。

建议与总结

建议客户使用V2接口提交作业接口。

15.16.13 集群反复出现 43006 告警

用户问题

集群反复出现“ALM-43006 JobHistory进程堆内存使用超出阈值”告警，且按照告警参考设置无效。

问题现象

集群出现告警“ALM-43006 JobHistory进程堆内存使用超出阈值”并且按照指导设置以后，运行一段时间又会出现同样的告警。

原因分析

可能存在JobHistory内存泄露问题，需要安装相应的补丁修复。

处理步骤

- 适当调大JobHistory进程堆内存。
- 如果已经调大堆内存，可以通过重启JobHistory实例规避。

15.16.14 在 spark-beeline 中创建/删除表失败

用户问题

客户在spark-beeline频繁创建和删除大量用户的场景下，个别用户偶现创建/删除表失败。

问题现象

创建表过程：

```
CREATE TABLE wlg_test001 (start_time STRING,value INT);
```

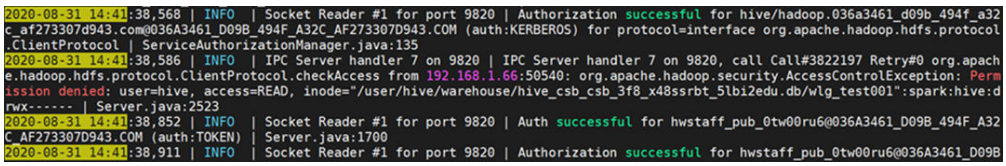
报错:

```
Error: org.apache.spark.sql.AnalysisException:
org.apache.hadoop.hive.ql.metadata.HiveException: MetaException(message:Failed to grant permission on
HDFSjava.lang.reflect.UndeclaredThrowableException); (state=,code=0)
```

原因分析

1. 查看metastore日志

2. 查看hdfs日志



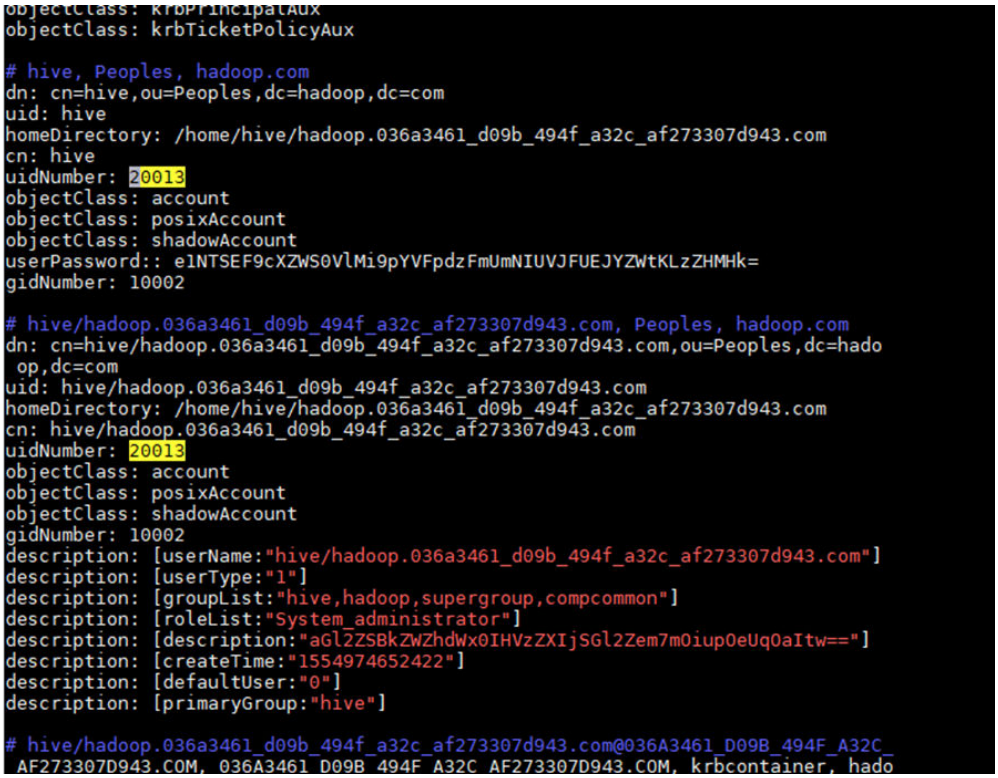
3. 权限对比（test001为异常用户创建表，test002为正常用户创建表）

4. drop表时报类似下面的错

```
0: jdbc:hive2://192.168.1.42:10000/> drop table
dataplan_modela_csbch2;
Error: Error while compiling statement: FAILED:
SemanticException Unable to fetch table dataplan_modela_csbch2.
java.security.AccessControlException: Permission denied: user=CSB_csb_3f8_x48ssrbt,
access=READ,
inode="/user/hive/warehouse/hive_csb_csb_3f8_x48ssrbt_5lbi2edu.db/
dataplan_modela_csbch2":spark:hive:drwx------
```

5. 根因分析。

   创建集群时创建的默认用户使用了相同的uid，造成用户错乱。在大量创建用户的场景下，触发了该问题，导致在创建表时偶现hive用户没有权限。



处理步骤

重启集群sssd进程。

以root用户执行**service sssd restart**命令重启sssd服务，执行**ps -ef | grep sssd**命令，查看sssd进程是否正常。

正常状态为：存在/usr/sbin/sssd进程和三个子进程/usr/libexec/sssd/sssd_be、/usr/libexec/sssd/sssd_nss、/usr/libexec/sssd/sssd_pam。

15.16.15 集群外节点提交 Spark 作业到 Yarn 报错连不上 Driver

用户问题

在集群外节点使用client模式提交Spark任务到Yarn上，任务失败，报错为连不上Driver。

问题现象

集群外节点和集群各个节点网络已经互通，在集群外节点使用client模式提交Spark任务到Yarn上，任务失败，报错为连不上Driver。

原因分析

使用client模式提交Spark任务的时候，Spark的driver进程是在客户端这边，而后面的executor都需要和Driver进行交互来运行作业。

如果NodeManager连不上客户端所在的节点，就会报错：

处理步骤

在客户端的Spark配置中指定Driver的IP地址：

“<客户端安装位置>/Spark/spark/conf/spark-defaults.conf”中添加参数“spark.driver.host=driverIP”，重新运行Spark任务即可。

建议与总结

建议客户通过cluster模式提交作业。

15.16.16 运行 Spark 任务发现大量 shuffle 结果丢失

用户问题

Spark任务运行失败，查看任务日志发现大量打印shuffle文件丢失。

问题现象

Spark任务运行失败，查看任务日志发现大量打印shuffle文件丢失。

原因分析

spark运行的时候会将临时产生的shuffle文件放在executor的临时目录中，方便后面获取。

而当某个executor异常退出时，NodeManager会把这个executor所在的container临时目录删除，随后其他executor再来申请这个executor的shuffle结果就会报文件找不到。

因此，遇到这样的问题需要确认是否executor异常退出，可以根据spark任务页面的executors便签页查看是否有dead状态的executor，查看各个dead状态的executor日志，确认异常退出的原因（其中可能有部分executor退出原因就是因为shuffle文件找不到，需要找到最早异常退出的executor）。

常见的异常退出：

- executor发生OOM
- executor运行时出现多个task任务失败
- executor所在节点被清理

处理步骤

根据executor异常退出的实际原因调整或者修改任务参数或代码，重新运行Spark任务即可。

15.16.17 JDBCServer 长时间运行导致磁盘空间不足

用户问题

连接Spark的JDBCServer服务提交spark-sql任务到yarn集群上，在运行一段时间以后会出现Core节点的数据盘被占满的情况。

问题现象

客户连接Spark的JDBCServer服务提交spark-sql任务到yarn集群上，在运行一段时间以后会出现Core节点的数据盘被占满的情况。

后台查看磁盘使用情况，主要是JDBCServer服务的APP临时文件（shuffle生成的文件）太多，并且没有进行清理占用了大量内存。

原因分析

查询Core节点有大量文件的目录，发现大部分都是类似“blockmgr-033707b6-fbbb-45b4-8e3a-128c9bcfa4bf”的目录，里面存放了计算过程中产生的shuffle临时文件。

因为JDBCServer启动了Spark的动态资源分配功能，已经将shuffle托管给NodeManager，NodeManager只会按照APP的运行周期来管理这些文件，并不会关注单个executor所在的container是否存在。因此，只有在APP结束的时候才会清理这些临时文件。任务运行时间较长时导致临时文件过多占用了大量磁盘空间。

处理步骤

启动一个定时任务来清理超过一定时间的shuffle文件，例如每个整点清理超过6个小时的文件：

步骤1 创建脚本“clean_appcache.sh”，若存在多个数据盘，请根据实际情况修改BASE_LOC中data1的值。

- 安全集群

```
#!/bin/bash
BASE_LOC=/srv/BigData/hadoop/data1/nm/localdir/usercache/spark/appcache/application_*/blockmgr*
find $BASE_LOC/ -mmin +360 -exec rmdir {} \;
find $BASE_LOC/ -mmin +360 -exec rm {} \;
```

- 普通集群

```
#!/bin/bash
BASE_LOC=/srv/BigData/hadoop/data1/nm/localdir/usercache/omm/appcache/application_*/blockmgr*
find $BASE_LOC/ -mmin +360 -exec rmdir {} \;
find $BASE_LOC/ -mmin +360 -exec rm {} \;
```

步骤2 修改脚本权限。

chmod 755 clean_appcache.sh

步骤3 增加一个定时任务来启动清理脚本，脚本路径请根据实际脚本存放位置修改。

查看定时任务：crontab -l

编辑定时任务：crontab -e

```
0 * * * * sh /root/clean_appcache.sh > /dev/null 2>&1
```

----结束

15.16.18 spark-shell 执行 sql 跨文件系统 load 数据到 hive 表失败

用户问题

使用spark-shell命令执行sql或者spark-submit提交的spark任务里面有sql的load命令，并且原数据和目标表存储位置不是同一套文件系统，上述两种方式MapReduce任务启动时会报错。

原因分析

当使用load导入数据到hive表的时候，属于需要跨文件系统的情况（例如原数据在hdfs上，而hive表数据存放在obs上），并且文件长度大于阈值（默认32M），则会触发使用distcp的MapReduce任务来执行数据迁移操作。这个MapReduce任务配置直接从spark任务配置里面提取，但是spark任务的net.topology.node.switch.mapping.impl配置项不是hadoop的默认值，需要使用spark的jar包，因此MapReduce会报类找不到。

处理步骤

方案一：

如果文件较小，则可以将默认长度设置得大于文件最大长度，例如最大的文件是95M，则设置：

```
hive.exec.copyfile.maxsize=104857600
```

方案二：

如果确实文件较大，需要使用distcp任务来提高数据迁移效率，则可以在spark任务启动的时候增加设置参数：

```
--conf spark.hadoop.net.topology.node.switch.mapping.impl=org.apache.hadoop.net.ScriptBasedMapping
```

15.16.19 Spark 任务提交失败

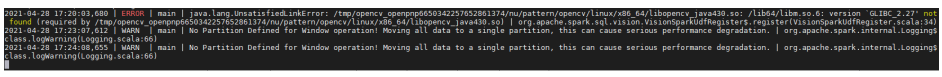
问题现象

- Spark提交任务直接提示无法提交任务。
- Spark提示无法获取到yarn的相关jar包。
- 提示多次提交一个文件。

原因分析

- 问题1：

  最常见的无法提交任务原因是认证失败，

  还有可能是参数设置不正确。

- 问题2：

  集群默认会把分析节点的hadoop相关jar包添加到任务的classpath中，如果提示yarn的包找不到，一般都是因为hadoop的相关配置没有设置。

- 问题3：

  常见的场景是使用--files上传了user.keytab，然后使用--keytab又指定了同一个文件，导致一个文件多次被上传。

  ```
  2021-04-29 10:08:56,973 | WARN  | main | Stopping a MetricsSystem that is not running | org.apache.spark.metrics.MetricsSystem.logWarning(Logging.scala:66)
  Exception in thread "main" java.lang.IllegalArgumentException: Attempt to add (file:///opt/user.keytab) multiple times to the distributed cache.
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10$$anonfun$apply$6.apply(Client.scala:646)
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10$$anonfun$apply$6.apply(Client.scala:637)
          at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
          at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:48)
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10.apply(Client.scala:637)
          at org.apache.spark.deploy.yarn.Client$$anonfun$prepareLocalResources$10.apply(Client.scala:636)
          at scala.collection.immutable.List.foreach(List.scala:392)
          at org.apache.spark.deploy.yarn.Client.prepareLocalResources(Client.scala:636)
          at org.apache.spark.deploy.yarn.Client.createContainerLaunchContext(Client.scala:913)
          at org.apache.spark.deploy.yarn.Client.submitApplication(Client.scala:205)
          at org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend.start(YarnClientSchedulerBackend.scala:57)
          at org.apache.spark.scheduler.TaskSchedulerImpl.start(TaskSchedulerImpl.scala:188)
          at org.apache.spark.SparkContext.<init>(SparkContext.scala:524)
          at org.apache.spark.SparkContext$.getOrCreate(SparkContext.scala:2695)
          at org.apache.spark.sql.SparkSession$Builder$$anonfun$7.apply(SparkSession.scala:956)
  ```

处理步骤

- 问题1：

  重新kinit一个用户并修改相应的配置参数。

- 问题2：

  查看hadoop相关的配置项是否正确，查看spark的conf目录下的core-site.xml，hdfs-site.xml，yarn-site.xml，mapred-site.xml等配置文件是否存在问题。

- 问题3：

  重新复制一个user.keytab，例如：

  cp user.keytab user2.keytab

  spark-submit --master yarn --files user.keytab --keytab user2.keytab .....

15.16.20 Spark 任务运行失败

问题现象

- 报错显示executor出现OOM
- 失败的task信息显示失败原因是lost task xxx

原因分析

- 问题1：一般出现executor OOM，都是因为数据量过大，也有可能是因为同一个executor上面同时运行的task太多。
- 问题2：有些task运行失败会报上述错误。当看到这个报错的时候，需要确认的是丢失的这个task在哪个节点上面运行，一般的情况是这个丢失的task异常退出导致的。

处理步骤

- 问题1：
  - 对于数据量过大，需要调整executor的内存大小的，使用--executor-memory指定内存大小；
  - 对于同时运行的task太多，主要看--executor-cores设置的vcore数量。
- 问题2：需要在相应的task的日志里面查找异常原因。如果有OOM的情况，请参照问题1。

15.16.21 JDBCServer 连接失败

问题现象

- 提示ha-cluster不识别（unknowHost或者必须加上端口）
- 提示连接JDBCServer失败

原因分析

- 问题1：使用**spark-beeline**命令连接JDBCServer，因为MRS_3.0以前的JDBCServer是ha模式，因此需要使用特定的url和MRS spark的自带的jar包来连接JDBCServer。
- 问题2：确认JDBCServer服务是否正常，查看对应的端口是否正常监听。

处理步骤

- 问题1：需要使用特定的url和MRS Spark的自带的jar包来连接JDBCServer。
- 问题2：确认JDBCServer服务是否正常，查看对应的端口是否正常监听。

15.16.22 查看 Spark 任务日志失败

问题现象

- 任务运行中查看日志失败
- 任务运行完成，但是查看不到日志

原因分析

- 问题1：可能原因是MapReduce服务异常
- 问题2：可能原因如下：
  - Spark的JobHistory服务异常。
  - 日志太大，NodeManager在做日志汇聚的时候出现超时。
  - HDFS存放日志目录权限异常（默认/tmp/logs/用户名/logs）。
  - 日志已被清理（spark的JobHistory默认存放7天的eventLog，配置项为spark.history.fs.cleaner.maxAge；MapReduce默认存放15天的任务日志，配置项为mapreduce.jobhistory.max-age-ms）。
  - 如果yarn页面上也找不到，可能是被yarn清理了（默认存放10000个历史任务，配置项为yarn.resourcemanager.max-completed-applications）。

处理步骤

- 问题1：确认MapReduce服务是否正常，如果异常，尝试重启服务。如果还是不能恢复，需要查看后台JobhistoryServer日志。
- 问题2：依次排查可能的情况：
  a. 查看Spark的JobHistory是否运行正常；
  b. 通过查看yarn的app详情页面，确认日志文件是否过大，如果日志汇聚失败，页面的“Log Aggregation Status:”应该显示为失败或者超时；
  c. 查看对应目录权限是否异常；

d. 查看目录下是否有对应的appid文件（spark的eventlog存放目录：MRS 3.x及以后版本的目录是hdfs://hacluster/spark2xJobHistory2x，MRS 3.x以前版本的目录是hdfs://hacluster/sparkJobHistory，任务运行日志存放目录是hdfs://hacluster/tmp/logs/用户名/logs）；

e. 查看appid和当前作业的id是否超过历史记录最大值。

15.16.23 Spark 连接其他服务认证问题

问题现象

- Spark连接HBase，报认证失败或者连接不到hbase表。
- Spark连接HBase报找不到jar包。

原因分析

- 问题1：HBase没有获取到当前任务的认证信息，导致连接HBase的时候认证失败，无法读取到相应数据
- 问题2：Spark默认没有加载HBase相关的jar包，需要使用--jars添加到任务中

处理步骤

- 问题1：可以尝试开启hbase认证开关：spark.yarn.security.credentials.hbase.enabled=true。但不建议直接用HBase客户端的hbase-site.xml替换Spark客户端下的hbase-site.xml，两者并不是完全相同。
- 问题2：需要将HBase相关的包使用--jars上传。

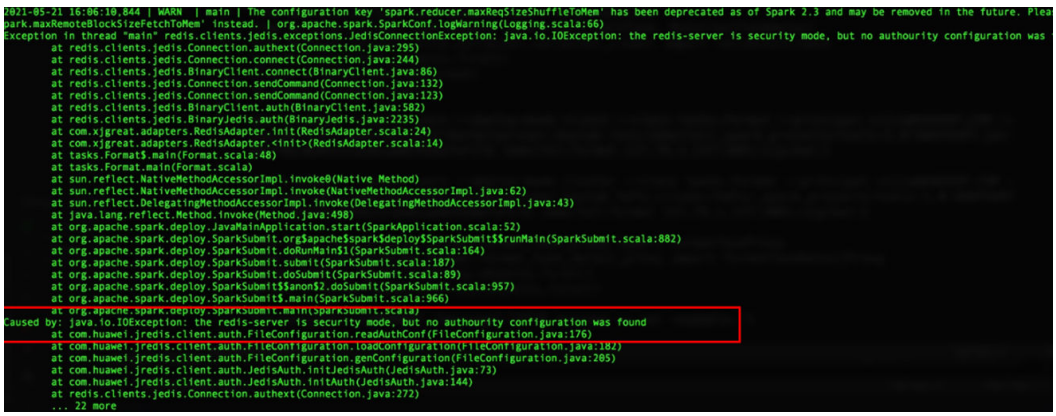
15.16.24 spark 连接 redis 报错

用户问题

使用MRS 3.x版本安全集群的spark组件访问redis报错。

问题现象

使用MRS_3.0版本安全集群的spark组件访问redis，会出现如下错误：



原因分析

Spark的jars目录下有一个MRS集群自带的jredisclient-xxx.jar包，客户使用spark任务连接redis的时候会因为加载了这个包从而出现该错误，需要手动去除redisclient包即可。

处理步骤

步骤1 清理Spark客户端下的jar包。

cd $SPARK_HOME/jars

mv jredisclient-*.jar /tmp

步骤2 清理Spark服务端下的jar包。

分别登录SparkResource2x所在的节点(一般有两个)。

mkdir /tmp/SparkResource2x

cd /opt/Bigdata/FusionInsight_Current/1_*_SparkResource2x/install/spark/jars/

mv jredisclient-*.jar /tmp/SparkResource2x

步骤3 清理HDFS上面的jredisclient文件。

1. 查看$SPARK_HOME/conf/spark-defaults.conf里面的配置项spark.yarn.archive，获取spark-archive-2x.zip包的地址。

   cat $SPARK_HOME/conf/spark-defaults.conf | grep "spark.yarn.archive"

2. 下载spark-archive-2x.zip包（本指导以MRS 3.0.5版本为例，具体命令请根据实际集群版本修改）。

   cd /opt

   mkdir sparkTmp

   cd sparkTmp

   hdfs dfs -get hdfs://hacluster/user/spark2x/jars/8.0.2.1/spark-archive-2x.zip

3. 解压spark-archive-2x.zip文件，并删除原文件。

   unzip spark-archive-2x.zip

   rm -f spark-archive-2x.zip

4. 移除jredisclient包。

   rm -f jredisclient-*.jar

5. 重新压缩spark-archive-2x.zip包。

   zip spark-archive-2x.zip ./*

6. 备份原有压缩包，上传新的压缩包。

   hdfs dfs -mv hdfs://hacluster/user/spark2x/jars/8.0.2.1/spark-archive-2x.zip /tmp

   hdfs dfs -put spark-archive-2x.zip hdfs://hacluster/user/spark2x/jars/8.0.2.1/spark-archive-2x.zip

7. 新的spark-archive-2x.zip中已删除jredisclient文件，需要重启JDBCServer服务，防止JDBCServer服务异常。

8. 清除临时文件。

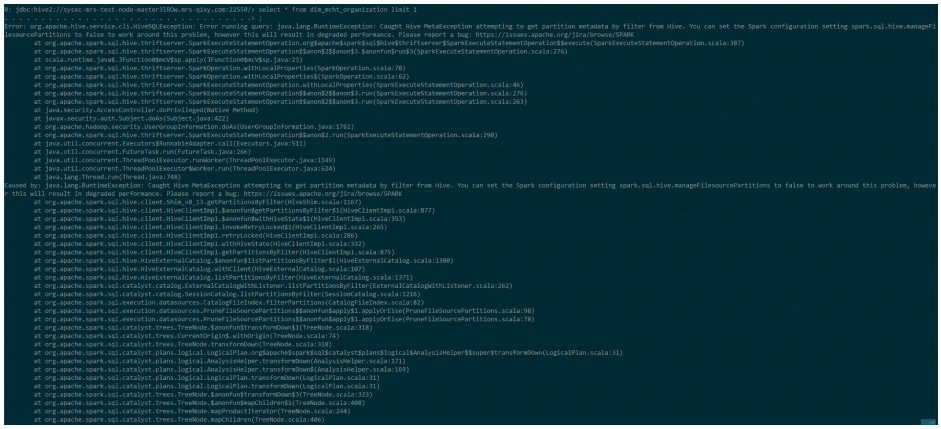
   rm -rf /opt/sparkTmp

----结束

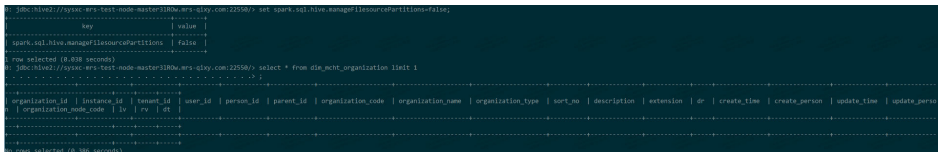
15.16.25 spark-beeline 查询 Hive 视图报错

用户问题

MRS 3.1.2版本，spark-beeline查询Hive视图报错，报错如下：



根据提示set spark.sql.hive.manageFilesourcePartitions=false后，查不到数据（但是实际在hive查询中是有数据的）。



原因分析

转换格式失败，spark-sql使用其内置的Metastore，而不是Hive中使用的Metastore，所以读取元数据时失败，spark-sql内部读取的是partque格式的数据，而Hive中默认是orc的。

处理步骤

步骤1 登录Spark客户端节点，执行如下命令，进入spark-sql：

cd *{客户端安装目录}*

source bigdata_env

source Spark2x/component_env

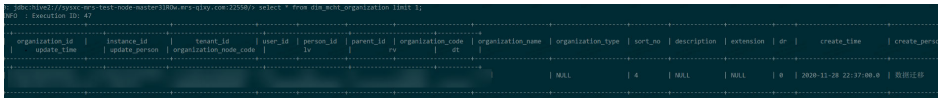
kinit *组件业务用户*（普通模式无需执行kinit）

spark-sql

步骤2 执行如下命令设置spark.sql.hive.convertMetastoreOrc=false。

set spark.sql.hive.convertMetastoreOrc=false;

步骤3 重新查询Hive视图，显示正常。



----结束

15.17 使用 Sqoop

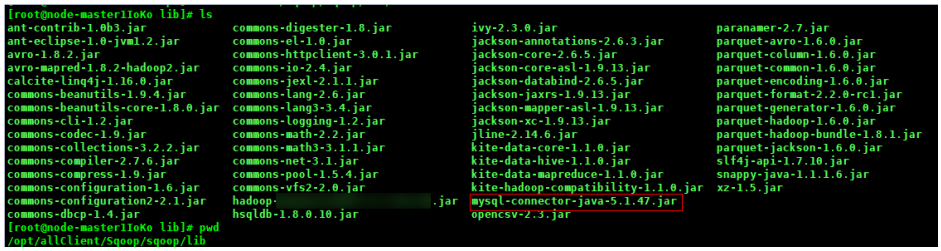
15.17.1 Sqoop 如何连接 mysql

用户问题

Sqoop如何连接mysql。

处理步骤

步骤1 在集群上安装客户端，查看客户端sqoop/lib下是否有mysql驱动包。



步骤2 在客户端目录下加载环境变量。

source bigdata_env

步骤3 Kerberos认证。

如果集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用跳过此步骤。

kinit MRS集群用户

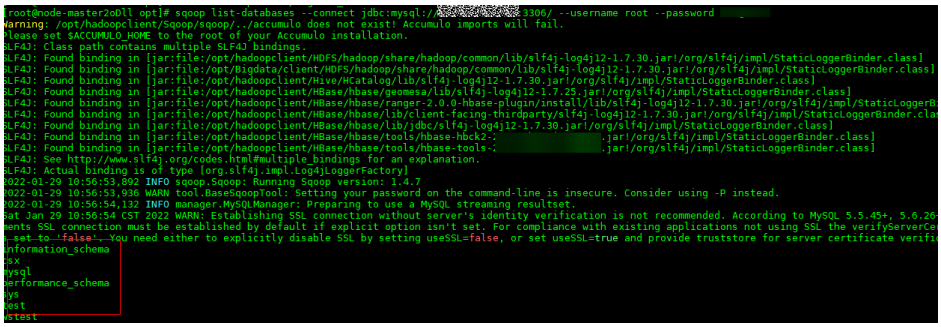
例如：

kinit admin

步骤4 连接数据库。

sqoop list-databases --connect jdbc:mysql://IP:3306/ --username 用户名 --password 密码

如下：



上图所示：则代表sqoop连接mysql成功。

----结束

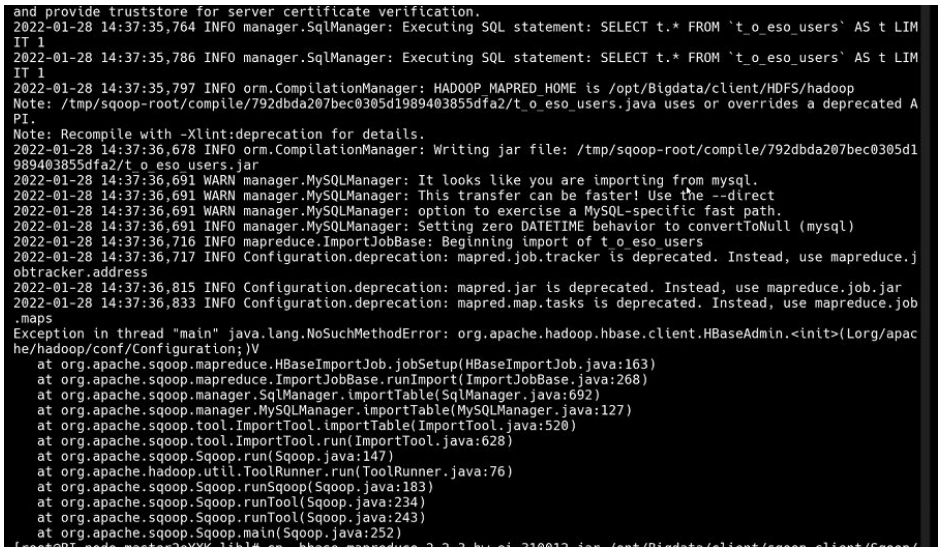
15.17.2 Sqoop 读取 MySQL 中数据到 HBase 报 HBaseAdmin.<init>方法找不到异常

问题

使用MRS的Sqoop客户端（1.4.7版本），从MySQL数据库中指定表抽取数据，存放到HBase（2.2.3版本）指定的表中，报出异常：

```
Trying to load data into HBASE through Sqoop getting below error.
Exception in thread "main" java.lang.NoSuchMethodError:
org.apache.hadoop.hbase.client.HBaseAdmin.<init>(Lorg/apache/hadoop/conf/Configuration;)V
```

完整异常信息如图所示：



执行Sqoop抽取数据命令样例：

```
sqoop import \
--connect jdbc:mysql://mysql服务器地址:端口号/database1 \
--username admin \
--password xxx \
--table table1 \
--hbase-table table2 \
--column-family info \
--hbase-row-key id \
--hbase-create-table --m 1
```

处理步骤

Sqoop客户端安装完成之后，没有直接引入HBase相关的依赖jar包，需要通过手动导入指定低版本的HBase相关依赖jar包。解决方法步骤如下：

步骤1 确认Sqoop客户端和HBase客户端是否在同一个路径下。

- 是，执行**步骤2**。
- 否，删除原有的Sqoop和HBase客户端文件，从FusionInsight Manager上下载完整的客户端安装在同一路径下。执行**步骤2**。

步骤2 以**root**用户登录Sqoop客户端安装节点。

步骤3 下载HBase 1.6.0版本的jar包上传到Sqoop客户端的“lib”目录下。

步骤4 上传包之后，修改包的权限，可以设置为755，具体执行命令为：

chmod 755 *包名称*

步骤5 在客户端目录下执行以下命令刷新Sqoop客户端：

source bigdata_env

重新执行sqoop命令

----结束

15.17.3 HUE 界面的 Sqoop 任务 HBase 到 HDFS 报错

本章节仅适用于MRS 1.9.2版本集群。

用户问题

利用HUE的sqoop操作把HBase中的数据导入HDFS中报错。

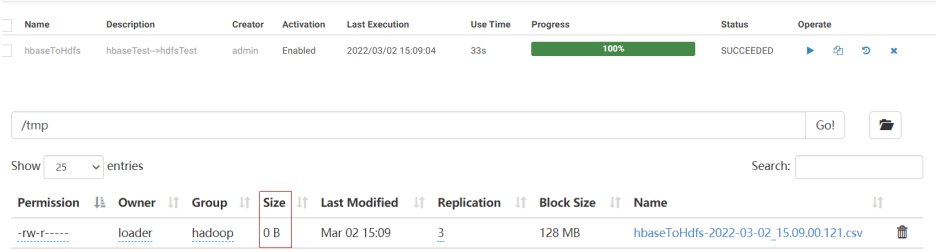
Caused by: java.lang.ClassNotFoundException: org.apache.htrace.Trace

```
2022-03-02 15:09:00,264 [main] ERROR org.apache.sqoop.connector.hbase.HBaseExtractor  - An exceptional condition has occurred.
org.apache.sqoop.common.SqoopException: HBASE_CONNECTOR_0011:Failed to open table.
        at org.apache.sqoop.connector.hbase.HBaseExtractor.openDB(HBaseExtractor.java:239)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.access$100(HBaseExtractor.java:34)
        at org.apache.sqoop.connector.hbase.HBaseExtractor$1.run(HBaseExtractor.java:86)
        at org.apache.sqoop.connector.hbase.HBaseExtractor$1.run(HBaseExtractor.java:76)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.extract(HBaseExtractor.java:114)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.extract(HBaseExtractor.java:34)
        at org.apache.sqoop.job.mr.SqoopMapper.runInternal(SqoopMapper.java:156)
        at org.apache.sqoop.job.mr.SqoopMapper.run(SqoopMapper.java:79)
        at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:787)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:188)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1840)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:182)
Caused by: java.lang.reflect.InvocationTargetException
```

```
Caused by: java.lang.reflect.InvocationTargetException
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.hadoop.hbase.client.ConnectionFactory.createConnection(ConnectionFactory.java:238)
        at org.apache.hadoop.hbase.client.ConnectionManager.createConnection(ConnectionManager.java:454)
        at org.apache.hadoop.hbase.client.ConnectionManager.createConnection(ConnectionManager.java:447)
        at org.apache.hadoop.hbase.client.ConnectionManager.getConnectionInternal(ConnectionManager.java:325)
        at org.apache.hadoop.hbase.client.HTable.<init>(HTable.java:184)
        at org.apache.hadoop.hbase.client.HTable.<init>(HTable.java:150)
        at org.apache.sqoop.connector.hbase.HBaseExtractor.openDB(HBaseExtractor.java:236)
        ... 14 more
Caused by: java.lang.NoClassDefFoundError: org/apache/htrace/Trace
        at org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper.exists(RecoverableZooKeeper.java:245)
        at org.apache.hadoop.hbase.zookeeper.ZKUtil.checkExists(ZKUtil.java:436)
        at org.apache.hadoop.hbase.zookeeper.ZKClusterId.readClusterIdZNode(ZKClusterId.java:65)
        at org.apache.hadoop.hbase.client.ZooKeeperRegistry.getClusterId(ZooKeeperRegistry.java:105)
        at org.apache.hadoop.hbase.client.ConnectionManager$HConnectionImplementation.retrieveClusterId(ConnectionManager.java:944)
        at org.apache.hadoop.hbase.client.ConnectionManager$HConnectionImplementation.<init>(ConnectionManager.java:720)
        ... 25 more
Caused by: java.lang.ClassNotFoundException: org.apache.htrace.Trace
        at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:419)
```

问题现象

Sqoop任务运行成功，但hdfs中的csv文件无内容。



原因分析

推测jar包冲突或者缺少jar包造成的。

处理步骤

步骤1 去sqoop的lib下grep。

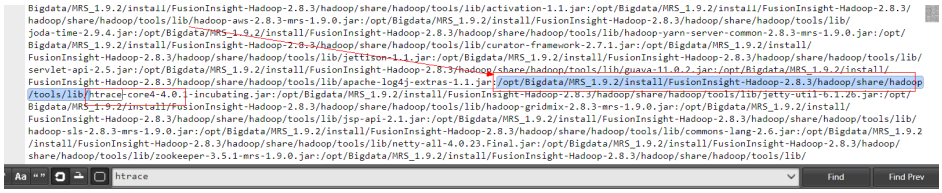
1. 进入sqoop的lib目录下，进行grep查找。

```
[root@node-master1PMPi lib]# pwd
/opt/Bigdata/MRS_1.9.2/install/FusionInsight-Sqoop-1.99.7/FusionInsight-Sqoop-1.99.7/server/lib
[root@node-master1PMPi lib]# grep org.apache.htrace.Trace *
Binary file htrace-core-3.1.0-incubating.jar matches
[root@node-master1PMPi lib]#
```

2. 进入yarn原生界面，查看运行的任务的报错具体信息。

```
Log Type: syslog
Log Upload Time: Thu Mar 03 15:19:29 +0800 2022
Log Length: 74284
2022-03-03 15:19:08,177 INFO [main] org.apache.hadoop.mapreduce.v2.app.MRAppMaster: Created MRAppMaster for application appattempt_1646291845172_
2022-03-03 15:19:08,357 INFO [main] org.apache.hadoop.mapreduce.v2.app.MRAppMaster:
/************************************************************
[system properties]
os.name: Linux
os.version: 3.10.0-327.62.59.83.h162.x86_64
java.home: /opt/Bigdata/jdk1.8.0_232/jre
java.runtime.version: 1.8.0_232-[redacted]_JDK_V100R001C00SPC173B001-b09
java.vendor: [redacted] Technologies Co., Ltd
java.version: 1.8.0_232
java.vm.name: OpenJDK 64-Bit Server VM
java.class.path: /srv/BigData/hadoop/data1/nm/localdir/usercache/loader/appcache/application_1646291845172_0001/container_e01_1646291845172_0001_
java.io.tmpdir: /srv/BigData/hadoop/data1/nm/localdir/usercache/loader/appcache/application_1646291845172_0001/container_e01_1646291845172_0001_0
user.dir: /srv/BigData/hadoop/data1/nm/localdir/usercache/loader/appcache/application_1646291845172_0001/container_e01_1646291845172_0001_01_0000
user.name: [redacted]
************************************************************/
```

3. 将java.class.path复制出来，搜索htrace-core。



4. 复制jar包到如下位置。

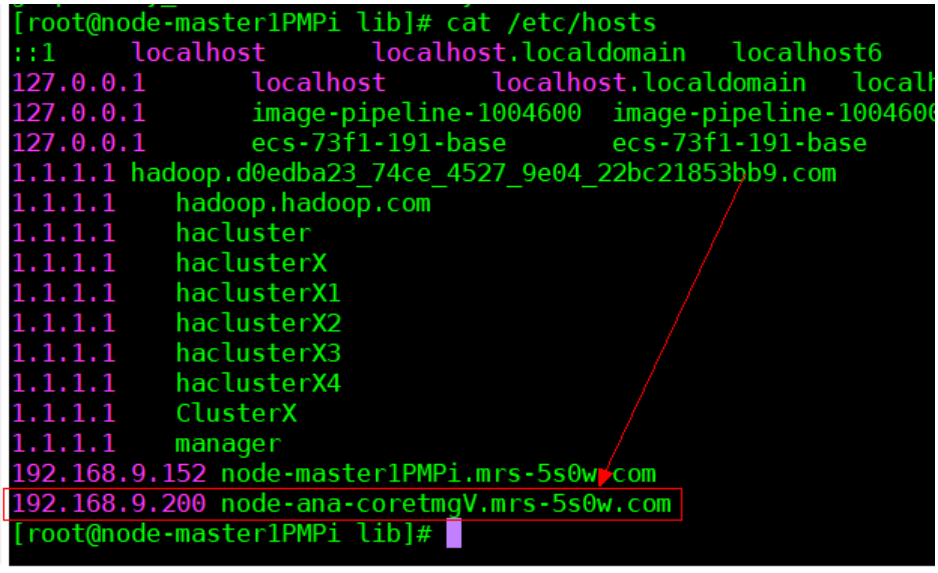
   cp /opt/Bigdata/MRS_1.9.2/install/FusionInsight-Sqoop-1.99.7/FusionInsight-Sqoop-1.99.7/server/lib/htrace-core-3.1.0-incubating.jar /opt/Bigdata/MRS_1.9.2/install/FusionInsight-Hadoop-2.8.3/hadoop/share/hadoop/common/lib/

5. 修改权限。

   chmod 777 *htrace-core-3.1.0-incubating.jar*（真实复制的jar包）

   chown omm:ficommon *htrace-core-3.1.0-incubating.jar*（真实复制的jar包）

6. 查看hosts文件，对其他所有节点进行同样的复制jar包操作。

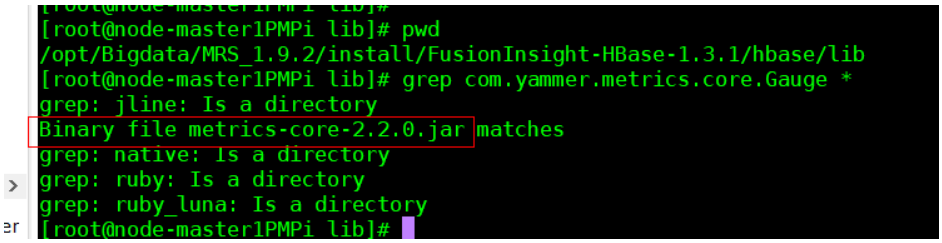


7. 重新运行sqoop任务，产生报错如下：

```
        at java.lang.Thread.run(Thread.java:745)
Caused by: com.google.protobuf.ServiceException: java.lang.NoClassDefFoundError: com/yammer/metrics/core/Gauge
        at org.apache.hadoop.hbase.ipc.AbstractRpcClient.callBlockingMethod(AbstractRpcClient.java:240)
        at org.apache.hadoop.hbase.ipc.AbstractRpcClient$BlockingRpcChannelImplementation.callBlockingMethod(AbstractRp
        at org.apache.hadoop.hbase.protobuf.generated.ClientProtos$ClientService$BlockingStub.scan(ClientProtos.java:35
        at org.apache.hadoop.hbase.client.ClientSmallReversedScanner$SmallReversedScannerCallable.call(ClientSmallRever
        ... 9 more
Caused by: java.lang.NoClassDefFoundError: com/yammer/metrics/core/Gauge
        at org.apache.hadoop.hbase.ipc.AbstractRpcClient.callBlockingMethod(AbstractRpcClient.java:225)
        ... 12 more
Caused by: java.lang.ClassNotFoundException: com.yammer.metrics.core.Gauge
        at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:419)
        at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:352)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:352)
        ... 13 more
2022-03-03 15:45:01,714 [main] INFO  org.apache.sqoop.job.mr.SqoopMapper  - Extractor has finished
2022-03-03 15:45:01,715 [main] INFO  org.apache.sqoop.job.mr.SqoopMapper  - Stopping progress service
2022-03-03 15:45:01,727 [main] INFO  org.apache.sqoop.job.mr.SqoopOutputFormatLoadExecutor  - SqoopOutputFormatLoadExec
2022-03-03 15:45:01,776 [OutputFormatLoader-consumer] INFO  org.apache.sqoop.job.mr.SqoopOutputFormatLoadExecutor  - Lo
2022-03-03 15:45:01,777 [main] INFO  org.apache.sqoop.job.mr.SqoopOutputFormatLoadExecutor  - SqoopOutputFormatLoadExec

Log Type: stdout
Log Upload Time: Thu Mar 03 15:45:15 +0800 2022
Log Length: 0

Log Type: syslog
```

步骤2 去hbase的lib下grep。

1. 进入hbase的lib目录下，进行grep查找。



2. 继续复制jar包过去。

   cp /opt/Bigdata/MRS_1.9.2/install/FusionInsight-HBase-1.3.1/hbase/lib/metrics-core-2.2.0.jar /opt/Bigdata/MRS_1.9.2/install/FusionInsight-Hadoop-2.8.3/hadoop/share/hadoop/common/lib/

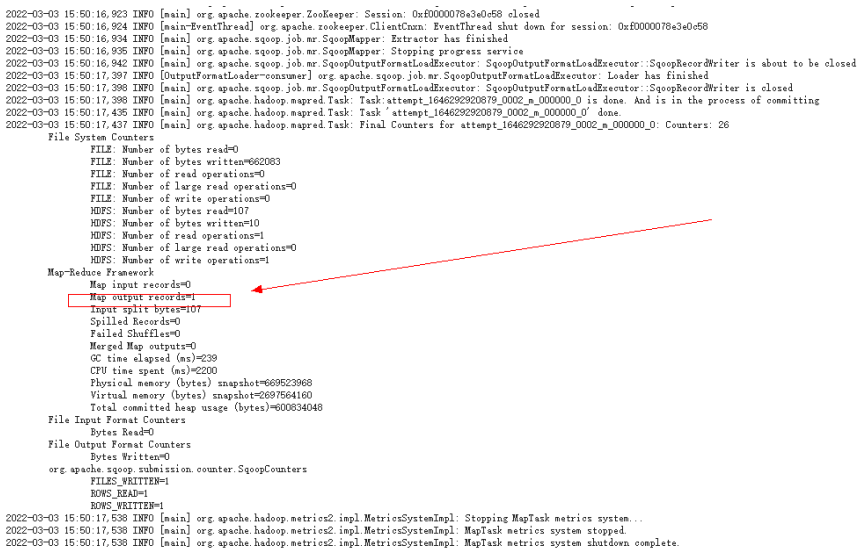
3. 修改文件权限。

   chmod 777 *metrics-core-2.2.0.jar*（真实复制的jar包）

   chown omm:ficommon *metrics-core-2.2.0.jar*（真实复制的jar包）

4. 查看hosts文件，对其他所有节点进行同样的复制jar包操作。

5. 继续运行sqoop任务，成功。



----结束

处理总结

1. 将sqoop的lib下htrace-core-3.1.0-incubating.jar和hbase的lib下的metrics-core-2.2.0.jar，复制到/opt/Bigdata/MRS_1.9.2/install/FusionInsight-Hadoop-2.8.3/hadoop/share/hadoop/common/lib/下。
2. 修改jar包的文件权限为777 和 omm:ficommon。
3. 所有节点均采取以上操作，重新运行sqoop任务即可。

15.17.4 Sqoop 从 hive 到 mysql8.0 报格式错误

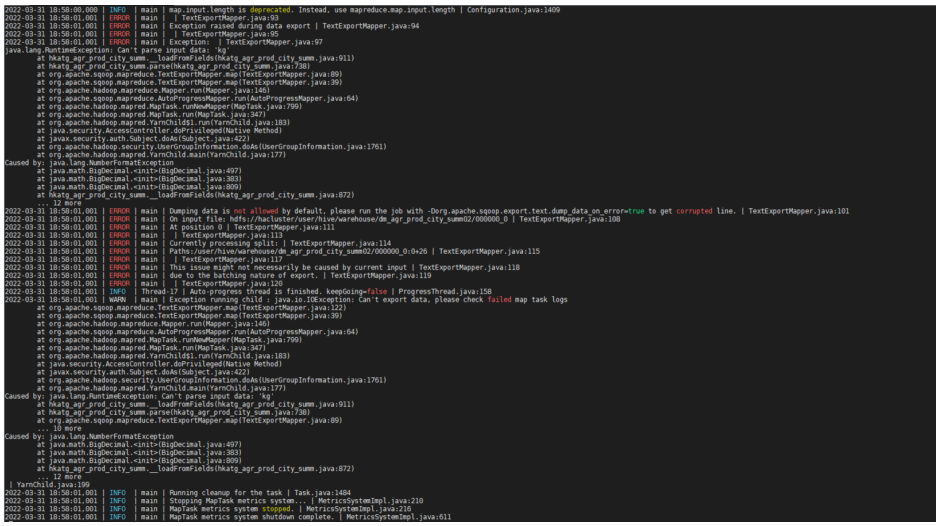
本章节仅适用于MRS 3.1.0版本集群。

用户问题

310集群Sqoop从hive到mysql8.0报格式错误。

问题现象

```
2022-03-31 19:56:44,581 ERROR mapreduce.ExportJobBase: Export job failed!
2022-03-31 19:56:44,581 ERROR tool.ExportTool: Error during export:
Export job failed!
        at org.apache.sqoop.mapreduce.ExportJobBase.runExport(ExportJobBase.java:445)
        at org.apache.sqoop.manager.SqlManager.exportTable(SqlManager.java:931)
        at org.apache.sqoop.tool.ExportTool.exportTable(ExportTool.java:80)
        at org.apache.sqoop.tool.ExportTool.run(ExportTool.java:99)
        at org.apache.sqoop.Sqoop.run(Sqoop.java:147)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:76)
        at org.apache.sqoop.Sqoop.runSqoop(Sqoop.java:183)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:234)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
        at org.apache.sqoop.Sqoop.main(Sqoop.java:252)
```



原因分析

通过日志判断得知是格式异常。

处理步骤

确认分隔符、表字段的格式无问题，在sqoop语句中添加--columns绑定对应字段。

sqoop export --connect jdbc:mysql://IP:端口号/数据库名 --username 用户名 --password 密码 --table 表名 --columns 列字段（多个列用英文逗号分开） -export-dir 导出地址 --fields-terminated-by 分隔符 --input-null-string '\\N' --input-null-non-string '\\N' -m 1

样例：

sqoop export --connect jdbc:mysql://172.16.0.6:3306/lidengpeng --username root --password Mrs@2021 --table hkatg_agr_prod_city_summ --columns year,city_name,city_code,prod_code,prod_name,prod_type,sown_area,area_unit,yield_wegt,yield_unit,total_wegt,total_wegt_unit,data_sorc_code,etl_time -export-dir hdfs://hacluster/user/hive/warehouse/dm_agr_prod_city_summ02 --fields-terminated-by ',' --input-null-string '\\N' --input-null-non-string '\\N' -m 1

15.17.5 Sqoop import 从 pg 到 hive 报错

背景

使用sqoop import命令抽取开源postgre到MRS hdfs或hive等。

用户问题

使用sqoop命令查询postgre表可以，但是执行sqoop import命令导入导出时报错：

The authentication type 5 is not supported. Check that you have configured the pg_hba.conf file to include the client's IP address or subnet。

原因分析

1. 连接postgresql MD5认证不通过，需要在pg_hba.cnf配置白名单。
2. 在执行sqoop import命令时，会启动MapReduce任务，由于MRS Hadoop安装目/opt/Bigdata/FusionInsight_HD_*/1_*_DataNode/install/hadoop/share/hadoop/common/lib下自带了postgre驱动包gsjdbc4-*.jar，与开源postgre服务不兼容导致报错。

处理步骤

1. 客户在pg_hba.cnf配置白名单。
2. 驱动重复，集群自带，将其余驱动排除出去，所有core节点上的gsjdbc4jar包去掉，在sqoop/lib下添加postgrejar包即可。

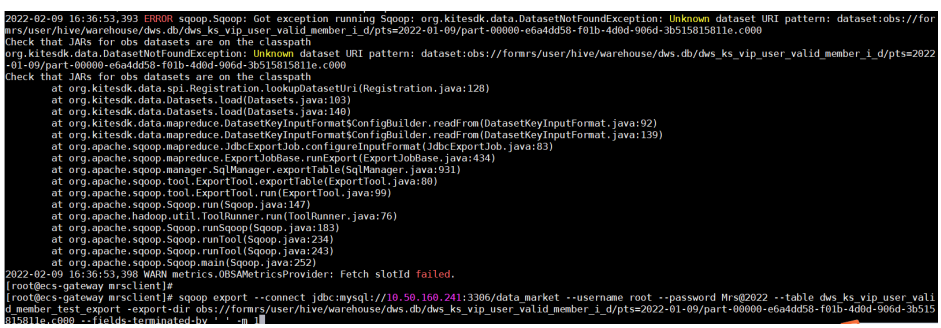
   mv /opt/Bigdata/FusionInsight_HD_*/1_*_DataNode/install/hadoop/share/hadoop/common/lib/gsjdbc4-*.jar /tmp

15.17.6 Sqoop 读 mysql，写 parquet 文件到 OBS 失败

用户问题

sqoop读mysql数据，然后直接写到obs，指定parquet格式时写入报错，不指定parquet时不报错。

问题现象



原因分析

parquet不支持hive3，用Hcatalog方式写入。

处理步骤

采用Hcatalog的方式：参数指定对应的hive库和表，需要修改SQL语句指定到具体字段（需要客户修改脚本）。

具体如下：

客户原来的脚本：

sqoop import --connect 'jdbc:mysql://10.160.5.65/xxx_pos_online_00?zeroDateTimeBehavior=convertToNull' --username root --password Mrs@2022

--split-by id

--num-mappers 2

--query 'select * from pos_remark where 1=1 and $CONDITIONS'

--target-dir obs://za-test/dev/xxx_pos_online_00/pos_remark

--delete-target-dir

--null-string '\\N‘

--null-non-string '\\N‘

--as-parquetfile

修改后的脚本（可以执行成功）：

sqoop import --connect 'jdbc:mysql://10.160.5.65/xxx_pos_online_00?zeroDateTimeBehavior=convertToNull' --username root --password Mrs@2022

--split-by id

--num-mappers 2

--query 'select id,pos_case_id,pos_transaction_id,remark,update_time,update_user,is_deleted,creator,modifier,gmt_created,gmt_modified,update_user_id,tenant_code from pos_remark where 1=1 and $CONDITIONS‘

--hcatalog-database xxx_dev

--hcatalog-table ods_pos_remark

15.18 使用 Storm

15.18.1 Storm 组件的 Storm UI 页面中 events 超链接地址无效

用户问题

Storm组件的Storm UI页面中events超链接地址无效。

问题现象

用户提交拓扑后无法查看拓扑数据处理日志，按钮events地址无效。

原因分析

MRS集群提交拓扑时默认不开启拓扑数据处理日志查看功能。

处理步骤

步骤1 进入服务页面：

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理 ”。
- MRS 2.0.1及之后，在MRS控制台单击集群名称，选择“组件管理 ”。

  说明

  若集群详情页面没有“组件管理”页签，请先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。
- MRS 3.x及后续版本，登录FusionInsight Manager，然后选择“集群 > *待操作的集群名*称 > 服务 ”。

步骤2 登录Storm WebUI：

- MRS 2.x及之前版本：选择“Storm”，在“Storm 概述”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

  说明

  第一次访问Storm WebUI，需要在浏览器中添加站点信任以继续打开页面。
- MRS 3.x及后续版本：选择“Storm > 概览”，在“基本信息”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

步骤3 单击“Topology Summary”区域的指定拓扑名称，打开拓扑的详细信息。

步骤4 在“Topology actions”区域单击“Kill”删除已经提交的Storm拓扑。

步骤5 重新提交Storm拓扑，并开启查看拓扑数据处理日志功能，在提交Storm拓扑时增加参数“topology.eventlogger.executors”，该参数设置为一个不为0的正整数。例如：

storm jar *拓扑包路径 拓扑Main方法的类名称 拓扑名称* -c topology.eventlogger.executors=X

步骤6 在Storm UI界面，单击“Topology Summary”区域的指定拓扑名称，打开拓扑的详细信息。

步骤7 在“Topology actions”区域单击“Debug”，输入采样数据的百分比数值，并单击“OK”开始采样。

步骤8 单击拓扑的“Spouts”或“Bolts”任务名称，在“Component summary”单击“events”即可打开处理数据日志。

说明

如需开启特定“Spouts”或“Bolts”任务的拓扑数据处理日志查看功能，请单击拓扑的“Spouts”或“Bolts”任务名称后，“Topology actions”区域单击“Debug”按钮，输入采样数据的百分比数值。

----结束

15.18.2 提交拓扑失败

问题背景与现象

使用MRS流式集群，主要安装ZooKeeper、Storm、Kafka。

使用客户端命令，提交Topology失败。

可能原因

- Storm服务异常。
- 客户端用户没有进行安全认证或者认证过期。
- 提交拓扑中包含storm.yaml文件和服务端冲突。

原因分析

用户提交拓扑失败，可能原因客户端侧问题或者Storm侧问题。

1. 查看Storm状态。

   MRS Manager：

   登录MRS Manager，在MRS Manager页面，选择“服务管理 > Storm”，查看Storm服务当前状态，发现状态为“良好”，且监控指标内容显示正确。

   FusionInsight Manager界面操作：

   对于MRS 3.x及后续版本集群：登录FusionInsight Manager。选择“集群 > 服务 > Storm”，查看Storm服务当前状态，发现状态为“良好”，且监控指标内容显示正确。

2. 查看客户端提交日志，发现打印KeeperExceptionSessionExpireException异常信息，如下所示：

   ```
   org.apache.zookeeper.KeeperException$SessionExpiredException: KeeperErrorCode = Session expired
           at org.apache.zookeeper.KeeperException.create(KeeperException.java:131) ~[zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.curator.framework.imps.CuratorFrameworkImpl.checkBackgroundRetry(CuratorFrameworkImpl.java:710) [curator-framework-2.5.0.jar:na]
           at org.apache.curator.framework.imps.CuratorFrameworkImpl.processBackgroundOperation(CuratorFrameworkImpl.java:510) [curator-framework-2.5.0.jar:na]
           at org.apache.curator.framework.imps.BackgroundSyncImpl$1.processResult(BackgroundSyncImpl.java:50) [curator-framework-2.5.0.jar:na]
           at org.apache.zookeeper.ClientCnxn$EventThread.processEvent(ClientCnxn.java:684) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn$EventThread.queuePacket(ClientCnxn.java:498) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn.finishPacket(ClientCnxn.java:731) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn.conLossPacket(ClientCnxn.java:765) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn.access$700(ClientCnxn.java:97) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn$SendThread.cleanup(ClientCnxn.java:1391) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
           at org.apache.zookeeper.ClientCnxn$SendThread.run(ClientCnxn.java:1314) [zookeeper-3.5.0.jar:3.5.0-V100R002C00B109]
   2016-08-31 09:25:24 | INFO  | [main] | Session: 0x100273947600ab4b closed | org.apache.zookeeper.ZooKeeper (ZooKeeper.java:968)
   Exception in thread "main" java.lang.RuntimeException: Exception while initializing NimbusLeaderElections
           at backtype.storm.nimbus.NimbusLeaderElections.init(NimbusLeaderElections.java:84)
           at backtype.storm.utils.NimbusClient.getConfiguredClient(NimbusClient.java:39)
           at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:189)
           at backtype.storm.StormSubmitter.submitTopologyWithProgressBar(StormSubmitter.java:256)
           at backtype.storm.StormSubmitter.submitTopologyWithProgressBar(StormSubmitter.java:236)
           at storm.starter.WordCountTopology.main(WordCountTopology.java:94)
   Caused by: org.apache.zookeeper.KeeperException$ConnectionLossException: KeeperErrorCode = ConnectionLoss for /storm/nimbus-leader
           at org.apache.zookeeper.KeeperException.create(KeeperException.java:99)
           at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
           at org.apache.zookeeper.ZooKeeper.exists(ZooKeeper.java:1501)
           at org.apache.curator.framework.imps.ExistsBuilderImpl$2.call(ExistsBuilderImpl.java:172)
           at org.apache.curator.framework.imps.ExistsBuilderImpl$2.call(ExistsBuilderImpl.java:161)
           at org.apache.curator.RetryLoop.callWithRetry(RetryLoop.java:107)
           at org.apache.curator.framework.imps.ExistsBuilderImpl.pathInForeground(ExistsBuilderImpl.java:157)
           at org.apache.curator.framework.imps.ExistsBuilderImpl.forPath(ExistsBuilderImpl.java:148)
           at org.apache.curator.framework.imps.ExistsBuilderImpl.forPath(ExistsBuilderImpl.java:36)
           at backtype.storm.nimbus.NimbusLeaderElections.init(NimbusLeaderElections.java:66)
           ... 5 more
   ```

   上述错误是由于在提交拓扑之前没有进行安全认证或者认证后TGT过期导致。

   解决方法参考**步骤1**。

3. 查看客户端提交日志，发现打印ExceptionInIntializerError异常信息，提示Found multiple storm.yaml resources。如下所示：

   ```
   Exception in thread "main" java.lang.ExceptionInInitializerError
           at backtype.storm.topology.TopologyBuilder.createTopology(TopologyBuilder.java:106)
           at com.huawei.streaming.storm.example.wordcount.WordCountTopology.cmdSubmit(WordCountTopology.java:117)
           at com.huawei.streaming.storm.example.wordcount.WordCountTopology.submitTopology(WordCountTopology.java:80)
           at com.huawei.streaming.storm.example.wordcount.WordCountTopology.main(WordCountTopology.java:71)
   Caused by: java.lang.RuntimeException: Found multiple storm.yaml resources. You're probably bundling the Storm jars with your topology jar.
           at backtype.storm.utils.Utils.findAndReadConfigFile(Utils.java:151)
           at backtype.storm.utils.Utils.readStormConfig(Utils.java:206)
           at backtype.storm.utils.Utils.<clinit>(Utils.java:70)
           ... 4 more
   ```

该错误是由于业务jar包中存在storm.yaml文件，和服务端的storm.yaml文件冲突导致的。

解决方法参考**步骤2**。

4. 如果不是上述原因，则请参考**提交拓扑失败，提示Failed to check principle for keytab**。

解决办法

步骤1 认证异常。

1. 登录客户端节点，进入客户端目录。
2. 执行以下命令重新提交任务。（业务jar包和Topology根据实际情况替换）

   source bigdata_env

   kinit *用户名*

   storm jar storm-starter-topologies-0.10.0.jar storm.starter.WordCountTopology test

步骤2 拓扑包异常。

排查业务jar，将业务jar中storm.yaml文件删除，重新提交任务。

----结束

15.18.3 提交拓扑失败，提示 Failed to check principle for keytab

问题背景与现象

使用MRS流式安全集群，主要安装ZooKeeper、Storm、Kafka等。

定义拓扑访问HDFS、HBase等组件，使用客户端命令，提交Topology失败。

可能原因

- 提交拓扑中没有包含用户的keytab文件。
- 提交拓扑中包含的keytab和提交用户不一致。
- 客户端/tmp目录下已存在user.keytab，且宿主非运行用户。

原因分析

1. 查看日志发现异常信息Can not found user.keytab in storm.jar。具体信息如下：

```
[main] INFO  b.s.StormSubmitter - Get principle for stream@HADOOP.COM success
[main] ERROR b.s.StormSubmitter - Can not found user.keytab in storm.jar.
Exception in thread "main" java.lang.RuntimeException: Failed to check principle for keytab
at backtype.storm.StormSubmitter.submitTopologyAs(StormSubmitter.java:219)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:292)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:176)
at com.xxx.streaming.storm.example.hbase.SimpleHBaseTopology.main(SimpleHBaseTopology.java:77)
```

   查看提交的拓扑运行Jar，发现没有包含keytab文件。

2. 查看日志发现异常信息The submit user is invalid,the principle is 。具体信息如下：

```
[main] INFO  b.s.StormSubmitter - Get principle for stream@HADOOP.COM success
[main] WARN  b.s.s.a.k.ClientCallbackHandler - Could not login: the client is being asked for a password, but the  client code does not currently support obtaining a password from the user. Make sure that the client is configured to use a ticket cache (using the JAAS configuration setting
```

```
'useTicketCache=true)' and restart the client. If you still get this message after that, the TGT in the ticket cache has expired and must be manually refreshed. To do so, first determine if you are using a password or a keytab. If the former, run kinit in a Unix shell in the environment of the user who is running this client using the command 'kinit <princ>' (where <princ> is the name of the client's Kerberos principal). If the latter, do 'kinit -k -t <keytab> <princ>' (where <princ> is the name of the Kerberos principal, and <keytab> is the location of the keytab file). After manually refreshing your cache, restart this client. If you continue to see this message after manually refreshing your cache, ensure that your KDC host's clock is in sync with this host's clock.
[main] ERROR b.s.StormSubmitter - The submit user is invalid,the principle is : stream@HADOOP.COM
Exception in thread "main" java.lang.RuntimeException: Failed to check principle for keytab
at backtype.storm.StormSubmitter.submitTopologyAs(StormSubmitter.java:219)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:292)
at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:176)
at com.xxx.streaming.storm.example.hbase.SimpleHBaseTopology.main(SimpleHBaseTopology.java:77)
```

业务提交拓扑时使用的认证用户为stream，但是在拓扑提交过程中提示submit user是无效用户，表明内部校验失败。

3. 查看提交的拓扑运行Jar，发现包含keytab文件。

   查看user.keytab文件，发现principal为zmk_kafka.

   ```
   [root@8-5-148-6 client]# klist -kt user.keytab
   Keytab name: FILE:user.keytab
   KVNO Timestamp         Principal
   ---- ----------------- --------------------------------------------------------
      1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
      1 12/19/16 16:28:17 zmk_kafka@HADOOP.COM
   ```

   发现认证用户和user.keytab文件中principal不对应。

4. 查看日志发现异常信息Delete the tmp keytab file failed，the keytab file is ：/tmp/user.keytab，具体信息如下：

   ```
   [main] WARN b.s.StormSubmitter - Delete the tmp keytab file failed, the keytab file is : /tmp/user.keytab
   [main] ERROR b.s.StormSubmitter - The submit user is invalid,the principle is : hbase1@HADOOP.COM
   Exception in thread "main" java.lang.RuntimeException: Failed to check principle for keytab
   at backtype.storm.StormSubmitter.submitTopologyAs(StormSubmitter.java:213)
   at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:286)
   at backtype.storm.StormSubmitter.submitTopology(StormSubmitter.java:170)
   at com.touchstone.storm.cmcc.CmccDataHbaseTopology.main(CmccDataHbaseTopology.java:183)
   ```

   查看系统/tmp目录，发现存在user.keytab文件，且文件宿主非运行用户。

解决办法

- 提交拓扑时携带用户user.keytab文件。
- 提交拓扑时的用户需要和user.keytab文件用户一致。
- 删除/tmp目录下不对应的user.keytab文件。

15.18.4 提交拓扑后 Worker 日志为空

现象描述

在Eclipse中远程提交拓扑成功之后，无法在Storm WebUI查看拓扑的详细信息，并且每个拓扑的Bolt和Spout所在Worker节点在一直变化。查看Worker日志，日志内容为空。

可能原因

Worker进程启动失败，触发Nimbus重新分配任务，在其他Supervisor上启动Worker。由于Worker启动失败后会继续重启，导致Worker节点在一直变化，且Worker日志内容为空。Worker进程启动失败的可能原因有两个：

- 提交的Jar包中包含“storm.yaml”文件。

  Storm规定，每个“classpath”中只能包含一个“storm.yaml”文件，如果多于一个那么就会产生异常。使用Storm客户端提交拓扑，由于客户端“classpath”配置和Eclipse远程提交方式“classpath”不一样，客户端会自动加载用户的Jar包到“classpath”，从而使“classpath”中存在两个“storm.yaml”文件。
- Worker进程初始化时间较长，超过Storm集群设置Worker启动超时时间，导致Worker被Kill从而一直进行重分配。

定位思路

1. 使用Storm客户端提交拓扑，检查出重复“storm.yaml”问题。
2. 重新打包Jar包，然后再提交拓扑。
3. 修改Storm集群关于Worker启动超时参数。

处理步骤

步骤1 使用Eclipse远程提交拓扑后Worker日志为空，则使用Storm客户端，提交拓扑对应的Jar包，查看提示信息。

例如，Jar包中包含两个不同路径下的“storm.yaml”文件，系统显示以下信息：

```
Exception in thread "main" java.lang.ExceptionInInitializerError
  at com.xxx.streaming.storm.example.WordCountTopology.createConf(WordCountTopology.java:132)
  at com.xxx.streaming.storm.example.WordCountTopology.remoteSubmit(WordCountTopology.java:120)
  at com.xxx.streaming.storm.example.WordCountTopology.main(WordCountTopology.java:101)
 Caused by: java.lang.RuntimeException: Found multiple storm.yaml resources. You're probably bundling the Storm jars with your topology jar. [jar:file:/opt/xxx/fi_client/Streaming/streaming-0.9.2/bin/stormDemo.jar!/storm.yaml, file:/opt/xxx/fi_client/Streaming/streaming-0.9.2/conf/storm.yaml]
  at backtype.storm.utils.Utils.findAndReadConfigFile(Utils.java:151)
  at backtype.storm.utils.Utils.readStormConfig(Utils.java:206)
  at backtype.storm.utils.Utils.<(Utils.java:70)>
```

步骤2 重新打包Jar包，且不能包含“storm.yaml”文件、“log4j”和“slf4j-log4j”相关的Jar包。

步骤3 使用IntelliJ IDEA远程提交新打包的Jar包。

步骤4 查看是否可以在WebUI查看拓扑的详细信息和Worker日志内容。

步骤5 在Manager页面修改Storm集群关于Worker启动超时参数（参数说明请参考**参考信息**），保存并重启Storm服务。

- MRS Manager界面操作入口：登录MRS Manager，依次选择 “服务管理 > Storm> 配置”。
- FusionInsight Manager界面操作入口：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Storm > 配置"

步骤6 重新提交待运行的Jar包。

----结束

参考信息

1. nimbus.task.launch.secs和supervisor.worker.start.timeout.secs这两个参数分别代表nimbus端和supervisor端对于拓扑启动的超时容忍时间，一般nimbus.task.launch.secs的值要大于等于supervisor.worker.start.timeout.secs的值（建议相等或略大，如果超出太多会影响任务重分配的效率）。

- nimbus.task.launch.secs：nimbus在超过该参数配置的时间内没有收到拓扑的task发的心跳时，会将该拓扑重新分配（分配给别的supervisor），同时会刷新zk中的任务信息，supervisor读到zk中的任务信息并且与自己当前所启动的拓扑进行比较，如果存在拓扑已经不属于自己，那么则会删除该拓扑的元数据，也就是/srv/Bigdata/streaming_data/stormdir/supervisor/stormdist/{worker-id}目录。
- supervisor.worker.start.timeout.secs：supervisor启动worker后，在该参数配置的时间内没有收到worker的心跳时，supervisor会主动停掉worker，等待worker的重新调度，一般在业务启动时间较长时适当增加该参数的值，保证worker能启动成功。

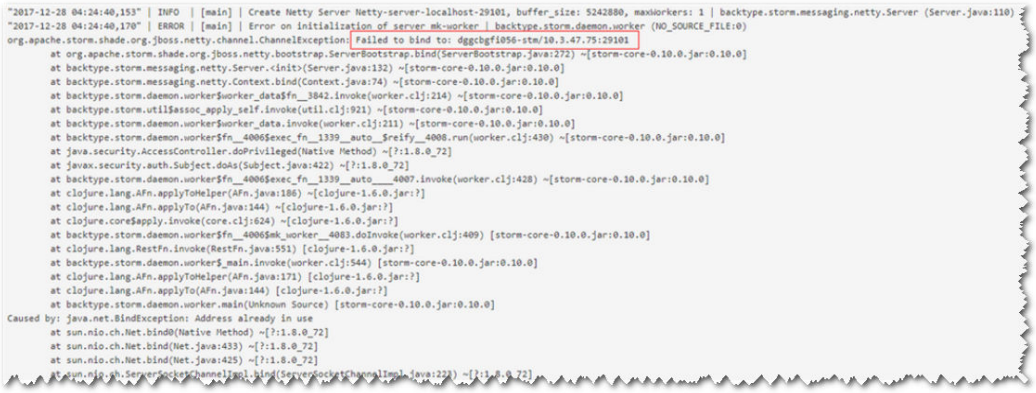
  如果supervisor.worker.start.timeout.secs配置的值比nimbus.task.launch.secs的值大，那么则会出现supervisor的容忍时间没到，仍然继续让worker启动，而nimbus却认定该业务启动超时，将该业务分配给了其他主机，这时supervisor的后台线程发现任务不一致，删除了拓扑的元数据，导致接下来worker在启动过程中要读取stormconf.ser时，发现该文件已经不存在了，就会抛出FileNotFoundException。

2. nimbus.task.timeout.secs和supervisor.worker.timeout.secs这两个参数则分别代表nimbus端和supervisor端对于拓扑运行过程中心跳上报的超时容忍时间，一般nimbus.task.timeout.secs的值要大于等于supervisor.worker.timeout.secs的值（建议相等或略大），原理同上。

15.18.5 提交拓扑后 Worker 运行异常，日志提示 Failed to bind to：host:ip

现象描述

提交业务拓扑后，发现Worker无法正常启动。查看Worker日志，日志提示Failed to bind to：host:ip。



可能原因

随机端口范围配置错误。

定位思路

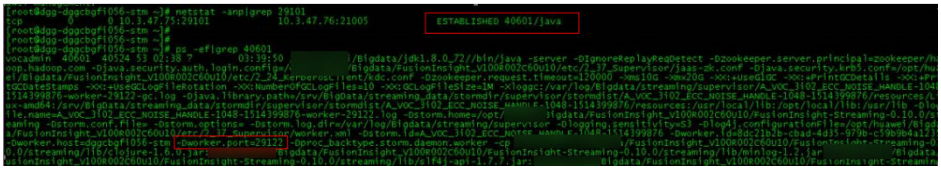
1、检查worker相关信息日志。

2、检查绑定端口的进程信息。

3、检查随机端口范围配置。

原因分析

1. 通过SSH登录Worker启动失败主机，通过**netstat -anp | grep <port>**命令，查看占用端口的进程ID信息。其中port修改为实际端口号。
2. 通过**ps -ef | grep <pid>**命令查看进程的详细信息，其中pid为查询出的实际进程ID。



发现占用端口的进程为worker进程，该进程为另一个拓扑业务进程。同时根据进程详细信息发现，分配给该进程的端口为29122。

3. 通过**lsof -i:<port>**命令，查看连接详细信息。其中port为实际端口号。

发现29101端口连接对端端口为21005，而21005为Kafka服务端端口。

说明业务层作为客户端连接Kafka获取消息，业务端口分配通过OS的随机端口分配范围来确定。

4. 通过**cat /proc/sys/net/ipv4/ip_local_port_range**命令查看随机端口范围。

5. 发现随机端口范围过大，和MRS的服务端口范围存在冲突。

说明

MRS的服务端口范围：20000-30000。

处理步骤

步骤1 修改随机端口范围。

```
vi /proc/sys/net/ipv4/ip_local_port_range
32768 61000
```

步骤2 停止占用服务端口的业务进程，释放端口。（停止业务拓扑）

----结束

15.18.6 使用 jstack 命令查看进程堆栈提示 well-known file is not secure

问题背景与现象

使用jstack命令查看进程堆栈信息报：well-known file is not secure。

```
omm@hadoop02:~> jstack 62517
62517: well-known file is not secure
```

原因分析

1. 由于执行命令的用户与当前查看pid信息的进程提交用户不一致导致。
2. Storm引入区分用户执行任务特性，在启动worker进程时将给进程的uid和gid改为提交用户和ficommon，目的是为了logviewer可以访问到worker进程的日志同时日志文件只开放权限到640。这样会导致切换到提交用户后对Worker进程执行jstack和jmap等命令执行失败，原因是提交用户的默认gid并不是ficommon，需要通过ldap命令修改提交用户的gid为9998（ficommon）才可执行。

解决办法

共有两种方式解决该问题。

方式一：通过storm原生页面查看进程堆栈

步骤1 登录Storm原生界面。

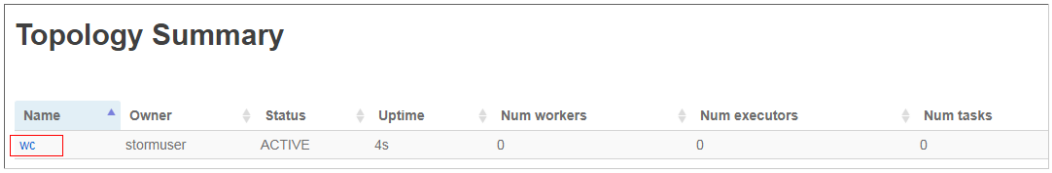
MRS Manager界面操作：

1. 访问MRS Manager。
2. 在Manager选择“服务管理 > Storm”，在“Storm 概述”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

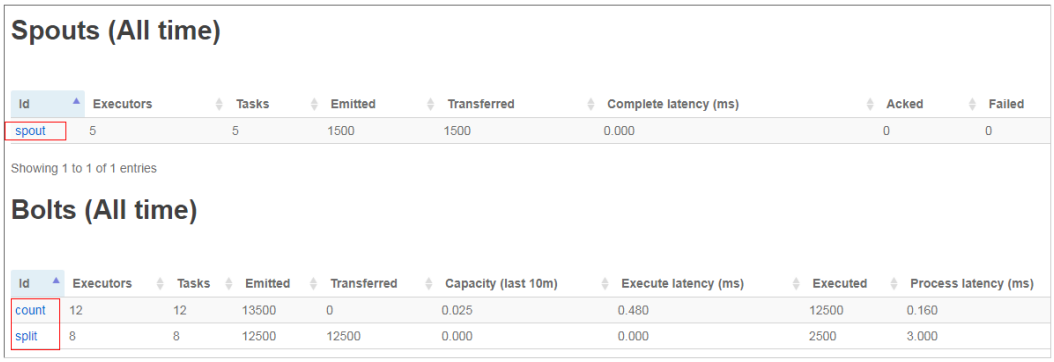
FusionInsight Manager界面操作：

1. 访问FusionInsight Manager。
2. 在Manager选择“集群 > 服务 > Storm”，在“概览”的“Storm WebUI”，单击任意一个UI链接，打开Storm的WebUI。

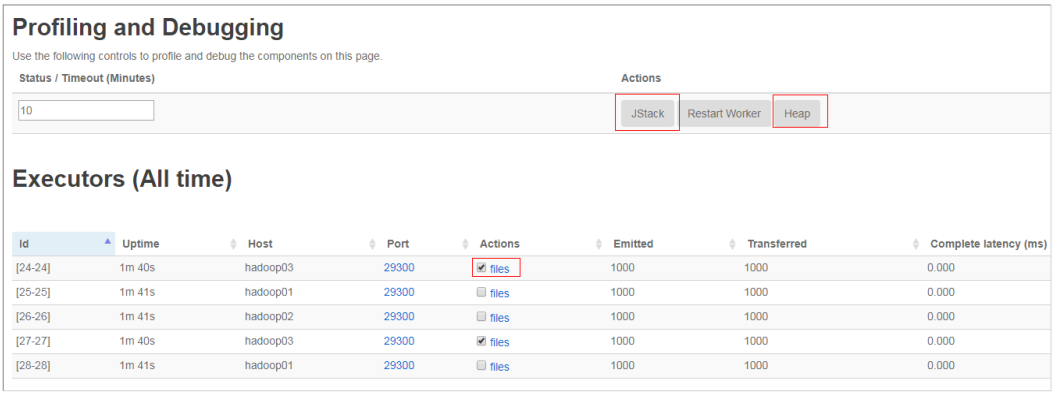
步骤2 选择要查看的拓扑。



步骤3 选择要查看的spout或者bolt。



步骤4 选择要查看的节点日志文件，再选择JStack或者Heap按钮，其中JStack对应的是堆栈信息，Heap对应的是堆信息：



----结束

方式二：通过修改自定义参数查看进程堆栈

步骤1 进入Storm服务参数配置界面。

MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Storm > 服务配置”，“参数类别”选择“全部配置”。

FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 服务 > Yarn”，单击“配置”，选择“全部配置”。

步骤2 在左侧导航栏选择“supervisor > 自定义”，添加一个变量supervisor.run.worker.as.user=false。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

步骤4 重新提交拓扑。

步骤5 后台节点切为omm用户执行jps命令即可查看worker的pid。

```
omm@hadoop02:~> jps | grep worker
22485 worker
111402 worker
```

步骤6 执行jstack pid，即可查看jstack信息。

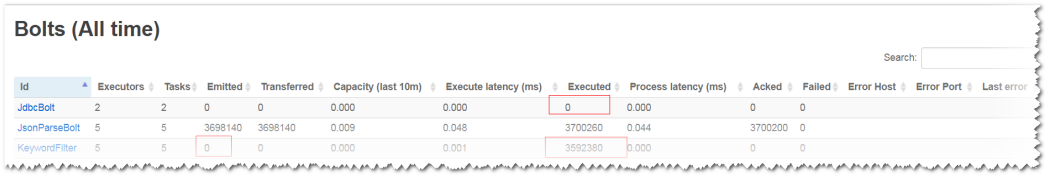
```
omm@hadoop02:~> jstack 22485
2018-05-26 08:46:24
Full thread dump Java HotSpot(TM) 64-Bit Server VM (25.144-b01 mixed mode):

"Attach Listener" #82 daemon prio=9 os_prio=0 tid=0x0000000001c95000 nid=0xb840 waiting on condition [0x0000000000000000]
   java.lang.Thread.State: RUNNABLE

"pool-14-thread-1" #81 daemon prio=5 os_prio=0 tid=0x00007f7ebc931000 nid=0x6113 waiting on condition [0x00007f7eb5ddf000]
   java.lang.Thread.State: TIMED_WAITING (parking)
        at sun.misc.Unsafe.park(Native Method)
        - parking to wait for  <0x00000000dfe820a0> (a java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject)
        at java.util.concurrent.locks.LockSupport.parkNanos(LockSupport.java:215)
        at java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject.awaitNanos(AbstractQueuedSynchronizer.java:2078)
        at java.util.concurrent.ScheduledThreadPoolExecutor$DelayedWorkQueue.take(ScheduledThreadPoolExecutor.java:1093)
        at java.util.concurrent.ScheduledThreadPoolExecutor$DelayedWorkQueue.take(ScheduledThreadPoolExecutor.java:809)
        at java.util.concurrent.ThreadPoolExecutor.getTask(ThreadPoolExecutor.java:1074)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1134)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
```

----结束

15.18.7 使用 Storm-JDBC 插件开发 Oracle 写入 Bolt，发现数据无法写入

现象描述

使用Storm-JDBC插件开发Oracle写入Bolt，发现能连上Oracle数据库，但是无法向Oracle数据库里面写数据。

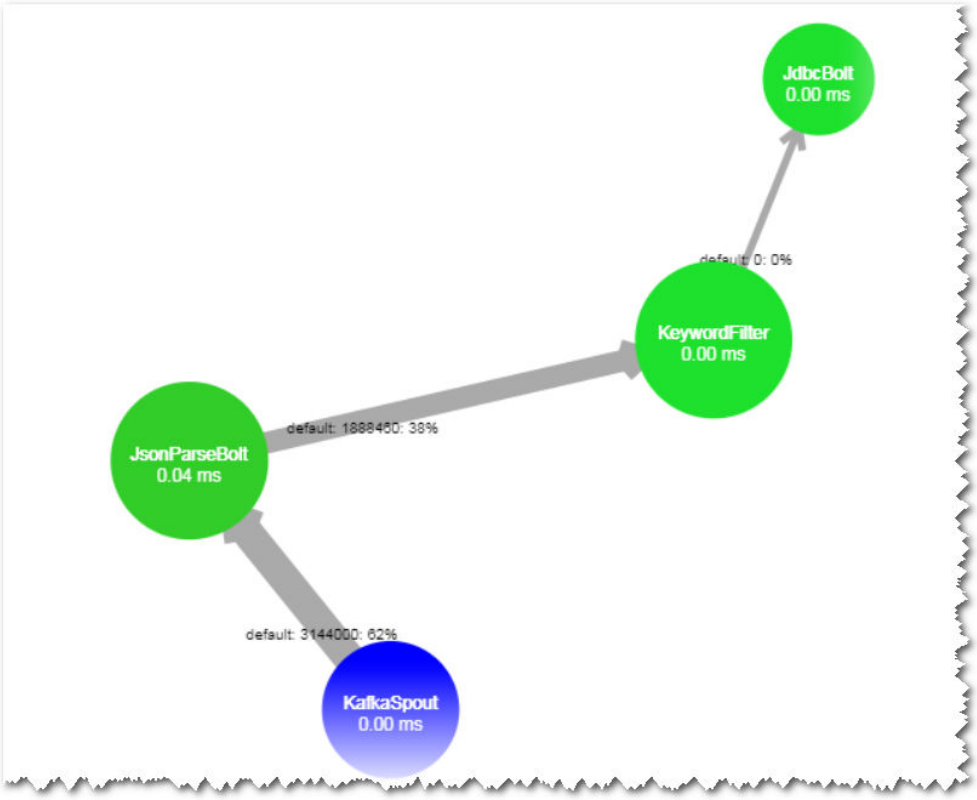


可能原因

- 拓扑定义异常。
- 数据库表结果定义异常。

原因分析

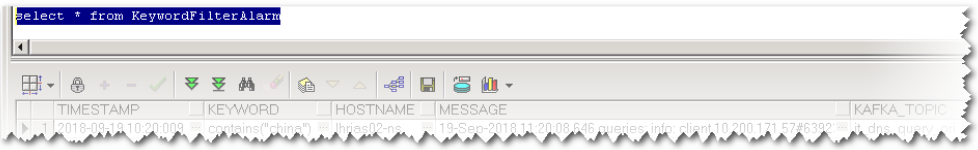
1. 通过Storm WebUI 查看拓扑DAG图，发现DAG图与拓扑定义一致。



2. 查看KeyWordFilter Bolt输出流字段定义和发送消息字段发现一致。

```
@Override
public void declareOutputFields(OutputFieldsDeclarer declarer)
{
    declarer.declare(new Fields("timestamp","keyword","hostname","message","kafka_topic") );
```

3. 查看Oracle数据库中表定义，发现字段名为大写，与流定义字段名称不一致。



4. 单独调试execute方法，发现抛出字段不存在。

处理步骤

修改流定义字段名称为大写，与数据库表定义字段一致。

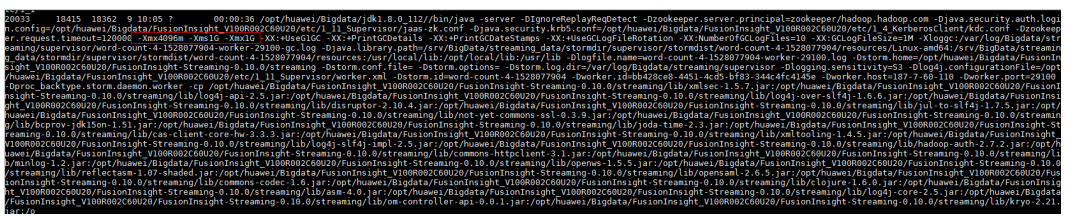
15.18.8 业务拓扑配置 GC 参数不生效

问题背景与现象

业务拓扑代码中配置参数topology.worker.childopts不生效，关键日志如下：

```
[main] INFO b.s.StormSubmitter - Uploading topology jar /opt/jar/example.jar to assigned location: /srv/
BigData/streaming/stormdir/nimbus/inbox/stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar
Start uploading file '/opt/jar/example.jar' to '/srv/BigData/streaming/stormdir/nimbus/inbox/
stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar' (65574612 bytes)
[==========================================================] 65574612 / 65574612
File '/opt/jar/example.jar' uploaded to '/srv/BigData/streaming/stormdir/nimbus/inbox/
stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar' (65574612 bytes)
[main] INFO b.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /srv/BigData/
streaming/stormdir/nimbus/inbox/stormjar-8d3b778d-69ea-4fbe-ba88-01aa2036d753.jar
[main] INFO b.s.StormSubmitter - Submitting topology word-count in distributed mode with conf
{"topology.worker.childopts":"-
Xmx4096m","storm.zookeeper.topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-59
15065013522446406:-6421330379815193999","topology.workers":1}
[main] INFO b.s.StormSubmitter - Finished submitting topology: word-count
```

通过ps -ef | grep worker命令查看worker进程信息如下：



原因分析

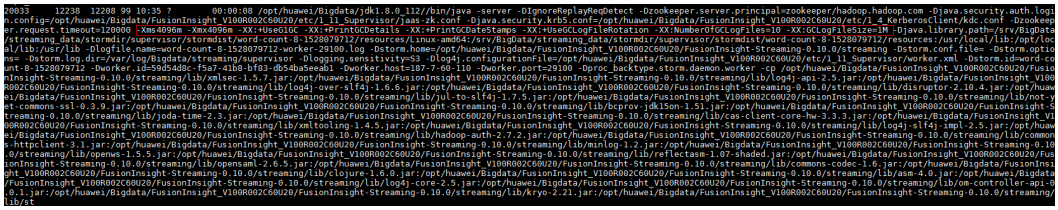
1. 由于topology.worker.gc.childopts、topology.worker.childopts和worker.gc.childopts(服务端参数)有优先级，优先级大小为：topology.worker.gc.childopts > worker.gc.childopts > topology.worker.childopts。
2. 如果设置了客户端参数topology.worker.childopts，则该参数会与服务端参数worker.gc.childopts共同配置，但是后面的相同参数会将前面的覆盖掉，如上面图有两个-Xmx，-Xmx1G会覆盖掉-Xmx4096m。
3. 如果配置了topology.worker.gc.childopts则服务端参数worker.gc.childopts会被替换。

解决办法

步骤1 如果想要修改拓扑的JVM参数，可以在命令中直接修改topology.worker.gc.childopts这个参数或者在服务端修改该参数，当topology.worker.gc.childopts为 "-Xms4096m -Xmx4096m -XX:+UseG1GC -XX:+PrintGCDetails -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M" 时，效果如下：

```
[main-SendThread(10.7.61.88:2181)] INO o.a.s.s.o.a.z.ClientCnxn - Socket connection established, initiating session, client: /10.7.61.88:44694, server: 10.7.61.88/10.7.61.88:2181
[main-SendThread(10.7.61.88:2181)] INFO o.a.s.s.o.a.z.ClientCnxn - Session establishment complete on server 10.7.61.88/10.7.61.88:2181, sessionid = 0x16037a6e5f092575, negotiated timeout = 40000
[main-EventThread] INFO o.a.s.s.o.a.c.f.s.ConnectionStateManager - State change: CONNECTED
[main] INFO b.s.u.StormBoundedExponentialBackoffRetry - The baseSleepTimeMs [1000] the maxSleepTimeMs [1000] the maxRetries [1]
[main] INFO o.a.s.s.o.a.z.Login - successfully logged in.
[main-EventThread] INFO o.a.s.s.o.a.z.ClientCnxn - EventThread shut down for session: 0x16037a6e5f092575
[main] INFO o.a.s.s.o.a.z.ZooKeeper - Session: 0x16037a6e5f092575 closed
[main] INFO b.s.StormSubmitter - Uploading topology jar /opt/jar/example.jar to assigned location: /srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar
Start uploading file '/opt/jar/example.jar' to '/srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar' (74143745 bytes)
[==================================================] 74143745 / 74143745
File '/opt/jar/example.jar' uploaded to '/srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar' (74143745 bytes)
[main] INFO b.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /srv/BigData/streaming/stormdir/nimbus/inbox/stormjar-86855b6b-133e-478d-b415-fa96e63e553f.jar
[main] INFO b.s.StormSubmitter - Submitting topology word-count in distributed mode with conf {"storm.zookeeper.topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-7360002804241426074:-6868950379453400421","topology.worker.gc.childopts":"-Xms4096m -Xmx4096m -XX:+UseG1GC -XX:+PrintGCDetails -XX:+PrintGCDateStamps -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=10 -XX:GCLogFileSize=1M","topology.workers":1}
[main] INFO b.s.StormSubmitter - Finished submitting topology: word-count
```

步骤2 通过ps -ef | grep worker命令查看worker进程信息如下：



----结束

15.18.9 UI 查看信息显示 Internal Server Error

问题背景与现象

使用MRS版本安装集群，主要安装ZooKeeper、Strom。

通过MRS Manager中的Storm Status页面UI连接访问信息时出现Internal Server Error。

UI页面出现如下信息：

```
Internal Server Error
org.apache.thrift7.transport.TTransportException: Frame size (306030) larger than max length (1048576)!
```

可能原因

- Storm服务中Nimbus异常。
- Storm集群信息较多超过系统默认Thrift传输大小的设置。

原因分析

1. 查看Storm服务状态及监控指标：
   - MRS Manager界面操作：登录MRS Manager，依次选择 "服务管理 > Storm ，查看当前Storm状态，发现状态为良好，且监控指标内容显示正确。
   - FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 待操作集群的名称 > 服务 > Storm，

     查看当前Storm状态，发现状态为良好，且监控指标内容显示正确。
2. 选择“实例”页签，查看Nimbus实例状态，显示正常。
3. 查看当前Storm集群thrift相关配置，发现nimbus.thrift.max_buffer_size参数配置为1048576（1M）。
4. 上述配置和异常信息中信息一致，说明当前配置的Thrift的buffer size小于集群信息所需的buffer size。

解决方法

调整Storm集群中Thrift的Buffer Size大小，具体大小根据错误信息进行实际调整。

步骤1 进入Storm服务参数配置界面。

- MRS Manager界面操作：登录MRS Manager页面，选择“服务管理 > Storm > 服务配置”，“参数类别”选择“全部配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 服务 > Yarn”，单击“配置”，选择“全部配置”。

步骤2 修改nimbus.thrift.max_buffer_size参数为10485760（10M）。

步骤3 保存配置，勾选“重新启动受影响的服务或实例。”并单击“确定”重启服务。

----结束

15.19 使用 Ranger

15.19.1 Hive 启用 Ranger 鉴权后，在 Hue 页面能查看到没有权限的表和库

用户问题

Hive启用Ranger鉴权后，在Hue页面能查看到没有权限的表和库

问题现象

普通集群（未开启Kerberos认证）中，Hive启用Ranger鉴权后，在Hue页面能查看到没有权限的表和库。

原因分析

Hive启用Ranger鉴权后，默认的Hive策略中有2个关于database的public组策略， 所有用户都属于public组，默认给public组配有default数据库的创表和所有其他数据库的**create**权限，因此默认所有的用户都有**show databases**和**show tables**的权限，如果不想让某些用户有**show databases**和**show tables**权限，可在Ranger WEBUI中删除该默认public组策略，并赋予需要查看的用户权限，具体请参考处理步骤。

处理步骤

步骤1 登录Ranger WebUI界面。

步骤2 在“Service Manager”区域内，单击Hive组件名称，进入Hive组件安全访问策略列表页面。

步骤3 分别单击“all - database”和“default database tables columns”策略所在行的按钮。

步骤4 删除“public”组策略。

图 15-58 all - database 策略

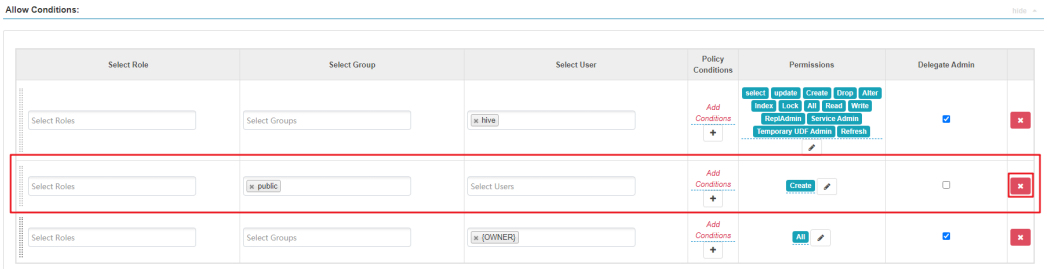
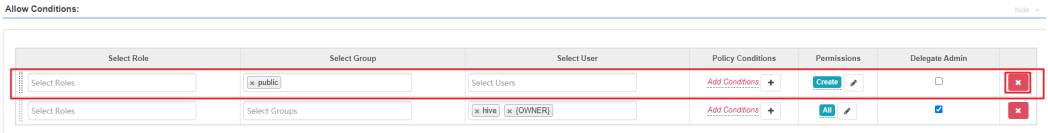


图 15-59 default database tables columns 策略



步骤5 在Hive组件安全访问策略列表页面，单击“Add New Policy”为相关用户或者用户组添加资源访问策略。

----结束

15.20 使用 Yarn

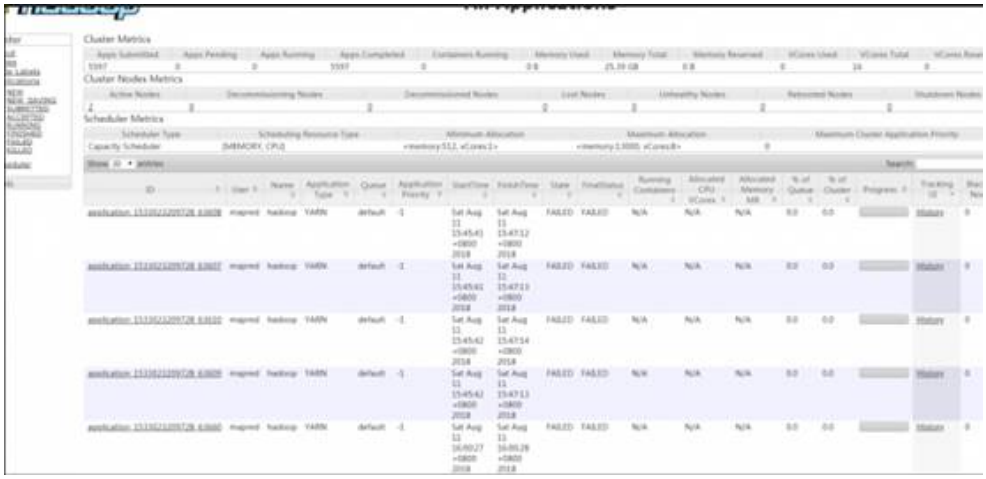
15.20.1 启动 Yarn 后发现一堆 job

用户问题

MRS 2.x及之前版本集群，构建MRS集群启动Yarn后，发现一堆job占用资源。

问题现象

客户使用MapReduce服务构建集群启动Yarn后 生成一堆job占用资源。



原因分析

- 疑似黑客攻击。
- 安全组入口方向的Any协议源地址配置为0.0.0.0/0。

处理步骤

步骤1 登录MRS集群页面，在“现有集群”中，单击对应的集群名称，进入集群详情页面。

步骤2 单击“集群管理页面”后面的“点击管理”，弹出“访问MRS Manager页面”。

步骤3 单击“管理安全组规则”，检查安全组规则配置。

步骤4 检查入口方向Any协议的源地址是否为0.0.0.0/0。

步骤5 如果是，修改入口方向Any协议的远端为指定IP地址。如果不是，则无需修改。

步骤6 修改成功后，重启集群虚机。

----结束

建议与总结

关闭入口方向的Any协议，或者指定入口方向的Any协议远端为指定IP。

参考信息

请参考“MapReduce服务用户指南 > 安全性 > 集群（未启用Kerberos认证）安全配置建议”。

15.20.2 通过客户端 hadoop jar 命令提交任务，客户端返回 GC overhead

问题背景与现象

通过客户端提交任务，客户端返回内存溢出的报错结果:

```
main path:hdfs://hacluster/user/wangyou
17/09/18 08:29:57 INFO hdfs.DFSClient: Created HDFS_DELEGATION_TOKEN token 22890097 for wangyou on ha-hdfs:hacluster
17/09/18 08:29:57 INFO security.TokenCache: Got dt for hdfs://hacluster; Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 2
2890097 for wangyou)
17/09/18 08:29:57 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
17/09/18 08:32:42 INFO retry.RetryInvocationHandler: Exception while invoking getListing of class ClientNamenodeProtocolTranslatorPB over fi1-cn-003/10.113.246.10:2500
0. Trying to fail over immediately.
java.io.IOException: com.google.protobuf.ServiceException: java.lang.OutOfMemoryError: GC overhead limit exceeded
        at org.apache.hadoop.ipc.ProtobufHelper.getRemoteException(ProtobufHelper.java:47)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.getListing(ClientNamenodeProtocolTranslatorPB.java:578)
        at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:497)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:191)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:102)
        at com.sun.proxy.$Proxy10.getListing(Unknown Source)
        at org.apache.hadoop.hdfs.DFSClient.listPaths(DFSClient.java:1757)
        at org.apache.hadoop.hdfs.DistributedFileSystem$DirListingIterator.hasNextNoFilter(DistributedFileSystem.java:1024)
        at org.apache.hadoop.hdfs.DistributedFileSystem$DirListingIterator.hasNext(DistributedFileSystem.java:999)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:304)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
        at org.apache.hadoop.mapreduce.lib.input.CombineFileInputFormat.getSplits(CombineFileInputFormat.java:217)
        at org.apache.hadoop.mapreduce.lib.input.DelegatingInputFormat.getSplits(DelegatingInputFormat.java:115)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:306)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:323)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:200)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1290)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1287)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1673)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1287)
```

原因分析

从报错堆栈可以看出是任务在提交过程中分片时在读取HDFS文件阶段内存溢出了，一般是由于该任务要读取的小文件很多导致内存不足。

解决办法

步骤1 排查启动的MR任务是否对应的HDFS文件个数很多，如果很多，减少文件数量，提前先合并小文件或者尝试使用combineInputFormat来减少任务读取的文件数量。

步骤2 增大hadoop命令执行时的内存，该内存在客户端中设置，修改对应路径“*客户端安装目录*/HDFS/component_env”文件中“CLIENT_GC_OPTS”的“-Xmx”参数，将该参数的默认值改大，比如改为512m。然后执行**source component_env**命令，使修改的参数生效。

```
export YARN_ROOT_LOGGER=INFO,console

#GC_OPTS for client operation.
CLIENT_GC_OPTS="-Xmx512m -Djava.io.tmpdir=${HADOOP_HOME}"

export HADOOP_CLIENT_OPTS="$CLIENT_GC_OPTS"
```

----结束

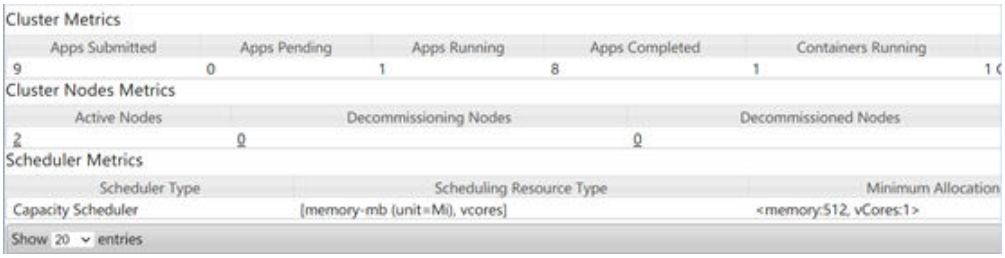
15.20.3 Yarn 汇聚日志过大导致磁盘被占满

用户问题

集群的磁盘使用率很高。

问题现象

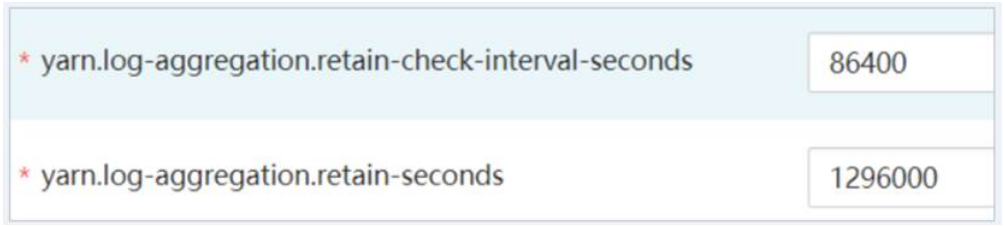
- Manager管理页面下主机管理显示磁盘使用率过高。
- Yarn WebUI界面上显示只有少量任务在运行。



- 登录到集群的Master节点执行**hdfs dfs -du -h /** 命令发现如下文件占用大量磁盘空间。

```
22.5 G   45.0 G   /tmp/logs/root/logs/application_1589278244866_0153
18.4 M   36.8 M   /tmp/logs/root/logs/application_1589278244866_0154
23.4 G   46.8 G   /tmp/logs/root/logs/application_1589278244866_0155
23.5 G   46.9 G   /tmp/logs/root/logs/application_1589278244866_0156
23.7 G   47.4 G   /tmp/logs/root/logs/application_1589278244866_0157
23.7 G   47.4 G   /tmp/logs/root/logs/application_1589278244866_0158
22.5 G   45.0 G   /tmp/logs/root/logs/application_1589278244866_0159
18.5 M   37.0 M   /tmp/logs/root/logs/application_1589278244866_0160
22.5 G   45.0 G   /tmp/logs/root/logs/application_1589278244866_0161
18.8 M   37.6 M   /tmp/logs/root/logs/application_1589278244866_0162
24.0 G   48.0 G   /tmp/logs/root/logs/application_1589278244866_0163
121.3 K  242.7 K  /tmp/logs/root/logs/application_1589278244866_0164
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0165
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0166
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0167
1.1 M    2.1 M    /tmp/logs/root/logs/application_1589278244866_0168
```

- Yarn服务的汇聚日志配置如下



原因分析

客户提交任务的操作过于频繁，且聚合后的日志文件被删除的时间配置为1296000，即聚合日志保留15天，导致汇聚的日志无法在短时时间内释放，从而引起磁盘被占满。

处理步骤

步骤1 登录Manager页面，进入MapReduce服务全部配置页面。

- MRS Manager界面操作：登录MRS Manager，选择“服务管理 > MapReduce > 服务配置 > 全部配置”。
- FusionInsight Manager界面操作：登录FusionInsight Manager，选择“集群 > 服务 > MapReduce > 配置 > 全部配置”。

步骤2 搜索“yarn.log-aggregation.retain-seconds”参数，并根据实际情况将yarn.log-aggregation.retain-seconds调小，比如调整为： 259200，即Yarn的聚合日志保留3天，到期后自动释放磁盘空间。

步骤3 保存配置，不勾选“重新启动受影响的服务或实例”。

步骤4 在业务空闲时执行该步骤重启MapReduce服务，重启服务会导致上层服务业务中断，影响集群的管理维护和业务，建议在空闲时执行。

1. 登录Manager页面。
2. 重启MapReduce服务。

----结束

15.20.4 MR 任务异常临时文件不删除

用户问题

MR任务异常临时文件为什么没有删除?

问题现象

HDFS临时目录文件过多，占用内存。

原因分析

MR任务提交时会将相关配置文件、jar包和-files添加的文件都放入hdfs上的临时目录，方便后面container启动以后获取相应的文件。由配置项yarn.app.mapreduce.am.staging-dir决定具体存放位置，默认值是/tmp/hadoop-yarn/staging。

正常运行的MR任务会在Job结束以后就清理这些临时文件，但是当Job对应的yarn任务是异常退出时，这些临时文件不会被清理，长时间积攒导致该临时目录下的文件数量越来越多，占用存储空间越来越多。

处理步骤

步骤1 登录集群。

1. 以root用户登录任意一个Master节点，用户密码为创建集群时用户自定义的密码。
2. 如果集群开启Kerberos认证，执行如下命令进入客户端安装目录并设置环境变量，再认证用户并按照提示输入密码，该密码请向管理员获取。

   cd *客户端安装目录*

   source bigdata_env

kinit hdfs

3. 如果集群未开启Kerberos认证，执行如下命令切换到omm用户，再进入客户端安装目录设置环境变量。

   su - omm

   cd *客户端安装目录*

   source bigdata_env

步骤2 获取文件列表。

hdfs dfs -ls /tmp/hadoop-yarn/staging/*/.staging/ | grep "^drwx" | awk '{print $8}' > job_file_list

job_file_list文件中就是所有任务的文件夹列表，文件内容参考：

```
/tmp/hadoop-yarn/staging/omm/.staging/job__<Timestamp>_<ID>
```

步骤3 统计当前运行中的任务。

mapred job -list 2>/dev/null | grep job_ | awk '{print $1}' > run_job_list

run_job_list文件里面就是当前正在运行的JobId列表，文件内容格式为：

```
job_<Timestamp>_<ID>
```

步骤4 删除job_file_list文件中正在运行中的任务。确保在删除过期数据时不会误删正在运行任务的数据。

cat run_job_list | while read line; do sed -i "/$line/d" job_file_list; done

步骤5 删除过期数据。

cat job_file_list | while read line; do hdfs dfs -rm -r $line; done

步骤6 清除临时文件。

rm -rf run_job_list job_file_list

----结束

15.20.5 提交任务的 Yarn 的 ResourceManager 报错 connection refused，且配置的 Yarn 端口为 8032

用户问题

请求提交任务的Yarn的ResourceManager报错connection refused，且配置的Yarn端口为8032。

问题现象

MRS的Yarn ResourceManager中的一个节点无法连接，且配置的Yarn端口为8032。

原因分析

该业务应用在集群外部运行，且使用的是老集群的客户端，配的Yarn端口是8032，与MRS服务的Yarn ResourceManager实际端口不同。从而使请求提交任务的Yarn的ResourceManager报错connection refused。

处理步骤

步骤1 更新MRS服务客户端。

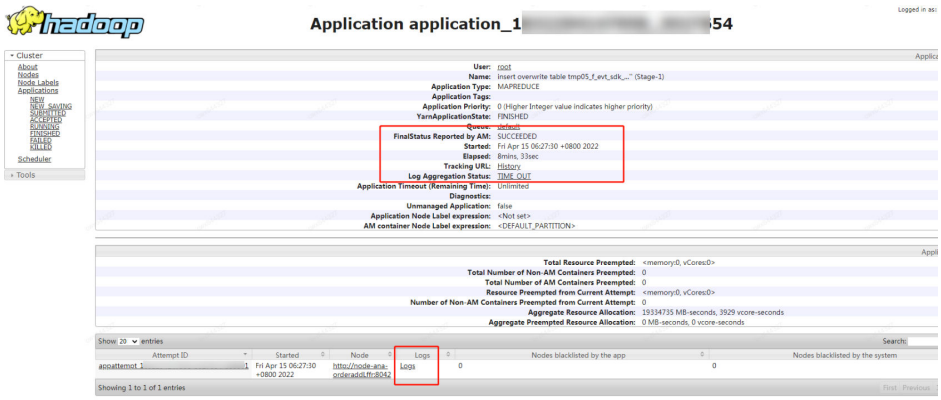
步骤2 然后重试提交作业。

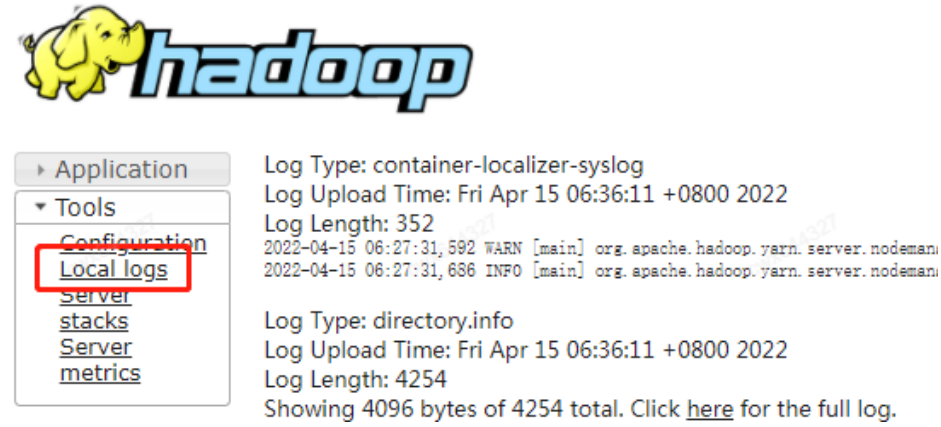
----结束

15.20.6 Yarn WebUI 作业查看日志提示“Could not access logs page!”

问题背景与现象

登录Yarn WebUI界面查看作业日志“Logs”，然后单击“Local logs”，界面提示“Could not access logs page!”





原因分析

该Local logs是用来访问服务的日志，但由于安全考虑，该功能暂不支持从Yarn WebUI界面访问。您可以登录ResourceManager主节点查看ResourceManager的日志。

处理步骤

步骤1 登录Manager界面，选择“集群 > 组件 > Yarn > 实例”，查看并获取“ResourceManager(主)”实例的业务IP地址。

步骤2 以**root**用户登录ResourceManager主节点。

步骤3 进入“/var/log/Bigdata/yarn/rm”路径查看ResourceManager的日志。

cd /var/log/Bigdata/yarn/rm

----结束

15.20.7 Yarn 页面单击队列名称报错

问题背景与现象

在Yarn使用Capacity调度器时，单击Yarn原生页面的队列名称会报500的错误。

HTTP ERROR 500 javax.servlet.ServletException: javax.servlet.ServletException: java.lang.IllegalArgumentException: Illegal character in query at index 81: https://[redacted]:20026/Yarn/ResourceManager/21/cluster/scheduler?openQueues=^default$

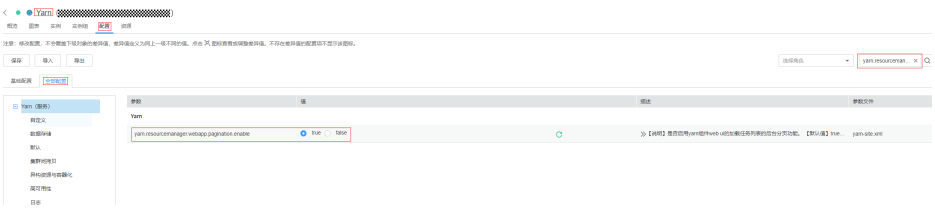
原因分析

页面链接无法识别符号“^”，导致页面访问失败。

处理步骤

步骤1 登录Manager页面，选择“集群 > 服务 > Yarn > 配置 > 全部配置”。

步骤2 在搜索框搜索“yarn.resourcemanager.webapp.pagination.enable”。



步骤3 如果该参数值为“true”（默认为“true”），请修改为“false”，并保存配置。

步骤4 在Yarn服务页面选择“实例”页签，勾选所有的ResourceManager实例，选择“更多 > 滚动重启实例”，等待实例启动完成。

----结束

15.21 使用 ZooKeeper

15.21.1 MRS 集群如何访问 ZooKeeper

用户问题

MRS集群如何访问ZooKeeper?

问题现象

客户在MRS的Master节点使用zkcli.sh访问ZooKeeper但是存在报错。

原因分析

客户使用命令有问题，造成报错的发生。

处理步骤

步骤1 获取ZooKeeper的IP地址。

步骤2 以root用户登录Master节点。

步骤3 初始化环境变量。

source /opt/client/bigdata_env

步骤4 执行**zkCli.sh -server** *ZooKeeper所在节点的IP*:**2181**即可连接上MRS的ZooKeeper。

zk所在节点的IP即为**步骤1**中查到的结果，多个IP之间以逗号间隔。

步骤5 使用**ls /**等常用的命令查看ZooKeeper上的信息。

----结束

15.22 访问 OBS

15.22.1 使用 MRS 多用户访问 OBS 功能时/tmp 目录没有权限

用户问题

在使用MRS多用户访问OBS功能，执行spark、hive、presto等作业时，出现/tmp目录没有权限的报错。

问题现象

在使用MRS多用户访问OBS功能，执行spark、hive、presto等作业时，出现/tmp目录没有权限的报错。

原因分析

作业执行过程中有临时目录，提交作业的用户对临时目录没有权限。

处理步骤

步骤1 在集群“概览”页签中，查询并记录集群所绑定的委托名称。

步骤2 登录IAM服务控制台。

步骤3 选择“权限 > 创建自定义策略”。

- 策略名称：请输入策略名称。
- 作用范围：请选择“全局级服务”。
- 策略配置方式：请选择“可视化视图”。
- 策略内容：

a. “允许”选择“允许”。
b. “云服务”选择“对象存储服务 (OBS)”。
c. “操作”勾选所有“写”、“列表”和“只读”权限。
d. “特定资源”选择：
   i. “object”选择“通过资源路径指定”，并单击“添加资源路径”分别在“路径”中输入***obs_bucket_name*/tmp/**和***obs_bucket_name*/tmp/***。此处以**/tmp**目录为例，如需其他目录权限请参考该步骤添加对应目录及该目录下所有对象的资源路径。
   ii. “bucket”选择“通过资源路径指定”，并单击“添加资源路径”在“路径”中输入***obs_bucket_name***。

   其中**obs_bucket-name**请使用实际的OBS桶名替换。若桶类型为“并行文件系统”需要再添加***obs_bucket_name*/tmp/**路径，桶类型为“对象存储”则不需要添加***obs_bucket_name*/tmp/**路径。
e. （可选）请求条件，暂不添加。

图 15-60 自定义策略

步骤4 单击“确定”完成策略添加。

步骤5 选择“委托”，并在**步骤1**中查询到的委托所在行的“操作”列单击“权限配置”。

步骤6 查询并勾选**步骤3**中创建的策略。

步骤7 单击“确定”完成委托权限配置。

----结束

15.22.2 Hadoop 客户端删除 OBS 上数据时.Trash 目录没有权限

用户问题

使用Hadoop客户端删除OBS上数据时出现.Trash目录没有权限的报错。

问题现象

执行**hadoop fs -rm obs://<obs_path>**出现如下报错：

```
exception [java.nio.file.AccessDeniedException: user/root/.Trash/Current/: getFileStatus on user/root/.Trash/Current/: status [403]
```

原因分析

hadoop删除文件时会将文件先移动到.Trash目录，若该目录没有权限则出现403报错。

处理步骤

方案一：

使用**hadoop fs -rm -skipTrash**来删除文件。

方案二：

在集群对应的委托中添加访问.Trash目录的权限。

步骤1 在集群“概览”页签中，查询并记录集群所绑定的委托名称。

步骤2 登录IAM服务控制台。

步骤3 选择“权限 > 创建自定义策略”。

- 策略名称：请输入策略名称。
- 作用范围：请选择“全局级服务”。
- 策略配置方式：请选择“可视化视图”。
- 策略内容：
  a. “允许”选择“允许”。
  b. “云服务”选择“对象存储服务 (OBS)”。
  c. “操作”勾选所有操作权限。
  d. “特定资源”选择：
     i. “object”选择“通过资源路径指定”，并单击“添加资源路径”分别在“路径”中输入.Trash目录，例如***obs_bucket_name*/user/root/.Trash/***。
     ii. “bucket”选择“通过资源路径指定”，并单击“添加资源路径”在“路径”中输入***obs_bucket_name***。

     其中**obs_bucket-name**请使用实际的OBS桶名替换。
  e. （可选）请求条件，暂不添加。

图 15-61 自定义策略

步骤4 单击“确定”完成策略添加。

步骤5 选择“委托”，并在**步骤1**中查询到的委托所在行的“操作”列单击“权限配置”。

步骤6 查询并勾选**步骤3**中创建的策略。

步骤7 单击“确定”完成委托权限配置。

步骤8 重新执行**hadoop fs -rm obs://<obs_path>**命令。

----结束

16 附录

16.1 MRS 所使用的裸金属服务器规格

针对不同的应用场景，MRS使用到如下实例的裸金属服务器。

鲲鹏 V1实例

规格命名规则

AB.C.D

例如m2.8xlarge.8

其中，

- A表示系列，例如：s表示通用型、c表示计算型、m表示内存型。
- B表示系列号，例如：s1中的1表示通用型I代，s2中2表示通用型II代。
- C表示规格，当前系列中的规格大小，例如：medium、large、xlarge。
- D表示内存、CPU比，以具体数字表示，例如4表示内存和CPU的比值为4。

规格

表 16-1 鲲鹏 V1 实例型裸金属服务器的规格

规格名称/ID	vCPU	内存（GB）	网络
physical.ks1ne.4xlarge	128	512	分布式
physical.ks1ne.8xlarge	128	1024	

16.2 数据迁移方案介绍

16.2.1 准备工作

本迁移指导将指导适用于多种不同场景下的HDFS、HBase、Hive数据向MRS集群的迁移工作。由于数据迁移过程中可能存在数据覆盖、丢失、损坏等风险，因此本指导只作为参考，具体的数据迁移方案的制定及实施请支持人员协同完成。

数据迁移前源集群的准备工作，目的是防止在数据迁移过程中源集群产生新数据，导致源集群与迁移后的目标集群数据不一致。在数据迁移完成之前，目标集群应处于初始状态，期间不能运行除数据迁移作业外的其它任何业务。

停止集群业务及相关服务

- 如果您的集群涉及到Kafka业务，请先停止所有向Kafka中生产数据的作业，等待Kafka的消费作业消费完Kafka中的存量数据后，再执行下一步操作。
- 停止所有与HDFS、HBase、Hive相关的业务和作业，然后停止HBase、Hive服务。

打通数据传输通道

- 当源集群与目标集群部署在同一区域的不同VPC时，请创建两个VPC之间的网络连接，打通网络层面的数据传输通道。请参见“虚拟私有云 VPC > 用户指南> VPC对等连接”。
- 当源集群与目标集群部署在同一VPC但属于不同安全组时，在VPC管理控制台，为每个安全组分别添加安全组规则。规则的“协议”为“ANY”，“方向”为“入方向”，“源地址”为“安全组”且是对端集群的安全组。
  - 为源集群的安全组添加入方向规则，源地址选择目标集群的安全组。
  - 为目标集群的安全组添加入方向规则，源地址选择源集群的安全组。
- 当源集群与目标集群部署在同一VPC同一安全组且两个集群都开启了Kerberos认证，需要为两个集群配置互信。

16.2.2 元数据导出

为了保持迁移后数据的属性及权限等信息在目标集群上与源集群一致，需要将源集群的元数据信息导出，以便在完成数据迁移后进行必要的元数据恢复。需要导出的元数据包括HDFS文件属主/组及权限信息、Hive表描述信息。

HDFS 元数据导出

HDFS数据需要导出的元数据信息包括文件及文件夹的权限和属主/组信息。可通过如下HDFS客户端命令导出。

```
$HADOOP_HOME/bin/hdfs dfs –ls –R <migrating_path> > /tmp/hdfs_meta.txt
```

其中，各参数的含义如下。

- *$HADOOP_HOME*：源集群Hadoop客户端安装目录
- *<migrating_path>*：HDFS上待迁移的数据目录
- /tmp/hdfs_meta.txt：导出的元数据信息保存在本地的路径。

说明

如果源集群与目标集群网络互通，且以超级管理员身份运行hadoop distcp命令进行数据拷贝，可以添加参数“-p”让distcp在拷贝数据的同时在目标集群上分别恢复相应文件的元数据信息。因此在这种场景下可直接跳过本步骤。

Hive 元数据导出

Hive表数据存储在HDFS上，表数据及表数据的元数据由HDFS统一按数据目录进行迁移。而Hive表的元数据根据集群的不同配置，可以存储在不同类型的关系型数据库中（如MySQL，PostgreSQL，Oracle等）。本指导导出的Hive表元数据即存储在关系型数据库中的Hive表的描述信息。

业界主流大数据发行版均支持Sqoop的安装，如果是自建的社区版大数据集群，可下载社区版Sqoop进行安装。借助Sqoop来解耦导出的元数据与关系型数据库的强依赖，将Hive元数据导出到HDFS上，与表数据一同迁移后进行恢复。步骤如下：

步骤1 在源集群上下载并安装Sqoop工具。请参见**http://sqoop.apache.org/**。

步骤2 下载相应关系型数据库的jdbc驱动放置到${Sqoop_Home}/lib目录。

步骤3 执行如下命令导出所有Hive元数据表。所有导出数据保存在HDFS上的/user/<*user_name*>/<*table_name*>目录。

```
$Sqoop_Home/bin/sqoop import --connect jdbc:<driver_type>://<ip>:<port>/<database> --table <table_name> --username <user> -password <passwd> -m 1
```

其中，各参数的含义如下。

- *$Sqoop_Home*：Sqoop的安装目录
- <*driver_type*>：数据库类型
- <*ip*>：源集群数据库的IP地址
- <*port*>：源集群数据库的端口号
- <*table_name*>：待导出的表名称
- <*user*>：用户名
- <*passwd*>：用户密码

----结束

16.2.3 数据拷贝

根据源集群与目标集群分别所处的区域及网络连通性，可分为以下几种数据拷贝场景：

同 Region

当源集群与目标集群处于同一Region时，根据**打通数据传输通道**进行网络配置，打通网络传输通道。使用Distcp工具执行如下命令将源集群的HDFS、HBase、Hive数据文件以及Hive元数据备份文件拷贝至目的集群。

```
$HADOOP_HOME/bin/hadoop distcp <src> <dist> -p
```

其中，各参数的含义如下。

- *$HADOOP_HOME*：目的集群Hadoop客户端安装目录
- <src>：源集群HDFS目录
- <dist>：目的集群HDFS目录

不同 Region

当源集群与目标集群处于不同Region时，用Distcp工具将源集群数据拷贝到OBS，借助OBS跨区域复制功能（请参见“对象存储服务 OBS > 控制台指南 > 跨区域复制”）

将数据复制到对应目的集群所在Region的OBS，然后通过Distcp工具将OBS数据拷贝到目的集群的HDFS上。由于执行Distcp无法为OBS上的文件设置权限、属主/组等信息，因此当前场景在进行数据导出时也需要将HDFS的元数据信息进行导出并拷贝，以防HDFS文件属性信息丢失。

线下集群向云迁移

线下集群可以通过如下方式将数据迁移至云：

- 云专线（DC）

  为源集群与目标集群之间建立云专线，打通线下集群出口网关与线上VPC之间的网络，然后参考**同Region**执行Distcp进行拷贝。

16.2.4 数据恢复

HDFS 文件属性恢复

根据导出的权限信息在目的集群的后台使用HDFS命令对文件的权限及属主/组信息进行恢复。

```
$HADOOP_HOME/bin/hdfs dfs –chmod <MODE> <path>
$HADOOP_HOME/bin/hdfs dfs –chown <OWNER> <path>
```

Hive 元数据恢复

在目的集群中安装并使用Sqoop命令将导出的Hive元数据导入MRS集群dbservice。

```
$Sqoop_Home/bin/sqoop export --connect jdbc:postgresql://<ip>:20051/hivemeta --table <table_name> --username hive -password <passwd> --export-dir <export_from>
```

其中，各参数的含义如下。

- *$Sqoop_Home*：目的集群上Sqoop的安装目录
- <*ip*>：目的集群上数据库的IP地址
- <*table_name*>：待恢复的表名称
- <*passwd*>：hive用户的密码
- <*export_from*>：元数据在目的集群的HDFS地址。

HBase 表重建

重启目的集群的HBase服务，使数据迁移生效。在启动过程中，HBase会加载当前HDFS上的数据并重新生成元数据。启动完成后，在Master节点客户端执行如下命令加载HBase表数据。

```
$HBase_Home/bin/hbase hbck -fixMeta -fixAssignments
```

命令执行完成后，重复执行如下命令查看HBase集群健康状态直至正常。

```
hbase hbck
```

16.3 MRS 3.x 版本操作注意事项

概述

MRS 3.x之前的版本的MRS集群使用MRS Manager对集群进行管理、监控，同时用户可通过MRS管理控制台的集群管理页面，进行集群概览查看、节点管理、组件管理、告警管理、补丁管理、文件管理、作业管理、租户管理、备份恢复、引导操作设置及标签管理。

MRS 3.x版本的MRS集群使用FusionInsight Manager对集群进行管理、监控，同时用户可通过MRS管理控制台的集群管理页面，进行集群概览查看、节点管理、组件管理、告警管理、文件管理、作业管理、引导操作设置及标签管理。

MRS 3.x版本集群的部分维护操作与历史版本有部分差异，更多详细操作可参考本**MRS Manager操作指导（适用于2.x及之前）**与**FusionInsight Manager操作指导（适用于3.x）**。

访问 MRS 集群 Manager

- 访问MRS 3.x之前的版本的MRS Manager请参考**访问MRS Manager（MRS 2.x及之前版本）**。
- 访问MRS 3.x版本的FusionInsight Manager请参考**访问FusionInsight Manager（MRS 3.x及之后版本）**。

修改 MRS 集群服务配置参数

- MRS 3.x之前的版本，用户可直接通过MRS管理控制台的集群管理页面修改各服务配置参数：
  a. 登录MRS控制台，在左侧导航栏选择“集群列表> 现有集群”，单击集群名称。
  b. 选择“组件管理 > *服务名称* > 服务配置”。
     默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。
  c. 在导航树选择指定的参数分类，并在右侧修改参数值。
     不确定参数的具体位置时，支持在右上角输入参数名，系统将实时进行搜索并显示结果。
  d. 单击“保存配置”，并在确认对话框中单击“确定”。
  e. 等待界面提示“操作成功”，单击“完成”，配置已修改。
     查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。也可在保存配置时直接勾选“重新启动受影响的服务或实例。”。
- MRS 3.x版本，服务配置参数需登录FusionInsight Manager修改：
  a. 登录FusionInsight Manager。
  b. 选择“集群 > 服务”。
  c. 单击服务视图中指定的服务名称。
  d. 单击“配置”。

默认显示“基础配置”，如果需要修改更多参数，请选择“全部配置”，界面上将显示该服务的全部配置参数导航树，导航树从上到下的一级节点分别为服务名称和角色名称。展开一级节点后显示参数分类。

e. 在导航树选择指定的参数分类，并在右侧修改参数值。

不确定参数的具体位置时，支持在右上角输入参数名，Manager将实时进行搜索并显示结果。

f. 单击“保存”，并在确认对话框中单击“确定”。

g. 等待界面提示“操作成功”，单击“完成”，配置已修改。

查看集群是否存在配置过期的服务，如果存在，需重启对应服务或角色实例使配置生效。

16.4 安装 Flume 客户端

16.4.1 安装 MRS 3.x 之前版本 Flume 客户端

操作场景

使用Flume搜集日志时，需要在日志主机上安装Flume客户端。用户可以创建一个新的ECS并安装Flume客户端。

本章节适用于MRS 3.x之前版本。

前提条件

- 已创建包含Flume组件的流集群。
- 日志主机需要与MRS集群在相同的VPC和子网。
- 已获取日志主机的登录方式。

操作步骤

步骤1 根据前提条件，创建一个满足要求的弹性云服务器。

步骤2 登录集群详情页面。

- MRS 2.0.1之前版本，登录MRS Manager，选择“服务管理”。
- MRS 2.0.1及之后版本，在MRS控制台单击集群名称，选择“组件管理 ”。

步骤3 单击“下载客户端”。

1. 在“客户端类型”选择“完整客户端”。
2. 在“下载路径”选择“远端主机”。
3. 将“主机IP”设置为ECS的IP地址，设置“主机端口”为“22”，并将“保存路径”设置为。
   - 如果使用SSH登录ECS的默认端口“22”被修改，请将“主机端口”设置为新端口。
   - “保存路径”最多可以包含256个字符。
4. “登录用户”设置为“root”。

   如果使用其他用户，请确保该用户对保存目录拥有读取、写入和执行权限。

5. 单击“确定”开始生成客户端文件。

   若界面显示以下提示信息表示客户端包已经成功保存。

   ```
   下载客户端文件到远端主机成功。
   ```

   若界面显示以下提示信息，请检查用户名密码及远端主机的安全组配置，确保用户名密码正确，及远端主机的安全组已增加SSH(22)端口的入方向规则。然后从**步骤3**执行重新开始下载客户端。

   ```
   连接到服务器失败，请检查网络连接或参数设置。
   ```

步骤4 选择“Flume”服务，单击“实例”，查看任意一个Flume实例和两个MonitorServer实例的“业务IP”。

步骤5 使用VNC方式，登录弹性云服务器。参见)。

请先参考**SSH密钥方式登录**登录弹性云服务器，设置云服务器密码，然后再采用VNC方式登录。

步骤6 在弹性云服务器，切换到**root**用户，并将安装包复制到目录“/opt”。

sudo su - root

cp /MRS_Flume_Client.tar /opt

步骤7 在“/opt”目录执行以下命令，解压压缩包获取校验文件与客户端配置包。

tar -xvf MRS_Flume_Client.tar

步骤8 执行以下命令，校验文件包。

sha256sum -c MRS_Flume_ClientConfig.tar.sha256

界面显示如下信息，表明文件包校验成功：

```
MRS_Flume_ClientConfig.tar: OK
```

步骤9 执行以下命令，解压“MRS_Flume_ClientConfig.tar”。

tar -xvf MRS_Flume_ClientConfig.tar

步骤10 执行以下命令，安装客户端运行环境到新的目录，例如“/opt/Flumeenv”。安装时自动生成目录。

sh /opt/MRS_Flume_ClientConfig/install.sh /opt/Flumeenv

查看安装输出信息，如有以下结果表示客户端运行环境安装成功：

```
Components client installation is complete.
```

步骤11 执行以下命令，配置环境变量。

source /opt/Flumeenv/bigdata_env

步骤12 执行以下命令，解压Flume客户端。

cd /opt/MRS_Flume_ClientConfig/Flume

tar -xvf FusionInsight-Flume-1.6.0.tar.gz

步骤13 执行以下命令，查看当前用户密码是否过期。

chage -l root

“Password expires”时间早于当前则表示过期。此时需要修改密码，或执行**chage -M -1 root**设置密码为未过期状态。

步骤14 执行以下命令，安装Flume客户端到新目录，例如“/opt/FlumeClient”。安装时自动生成目录。

sh /opt/MRS_Flume_ClientConfig/Flume/install.sh -d /opt/FlumeClient -f *MonitorServer实例的业务IP地址* -c Flume配置文件路径 -l /var/log/ -e *Flume的业务IP地址* -n *Flume客户端名称*

各参数说明如下：

- “-d”：表示Flume客户端安装路径。
- “-f”：可选参数，表示两个MonitorServer角色的业务IP地址，中间用英文逗号分隔，若不设置则Flume客户端将不向MonitorServer发送告警信息，同时在MRS Manager界面上看不到该客户端的相关信息。
- “-c”：可选参数，表示Flume客户端在安装后默认加载的配置文件“properties.properties”。如不添加参数，默认使用客户端安装目录的“fusioninsight-flume-1.6.0/conf/properties.properties”。客户端中配置文件为空白模板，根据业务需要修改后Flume客户端将自动加载。
- “-l”：可选参数，表示日志目录，默认值为“/var/log/Bigdata”。
- “-e”：可选参数，表示Flume实例的业务IP地址，主要用于接收客户端上报的监控指标信息。
- “-n”：可选参数，表示自定义的Flume客户端的名称。
- IBM的JDK不支持“-Xloggc”，需要修改“flume/conf/flume-env.sh”，将“-Xloggc”修改为“-Xverbosegclog”，若JDK为32位，“-Xmx”不能大于3.25GB。
- “flume/conf/flume-env.sh”中，“-Xmx”默认为4GB。若客户端机器内存过小，可调整为512M甚至1GB。

例如执行：**sh install.sh -d /opt/FlumeClient**

系统显示以下结果表示客户端运行环境安装成功：

```
install flume client successfully.
```

----结束

16.4.2 安装 MRS 3.x 及之后版本 Flume 客户端

操作场景

使用Flume搜集日志时，需要在日志主机上安装Flume客户端。用户可以创建一个新的ECS并安装Flume客户端。

本章节适用于MRS 3.x及之后版本。

前提条件

- 已创建包含Flume组件的集群。
- 日志主机需要与MRS集群在相同的VPC和子网。
- 已获取日志主机的登录方式。

- 安装目录可以不存在，会自动创建。但如果存在，则必须为空。目录路径不能包含空格。

操作步骤

步骤1 获取软件包。

登录FusionInsight Manager，选择“集群 > *待操作集群的名称* > 服务 > Flume”进入Flume服务界面，在右上角选择“更多 > 下载客户端”，选择“选择客户端类型”为“完整客户端”，下载Flume服务客户端文件。

客户端文件名称为“FusionInsight_Cluster_*<集群ID>*_Flume_Client.tar”，本章节以“FusionInsight_Cluster_1_Flume_Client.tar”为例进行描述。

步骤2 上传软件包。

以**user**用户将软件包上传到将要安装Flume服务客户端的节点目录上，例如“/opt/client”。

说明

user用户为安装和运行Flume客户端的用户。

步骤3 解压软件包。

以**user**用户登录将要安装Flume服务客户端的节点。进入安装包所在目录，例如“/opt/client”，执行如下命令解压安装包到当前目录。

cd /opt/client

tar -xvf FusionInsight_Cluster_1_Flume_Client.tar

步骤4 校验软件包。

执行**sha256sum -c**命令校验解压得到的文件，返回“OK”表示校验通过。例如：

sha256sum -c FusionInsight_Cluster_1_Flume_ClientConfig.tar.sha256

```
FusionInsight_Cluster_1_Flume_ClientConfig.tar: OK
```

步骤5 解压文件。

tar -xvf FusionInsight_Cluster_1_Flume_ClientConfig.tar

步骤6 在Flume客户端安装目录下执行以下命令，安装客户端到指定目录（绝对路径），例如安装到“/opt/FlumeClient”目录。客户端安装成功后安装结束。

cd /opt/client/FusionInsight_Cluster_1_Flume_ClientConfig/Flume/FlumeClient

./install.sh -d */opt/FlumeClient* **-f** *MonitorServer角色的业务IP或主机名* **-c** *用户业务配置文件properties.properties放置路径* **-s** *cpu阈值* **-l /var/log/Bigdata -e** *FlumeServer的业务IP或主机名* **-n** *Flume*

说明

- “-d”：Flume客户端安装路径。
- “-f”（可选）：两个MonitorServer角色的业务IP或主机名，中间用逗号分隔，若不设置则Flume客户端将不向MonitorServer发送告警信息，同时在FusionInsight Manager界面上看不到该客户端的相关信息。
- “-c”（可选）：指定业务配置文件，该文件需要用户根据自己业务生成，具体操作可在Flume服务端中“配置工具”页面生成，并上传到待安装客户端节点上的任一目录下。若安装时未指定（即不配置该参数），可在安装后上传已经生成的业务配置文件properties.properties到“/opt/FlumeClient/fusioninsight-flume-1.9.0/conf”目录下。
- “-s”（可选）：Cgroup阈值，阈值取值范围为1~100*N之间的整数，N表示机器cpu核数。默认阈值为“-1”，表示加入到Cgroup的进程不受cpu使用率限制。
- “-l”（可选）：日志路径，默认值为“/var/log/Bigdata”（“user”用户需要对此目录有写权限）。首次安装客户端会生成名为flume-client的子目录，之后安装会依次生成名为“flume-client-*n*”的子目录，n代表一个序号，从1依次递增。在Flume客户端安装目录下的conf目录中，编辑ENV_VARS文件，搜索FLUME_LOG_DIR属性，可查看客户端日志路径。
- “-e”（可选）：FlumeServer的业务IP地址或主机名，主要用于接收客户端上报的监控指标信息。
- “-n”（可选）：Flume客户端的名称，可以通过在FusionInsight Manager上选择“集群 > *待操作集群名称* > 服务 > Flume > Flume管理”查看对应节点上客户端的名称。
- 若产生以下错误提示，可执行命令**export JAVA_HOME=***JDK路径*进行处理。
  ```
  JAVA_HOME is null in current user,please install the JDK and set the JAVA_HOME
  ```
- IBM的JDK不支持“-Xloggc”，需要修改“flume/conf/flume-env.sh”，将“-Xloggc”修改为“-Xverbosegclog”，若JDK为32位，“-Xmx”不能大于3.25GB。
- 集群混搭时，安装跨平台客户端时，请进入/opt/client/FusionInsight_Cluster_1_Flume_ClientConfig/Flume/FusionInsight-Flume-1.9.0.tar.gz路径下进行Flume客户端安装。

----结束

17 修订记录

发布日期	更新特性
2023-04-04	第八次正式发布。 上线MRS 3.2.0-LTS.1版本，详见**创建自定义集群**。
2022-11-30	第七次正式发布。 • 用户账号一览表适配3.1.2-LTS.3版本，详见**用户账号一览表**。 • 新增**修改DBService数据库compdbuser用户密码**章节。
2022-07-30	第六次正式发布。 上线MRS 3.1.2-LTS.3版本，详见**创建自定义集群**。
2021-06-30	第五次正式发布。 上线MRS 3.1.0-LTS.1版本，详见**创建自定义集群**。
2020-10-24	第四次正式发布。 新增如下章节： • **创建方式简介** • **快速创建Hadoop分析集群** • **快速创建HBase查询集群** • **快速创建Kafka流式集群** 修改如下章节： • **创建自定义集群** • **查看集群基本信息** • **查看历史集群信息**
2020-01-17	第三次正式发布。 修改如下章节： • **创建自定义集群** • **查看集群基本信息** • **查看历史集群信息**

发布日期	更新特性
2019-03-20	第二次正式发布。 新增如下章节： • **集群（未启用Kerberos认证）安全配置建议** • **运维授权** • **运维授权** • **添加集群标签** • **通过引导操作安装第三方软件** • **修复隔离主机补丁** • **支持滚动重启** 修改如下章节： • **创建自定义集群** • **查看集群基本信息** • **配置弹性伸缩规则** • **查看历史集群信息**
2018-10-12	第一次正式发布。