

数据治理中心

常见问题

文档版本 01
发布日期 2024-04-29



版权所有 © 华为云计算技术有限公司 2024。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目录

1 咨询与计费	1
1.1 区域和可用区	1
1.2 数据库、数据仓库、数据湖与华为智能数据湖方案	2
1.3 DataArts Studio 和沃土是什么关系?	5
1.4 DataArts Studio 和 ROMA 是什么关系?	5
1.5 DataArts Studio 是否支持私有化部署到本地或私有云?	5
1.6 如何在 IAM 中创建细粒度权限策略?	5
1.7 用户已添加权限, 还是无法查看工作空间?	6
1.8 IAM 用户操作时报错“无 xx 权限”怎么办?	7
1.9 DataArts Studio 的工作空间可以删除吗?	8
1.10 实例试用/购买成功后, 可以转移到其他账号下吗?	9
1.11 DataArts Studio 是否支持版本升级?	9
1.12 DataArts Studio 是否支持版本降级?	9
1.13 如何查看 DataArts Studio 的版本?	9
1.14 购买 DataArts Studio 实例, 选不到指定的 IAM 项目下面, 怎么办?	10
1.15 DataArts Studio 的会话超时时间是多少, 是否支持修改?	10
1.16 套餐包到期未续订或按需资源欠费时, 我的数据会保留吗?	11
1.17 如何查看套餐包的剩余时长?	11
1.18 DataArts Studio 实例中的 CDM 没有计费是什么原因?	11
1.19 为什么会提示每日执行节点个数超过上限, 应该怎么处理?	11
2 管理中心	13
2.1 DataArts Studio 支持治理哪些数据湖?	13
2.2 创建数据连接需要注意哪些事项?	13
2.3 为什么 DWS/Hive/HBase 数据连接突然无法获取数据库或表的信息?	13
2.4 为什么在创建数据连接的界面上 MRS Hive/HBase 集群不显示?	14
2.5 创建 DWS 数据连接, 开启 SSL 连接时测试连接失败?	14
2.6 通过代理方式创建数据连接, 一个空间可以创建多个连接吗?	14
2.7 创建 DWS 连接的时候, 连接方式是直接连还是通过代理连比较好?	14
2.8 如何将一个空间的数据开发作业和数据连接迁移到另一空间?	15
3 数据集成	16
3.1 CDM 与其他数据迁移服务有什么区别, 如何选择?	16
3.2 CDM 有哪些优势?	19

3.3 CDM 有哪些安全防护?	20
3.4 如何降低 CDM 使用成本?	20
3.5 CDM 未使用数据传输功能时, 是否会计费?	21
3.6 已购买包年包月的 CDM 套餐包, 为什么还会产生按需计费的费用?	21
3.7 如何查看套餐包的剩余时长?	21
3.8 CDM 可以跨账户使用吗?	21
3.9 CDM 集群是否支持升级操作?	22
3.10 CDM 迁移性能如何?	22
3.11 CDM 不同集群规格对应并发的作业数是多少?	22
3.12 是否支持增量迁移?	24
3.13 是否支持字段转换?	24
3.14 Hadoop 类型的数据源进行数据迁移时, 建议使用的组件版本有哪些?	31
3.15 数据源为 Hive 时支持哪些数据格式?	31
3.16 是否支持同步作业到其他集群?	31
3.17 是否支持批量创建作业?	32
3.18 是否支持批量调度作业?	32
3.19 如何备份 CDM 作业?	32
3.20 如果 HANA 集群只有部分节点和 CDM 集群网络互通, 应该如何配置连接?	32
3.21 如何使用 Java 调用 CDM 的 Rest API 创建数据迁移作业?	32
3.22 如何将云下内网或第三方云上的私网与 CDM 连通?	38
3.23 CDM 是否支持参数或者变量?	40
3.24 CDM 迁移作业的抽取并发数应该如何设置?	40
3.25 CDM 是否支持动态数据实时迁移功能?	42
3.26 CDM 是否支持集群关机功能?	42
3.27 如何使用表达式方式获取当前时间?	42
3.28 日志提示解析日期格式失败时怎么处理?	42
3.29 字段映射界面无法显示所有列怎么处理?	45
3.30 CDM 迁移数据到 DWS 时如何选取分布列?	48
3.31 迁移到 DWS 时出现 value too long for type character varying 怎么处理?	49
3.32 OBS 导入数据到 SQL Server 时出现 Unable to execute the SQL statement 怎么处理?	51
3.33 获取集群列表为空/没有权限访问/操作时报当前策略不允许执行?	51
3.34 Oracle 迁移到 DWS 报错 ORA-01555.....	52
3.35 MongoDB 连接迁移失败时如何处理?	53
3.36 Hive 迁移作业长时间卡住怎么办?	53
3.37 使用 CDM 迁移数据由于字段类型映射不匹配导致报错怎么处理?	53
3.38 MySQL 迁移时报错“JDBC 连接超时”怎么办?	54
3.39 创建了 Hive 到 DWS 类型的连接, 进行 CDM 传输任务失败时如何处理?	55
3.40 如何使用 CDM 服务将 MySQL 的数据导出成 SQL 文件, 然后上传到 OBS 桶?.....	55
3.41 如何处理 CDM 从 OBS 迁移数据到 DLI 出现迁移中断失败的问题?	55
3.42 如何处理 CDM 连接器报错“配置项 [linkConfig.iamAuth] 不存在”?	56
3.43 创建数据连接时报错“配置项[linkConfig.createBackendLinks]不存在”或创建作业时报错“配置项 [throttlingConfig.concurrentSubJobs] 不存在”怎么办?.....	56
3.44 新建 MRS Hive 连接时, 提示: CORE_0031:Connect time out. (Cdm.0523) 怎么解决?	56

3.45 迁移时已选择表不存在时自动创表，提示“CDM not support auto create empty table with no column”怎么处理？	56
3.46 创建 Oracle 关系型数据库迁移作业时，无法获取模式名怎么处理？	57
3.47 MySQL 迁移时报错：invalid input syntax for integer: "true".....	57
4 数据架构.....	58
4.1 码表和数据标准有什么关系？	58
4.2 关系建模和维度建模的区别？	58
4.3 数据架构支持哪些数据建模方法？	58
4.4 规范化的数据如何使用？	59
4.5 数据架构支持逆向数据库吗？	59
4.6 数据架构中的指标与数据质量的指标的区别？	59
4.7 为什么关系建模或维度建模修改字段后，数据库中表无变化？	59
4.8 表是否可配置生命周期管理？	60
5 数据开发.....	61
5.1 数据开发可以创建多少个作业，作业中的节点数是否有限制？	61
5.2 DataArts Studio 支持自定义的 Python 脚本吗？	61
5.3 作业关联的 CDM 集群删除后，如何快速修复？	61
5.4 作业的计划时间和开始时间相差大，是什么原因？	61
5.5 相互依赖的几个作业，调度过程中某个作业执行失败，是否会影响后续作业？这时该如何处理？	62
5.6 通过 DataArts Studio 调度大数据服务时需要注意什么？	62
5.7 环境变量、作业参数、脚本参数有什么区别和联系？	62
5.8 打不开作业日志，返回 404 报错?.....	64
5.9 配置委托时获取委托列表失败如何处理？	65
5.10 数据开发创建数据连接，为什么选不到指定的周边资源？	66
5.11 配置了 SMN 通知，却收不到作业失败告警通知？	66
5.12 作业配置了周期调度，但是实例监控没有作业运行调度记录？	67
5.13 Hive SQL 和 Spark SQL 脚本执行失败，界面只显示执行失败，没有显示具体的错误原因？	67
5.14 数据开发节点运行中报 TOKEN 不合法？	67
5.15 作业开发时，测试运行后如何查看运行日志？	67
5.16 月周期的作业依赖天周期的作业，为什么天周期作业还未跑完，月周期的作业已经开始运行？	67
5.17 执行 DLI 脚本，报 Invalid authentication 怎么办？	68
5.18 创建数据连接时，在代理模式下为什么选不到需要的 CDM 集群？	68
5.19 作业配置了每日调度，但是实例没有作业运行调度记录？	68
5.20 查看作业日志，但是日志中没有内容？	69
5.21 创建了 2 个作业，但是为什么无法建立依赖关系？	69
5.22 DataArts Studio 执行调度时报错：提示作业没有可以提交的版本怎么办？	69
5.23 DataArts Studio 执行调度时报错：作业中节点 XXX 关联的脚本没有提交的版本？	70
5.24 提交调度后的作业执行失败，报 depend job [XXX] is not running or pause 怎么办？	70
5.25 如何创建数据库和数据表，数据库对应的是不是数据连接?.....	71
5.26 为什么执行完 HIVE 任务什么结果都不显示？	71
5.27 在作业监控页面里的“上次实例状态”只有运行成功、运行失败，这是为什么？	71
5.28 如何创建通知配置对全量作业都进行结果监控？	71

5.29 数据开发的并行执行节点数是多少?	71
5.30 DataArts Studio 是否支持修改时区?	73
5.31 CDM 作业改名后, 在数据开发中如何同步?	73
5.32 执行 RDS SQL, 报错 hll 不存在, 在 DataArts Studio 可以执行成功?	74
5.33 创建 DWS 数据连接时报错提示: The account has been locaked?	74
5.34 作业实例取消了, 日志提示: The node start execute failed, so the current node status is set to cancel.....	74
5.35 调用数据开发接口报错, Workspace does not exists?.....	74
5.36 Postman 调用接口返回结果正常, 为什么测试环境调用接口的 URL 参数不生效?	74
5.37 执行 Python 脚本报错: Agent need to be updated?	74
5.38 节点状态为成功, 为什么日志显示运行失败?	74
5.39 调用数据开发 API 报错 Unknown Exception?	74
5.40 调用创建资源的 API 报错“资源名不合法”是什么原因?	75
5.41 补数据的作业实例都是成功的, 为什么补数据任务失败了?	75
5.42 DWS 数据连接可视化建表, 报错提示“表已存在”, 但是展开数据连接看不到该表?	75
5.43 调度 MRS spark 作业报错 The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute.....	75
5.44 执行 Python 脚本, 报错 UnicodeEncodeError : 'ascii' codec cant encode characters in position 63-64 : ordinal not in range (128).....	76
5.45 查看日志时, 系统提示“OBS 日志文件不存在, 请检查文件是否被删除或者没有 OBS 写入权限。”怎么办?	77
5.46 Shell/Python 节点执行失败, 后台报错 session is down.....	79
5.47 请求头中参数值长度超过 512 个字符时, 何如处理?	80
5.48 执行 DWS SQL 脚本时, 提示 id 不存在, 如何处理?	82
5.49 如何查看 CDM 作业被哪些作业进行调用?	82
5.50 使用 python 调用执行脚本的 api 报错: The request parameter invalid, 如何处理?	83
6 数据质量.....	85
6.1 质量作业和对账作业有什么区别?	85
6.2 如何确认质量作业或对账作业已经阻塞?	85
6.3 如何手工重启阻塞的质量作业或对账作业?	85
6.4 怎样查看质量规则模板关联的作业?	86
6.5 用户在执行质量作业时提示无 MRS 权限怎么办?	86
7 数据目录.....	89
7.1 数据目录组件有什么用?	89
7.2 数据目录支持采集哪些对象的资产?	89
7.3 什么是数据血缘关系?	89
7.4 数据目录如何可视化展示数据血缘?	90
8 数据服务.....	91
8.1 数据服务 SDK 支持的语言?	91
8.2 创建 API 时提示代理调用失败, 怎么办?	91
8.3 数据服务 API 接口, 访问“测试 APP”, 填写了相关参数, 但是后台报错要怎么处理?	91
8.4 使用 API 时报错, 请问有什么办法可以解决?	91

8.5 API 传参是否支持传递操作符?	91
8.6 数据服务专享版提供的 API 配额已满怎么解决?	91
8.7 数据服务专享版发布的 API 如何绑定公网和域名?	92
8.8 如何处理 API 对应的数据表数据量较大时, 获取数据总条数比较耗时的问题?	92

1 咨询与计费

1.1 区域和可用区

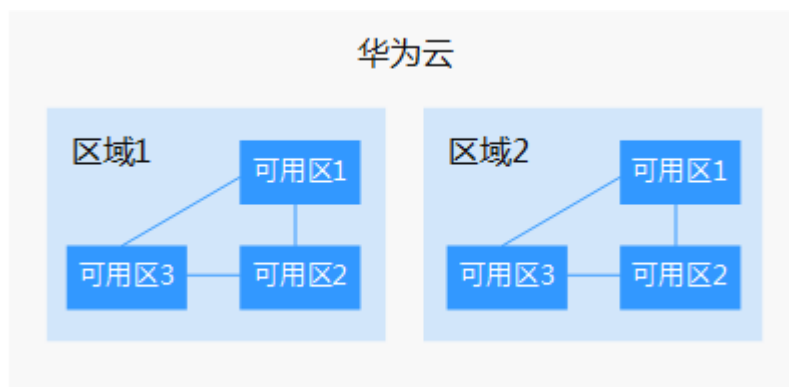
什么是区域、可用区？

我们用区域和可用区来描述数据中心的位置，您可以在特定的区域、可用区创建资源。

- 区域（Region）：从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
- 可用区（AZ，Availability Zone）是同一区域内，电力和网络互相隔离的物理区域，一个可用区不受其他可用区故障的影响。一个区域内可以有多个可用区，不同可用区之间物理隔离，但内网互通，既保障了可用区的独立性，又提供了低价、低时延的网络连接。

图1-1阐明了区域和可用区之间的关系。

图 1-1 区域和可用区



目前，华为云已在全球多个地域开放云服务，您可以根据需求选择适合自己的区域和可用区。更多信息请参见[华为云全球站点](#)。

如何选择区域？

选择区域时，您需要考虑以下几个因素：

- 地理位置

一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。不过，在基础设施、BGP网络品质、资源的操作与配置等方面，中国大陆各个区域间区别不大，如果您或者您的目标用户在中国大陆，可以不用考虑不同区域造成的网络时延问题。

曼谷等其他地区和国家提供国际带宽，主要面向非中国大陆地区的用户。如果您或者您的目标用户在中国大陆，使用这些区域会有较长的访问时延，不建议使用。

- 云服务之间的关系

如果多个云服务一起搭配使用，需要注意不同区域的云服务内网不互通。

例如DataArts Studio（包括管理中心、CDM等组件）需要与MRS、OBS等服务互通时，如果DataArts Studio与其他云服务处于不同区域的情况下，需要通过公网或者专线打通网络；而在同区域情况下，同子网、同安全组的不同实例默认网络互通。

- 资源的价格

不同区域的资源价格可能有差异，请参见[华为云服务价格详情](#)。

如何选择可用区？

DataArts Studio实例中的数据集成CDM集群所在可用区。DataArts Studio实例通过数据集成CDM集群与其他服务实现网络互通。

第一次购买DataArts Studio实例或增量包时，可用区无要求。再次购买DataArts Studio实例或增量包时，是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。

- 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。

实例可以转移到另一个区域/可用区吗？

- DataArts Studio服务套餐生效期间，可以根据需要，灵活退订已购买区域的数据集成CDM包年包月套餐，然后在新区域重新购买。支持[退订使用中的资源](#)。
- 实例购买成功后，无法转移到另一个区域/可用区。

区域和终端节点

终端节点（Endpoint）即调用API的[请求地址](#)，不同服务不同区域的终端节点不同。本服务的Endpoint可从[终端节点Endpoint](#)获取。

1.2 数据库、数据仓库、数据湖与华为智能数据湖方案

如今随着互联网以及物联网等技术的不断发展，越来越多的数据被生产出来，数据管理工具也得到了飞速的发展，大数据相关概念如雨后春笋一般应运而生，如从数据

库、数据仓库、数据湖、湖仓一体等。这些概念分别指的是什么，又有着怎样的联系，同时，华为对应的产品与方案又是什么呢？本文将一一进行对比介绍。

什么是数据库？

数据库是“按照数据结构来组织、存储和管理数据的仓库”。

广义上的数据库，在20世纪60年代已经在计算机中应用了。但这个阶段的数据库结构主要是层次或网状的，且数据和程序之间具备非常强的依赖性，应用较为有限。

现在通常所说的数据库指的是关系型数据库。关系数据库是指采用了关系模型来组织数据的数据库，其以行和列的形式存储数据，具有结构化程度高，独立性强，冗余度低等优点。1970年关系型数据库的诞生，真正彻底把软件中的数据和程序分离开来，成为主流计算机系统不可或缺的组成部分。关系型数据库已经成为目前数据库产品中最重要的一员，几乎所有的数据库厂商新出的数据库产品都支持关系型数据库，即使一些非关系数据库产品也几乎都有支持关系数据库的接口。

关系型数据库的主要用于联机事务处理OLTP（On-Line Transaction Processing）主要进行基本的、日常的事务处理，例如银行交易等场景。

什么是数据仓库？

随着数据库的大规模应用，使信息行业的数据爆炸式的增长。为了研究数据之间的关系，挖掘数据隐藏的价值，人们越来越多的需要使用联机分析处理OLAP（On-Line Analytical Processing）进行数据分析，探究一些深层次的关系和信息。但是不同的数据库之间很难做到数据共享，数据之间的集成与分析也存在非常大的挑战。

为解决企业的系统集成与分析问题，数据仓库之父比尔·恩门于1990年提出数据仓库（Data Warehouse）。数据仓库主要功能是将OLTP经年累月所累积的大量数据，通过数据仓库特有的数据储存架构进行OLAP，最终帮助决策者能快速有效地从大量数据中，分析出有价值的信息，提供决策支持。自从数据仓库出现之后，信息产业就开始从以关系型数据库为基础的运营式系统慢慢向决策支持系统发展。

数据仓库相比数据库，主要有以下两个特点：

- 数据仓库是面向主题集成的。数据仓库是为了支撑各种业务而建立的，数据来自于分散的操作型数据。因此需要将所需数据从多个异构的数据源中抽取出来，进行加工与集成，按照主题进行重组，最终进入数据仓库。
- 数据仓库主要用于支撑企业决策分析，所涉及的数据操作主要是数据查询。因此数据仓库通过表结构优化、存储方式优化等方式提高查询速度、降低开销。

表 1-1 数据仓库与数据库的对比

维度	数据仓库	数据库
应用场景	OLAP	OLTP
数据来源	多数据源	单数据源
数据标准化	非标准化Schema	高度标准化的静态Schema
数据读取优势	针对读操作进行优化	针对写操作进行优化

什么是数据湖？

在企业内部，数据是一类重要资产已经成为了共识。随着企业的持续发展，数据不断堆积，企业希望把生产经营中的所有相关数据都完整保存下来，进行有效管理与集中治理，挖掘和探索数据价值。

数据湖就是在这种背景下产生的。数据湖是一个集中存储各类结构化和非结构化数据的大型数据仓库，它可以存储来自多个数据源、多种数据类型的原始数据，数据无需经过结构化处理，就可以进行存取、处理、分析和传输。数据湖能帮助企业快速完成异构数据源的联邦分析、挖掘和探索数据价值。

数据湖的本质，是由“数据存储架构+数据处理工具”组成的解决方案。

- 数据存储架构：要有足够的扩展性和可靠性，可以存储海量的任意类型的数据，包括结构化、半结构化和非结构化数据。
- 数据处理工具，则分为两大类：
 - 第一类工具，聚焦如何把数据“搬到”湖里。包括定义数据源、制定数据同步策略、移动数据、编制数据目录等。
 - 第二类工具，关注如何对湖中的数据进行分析、挖掘、利用。数据湖需要具备完善的数据管理能力、多样化的数据分析能力、全面的数据生命周期管理能力、安全的数据获取和数据发布能力。如果没有这些数据治理工具，元数据缺失，湖里的数据质量就没法保障，最终会由数据湖变质为数据沼泽。

随着大数据和AI的发展，数据湖中数据的价值逐渐水涨船高，价值被重新定义。数据湖能给企业带来多种能力，例如实现数据的集中式管理，帮助企业构建更多优化后的运营模型，也能为企业提供其他能力，如预测分析、推荐模型等，这些模型能刺激企业能力的后续增长。

对于数据仓库与数据湖的不同之处，可以类比为仓库和湖泊的区别：仓库存储着来自特定来源的货物；而湖泊的水来自河流、溪流和其他来源，并且是原始数据。

表 1-2 数据湖与数据仓库的对比

维度	数据湖	数据仓库
应用场景	可以探索性分析 所有类型的数据 ，包括机器学习、数据发现、特征分析、预测等	通过历史的 结构化数据 进行数据分析
使用成本	起步成本低，后期成本较高	起步成本高，后期成本较低
数据质量	包含大量原始数据，使用前需要清洗和标准化处理	质量高，可作为事实依据
适用对象	数据科学家、数据开发人员为主	业务分析师为主

华为智能数据湖方案

华为数据使能服务DAYU，为大型政企客户量身定制跨越孤立系统、感知业务的数据资源智能管理解决方案，实现全域数据入湖，帮助政企客户从多角度、多层次、多粒度挖掘数据价值，实现数据驱动的数字化转型。

DAYU的核心主要是华为智能数据湖FusionInsight，包含数据库、数据仓库、数据湖等各计算引擎和数据治理中心DataArts Studio平台，提供了数据使能的全套能力，支持数据的采集、汇聚、计算、资产管理、数据开放服务的全生命周期管理。

华为FusionInsight解决方案，对应的各服务如下：

- 数据库：
 - 关系型数据库包括：[云数据库RDS](#)、[云数据库 GaussDB\(for MySQL\)](#)、[云数据库 GaussDB](#)、[云数据库 PostgreSQL](#)等。
 - 非关系型数据库包括：[文档数据库服务DDS](#)、[云数据库 GeminiDB](#)（兼容Influx、Redis、Mongo以及Cassandra多种协议）等。
- 数据仓库：[数据仓库服务DWS](#)。
- 数据湖：[云原生大数据MRS](#)、[数据湖探索DLI](#)等。
- 数据治理平台：[数据治理中心DataArts Studio](#)。

1.3 DataArts Studio 和沃土是什么关系？

DataArts Studio作为沃土平台数据使能模块，帮助接入沃土数字平台的企业更好的管理使用数据。

1.4 DataArts Studio 和 ROMA 是什么关系？

ROMA作为连接各个系统的管道，对接入数据没有治理和规划的功能，DataArts Studio对接入数据进行结构分析，重新建模，最终打破数据孤岛，帮助企业建立统一数据模型。

1.5 DataArts Studio 是否支持私有化部署到本地或私有云？

DataArts Studio必须基于华为云底座部署。资源隔离场景下，支持以全栈专属云模式部署，另外也支持以华为云Stack和HCS Online混合云模式部署。

关于全栈专属云、华为云Stack和HCS Online的适用场景和差异等更多信息，欢迎通过[咨询](#)了解。

1.6 如何在 IAM 中创建细粒度权限策略？

当前DataArts Studio不支持在IAM中创建细粒度权限策略。推荐通过DAYU策略+工作空间角色的方式进行权限控制。

DataArts Studio基于**DAYU系统角色+工作空间角色**实现授权的能力。为使IAM用户权限正常，IAM用户所在的用户组需要在IAM控制台中被授予DAYU User或DAYU Administrator的系统角色，另外也必须确保DAYU User角色的IAM用户已在对应的DataArts Studio工作空间中被设置为对应的工作空间角色。

工作空间角色决定了该用户在工作空间内的权限，当前有管理员、开发者、运维者和访客这四种预置角色可被分配。各角色权限的详细说明请参见[权限列表](#)章节。

- 管理员：工作空间管理员，拥有工作空间内所有的业务操作权限。建议将项目负责人、开发责任人、运维管理员设置为管理员角色。

- 开发者：开发者拥有工作空间内创建、管理工作项的业务操作权限。建议将任务开发、任务处理的用户设置为开发者。
- 运维者：运维者具备工作空间内运维调度等业务的操作权限，但无法更改工作项及配置。建议将运维管理、状态监控的用户设置为运维者。
- 访客：访客可以查看工作空间内的数据，但无法操作业务。建议将只查看空间内容、不进行操作的用户设置为访客。
- 部署者：企业模式独有，具备工作空间内任务包发布的相关操作权限。在企业模式中，开发者提交脚本或作业版本后，系统会对应产生发布任务。开发者确认发包后，需要部署者审批通过，才能将修改后的作业同步到生产环境。

1.7 用户已添加权限，还是无法查看工作空间？

DataArts Studio基于DAYU系统角色+工作空间角色实现授权的能力。为使IAM用户权限正常，IAM用户所在的用户组需要在IAM控制台中被授予DAYU User或DAYU Administrator的系统角色，另外也必须确保DAYU User角色的IAM用户已在对应的DataArts Studio工作空间中被设置为对应的工作空间角色。

如果您只给用户配置了DAYU User系统角色，未配置工作空间角色，则会出现无法查看工作空间的报错。请查看该工作空间下是否已添加用户，如果没有，请参考以下步骤添加该用户。

添加成员和角色

步骤1 参考[添加工作空间成员和角色](#)，登录DataArts Studio管理控制台。

步骤2 在“空间管理”页签，单击列表中相应工作空间后的“编辑”，弹出“空间信息”弹窗。

图 1-2 空间信息

空间信息

* 空间名称: default

描述: 请输入空间描述 (0/4,096)

* 空间模式: 简单模式 [升级]

* 企业项目: default [C]

作业日志OBS鉴证 [选择]

数据服务专享API配额: 已使用配额 9 / 已分配配额 10 [保存] / 总使用配额 9 / 总分配配额 10 / 总配额 6,000

* 空间成员

账号	用户类型	角色	加入时间	操作
[图标]	用户	管理员	2024/02/20 16:07:24 GMT+08:00	编辑
[图标]	用户	管理员	2024/01/27 16:33:00 GMT+08:00	编辑
[图标]	用户	管理员	2024/01/25 19:41:42 GMT+08:00	编辑
[图标]	用户	管理员	2024/01/18 14:47:06 GMT+08:00	编辑

[确定] [取消]

步骤3 单击空间成员下的“添加”，在弹出的“添加成员”对话框中选择“按用户添加”或“按用户组添加”，然后从“成员账号”的下拉选项中选择用户或用户组，并设置角色。

图 1-3 添加成员



步骤4 单击“确定”即可添加成功。添加完成后，您可以在空间成员列表中查看或修改已有的成员和对应角色，也可将空间成员从工作空间中删除。

----结束

1.8 IAM 用户操作时报错“无 xx 权限”怎么办？

DataArts Studio基于DAYU系统角色+工作空间角色实现授权的能力。为使IAM用户权限正常，IAM用户所在的用户组需要在IAM控制台中被授予DAYU User或DAYU Administrator的系统角色，另外也必须确保DAYU User角色的IAM用户已在对应的DataArts Studio工作空间中被设置为对应的工作空间角色。

如果您只给用户配置了工作空间的角色，则会出现无权限的报错。您需要检查IAM用户所在的用户组是否已经在IAM控制台中被授予DAYU User或DAYU Administrator的系统角色。IAM用户的创建和授权系统角色的具体操作如下：

步骤1 创建用户组并授权系统角色。

使用华为账号登录统一身份认证服务IAM控制台，创建用户组，并授予DataArts Studio的系统角色，如“DAYU Administrator”或“DAYU User”。

创建用户组并授权的具体操作，请参见[创建用户组并授权](#)。

📖 说明

- 配置用户组的DataArts Studio权限时，直接在搜索框中输入权限名“DAYU”进行搜索，然后勾选需要授予用户组的权限，如“DAYU User”。
- DataArts Studio部署时通过物理区域划分，为项目级服务。授权时，“授权范围方案”如果选择“所有资源”，则该权限在所有区域项目中都生效；如果选择“指定区域项目资源”，则该权限仅对此项目生效。IAM用户授权完成后，访问DataArts Studio时，需要先切换至授权区域。

步骤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[步骤1](#)中创建的用户组。

创建用户并加入用户组的具体操作，请参见[创建用户并加入用户组](#)。

📖 说明

仅当创建IAM用户时的访问方式勾选“编程访问”后，此IAM用户才能通过认证鉴权，从而使用API、SDK等方式访问DataArts Studio。

步骤3 为“DAYU User”系统角色用户自定义工作空间角色，并将其添加到工作空间成员、配置角色。

对于“DAYU User”权限的IAM用户而言，DataArts Studio工作空间角色决定了其在工作空间内的权限，当前有管理员、开发者、部署者、运维者和访客这五种预置角色可被分配。添加工作空间成员并配置角色的具体操作请参见[添加工作空间成员和角色](#)。

角色的权限说明请参见[权限列表](#)章节。

步骤4 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，例如：

- 在“服务列表”中选择数据治理中心，进入DataArts Studio实例卡片。从实例卡片进入控制台首页后，确认能否正常查看工作空间列表情况。
- 进入已添加当前用户的工作空间业务模块（例如管理中心），查看能否根据所配置的工作空间角色，正常进行业务操作。

----结束

1.9 DataArts Studio 的工作空间可以删除吗？

DataArts Studio已经支持删除工作空间。

📖 说明

为避免误删除导致的业务受损，删除工作空间需要**DAYU Administrator**或**Tenant Administrator**账号才能操作，且删除工作空间的前提是各组件内已无业务资源，各组件校验的资源如下：

- 管理中心组件：数据连接。
- 数据集成组件：数据集成集群。
- 数据架构组件：主题设计，逻辑模型，标准设计，物理模型，维度建模和指标。
- 数据开发组件：作业，作业目录，脚本，脚本目录和资源。
- 数据质量组件：质量作业和对账作业。
- 数据目录组件：技术资产中的表（Table）和文件（File）类型资产，以及元数据采集任务。
- 数据服务组件：数据服务集群，API和APP。
- 数据安全组件：敏感数据发现任务，脱敏策略，静态脱敏任务和数据水印任务。

如果当前任意组件内还有业务资源，则删除工作空间会弹出失败提示窗口，无法删除。

1. 登录DataArts Studio控制台，进入“空间管理”页签。
2. 在“空间管理”页面，找到所需删除的工作空间，单击其所在行的“更多 > 删除”。
3. 在“删除工作空间”对话框中，如果确认删除，请单击“确认”。

如果当前各组件内还有业务资源，则需要根据失败提示窗口，删除对应业务资源后再次重试删除。

图 1-4 删除失败提示



1.10 实例试用/购买成功后，可以转移到其他账号下吗？

不可以，实例试用/购买后不能转移到另一个账户。

如需在当前添加另一个账户，请参见[授权用户使用DataArts Studio](#)。

1.11 DataArts Studio 是否支持版本升级？

支持。如果您的业务量不断增长，已购版本无法满足您的业务需求，建议您升级版本。

您可以登录DataArts Studio控制台，找到需要升级的DataArts Studio实例卡片，单击“升级”，然后根据页面提示购买更高规格的套餐。

- 升级时，已经产生的费用按天结算。
- 升级成功后，按新订购套餐进行计费。
- 套餐升级后，系统会根据升级后的套餐创建一个新的批量数据迁移集群。原基础包中的批量数据迁移集群仍为您保留，但是不会再产生费用。您需要将原集群的数据连接、作业等信息手动迁移到新集群，具体操作请参见[是否支持同步作业到其他集群？](#)。

1.12 DataArts Studio 是否支持版本降级？

已购买DataArts Studio实例后，不支持降级版本。

1.13 如何查看 DataArts Studio 的版本？

您可以在DataArts Studio实例卡片中查看DataArts Studio版本，如下图所示。

图 1-5 DataArts Studio 实例卡片



1.14 购买 DataArts Studio 实例，选不到指定的 IAM 项目下面，怎么办？

请确认当前账户是否有开通企业项目。

企业项目和IAM项目是互斥的，开通企业项目后，只能在企业项目下购买DataArts Studio实例，且一个企业项目下只能购买一个DataArts Studio实例。

图 1-6 购买 DataArts Studio 实例



1.15 DataArts Studio 的会话超时时间是多少，是否支持修改？

会话超时时间指的是如果用户超过该时长未操作界面，会话将会失效，需要重新登录。

会话超时策略可以在IAM服务进行设置，如图所示。



会话超时策略默认开启，不能关闭，管理员可以设置会话超时的时长，会话超时时长默认为1个小时，可以在15分钟~24小时之间进行设置，该策略对账号以及账号下的IAM用户都生效。

1.16 套餐包到期未续订或按需资源欠费时，我的数据会保留吗？

云服务进入宽限期/保留期后，华为云将会通过邮件、短信等方式向您发送提醒，提醒您续订或充值。保留期到期仍未续订或充值，存储在云服务中的数据将被删除、云服务资源将被释放。

- 宽限期：指客户的包周期资源到期未续订或按需资源欠费时，华为云提供给客户进行续费与充值的时间，宽限期内客户可正常访问及使用云服务。
- 保留期：指宽限期到期后客户的包周期资源仍未续订或按需资源仍未缴清欠款，将进入保留期。保留期内客户不能访问及使用云服务，但对客户存储在云服务中的数据仍予以保留。

华为云宽限期和保留期时长设定请参考[资源停止服务或逾期释放说明](#)。

1.17 如何查看套餐包的剩余时长？

您可以进入华为云官网，在用户名下拉列表中选择“费用中心”，然后进入“订单管理-续费管理”查看对应套餐包的剩余时长。

1.18 DataArts Studio 实例中的 CDM 没有计费是什么原因？

购买DataArts Studio实例后，会赠送一个免费的CDM集群，不产生费用。

1.19 为什么会提示每日执行节点个数超过上限，应该怎么处理？

每日执行节点个数即DataArts Studio不同实例版本中的作业节点调度次数/天配额，各版本配额差异可参见[如何选择DataArts Studio版本](#)。

原因分析

当每日作业节点调度的已使用次数+运行中次数+本日将运行次数之和达到配额时，将会提示每日执行节点个数超过上限。

解决方案

每日执行节点个数超过上限，一般是由于作业调度过于频繁导致的。

您可通过如下方式排查哪些作业调度节点次数较高，然后适当调整调度周期或停止调度即可。

1. 在数据开发模块控制台的左侧导航栏，选择“运维调度 > 实例监控”，日期选择当天，查看哪些作业调度较多。

- 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”，查看调度较多的作业设置的调度周期是否合理。如果调度周期不合理，建议适当调整这些调度周期或停止调度。一般每日执行节点个数超过上限都是由于分钟级别的作业导致的。

图 1-7 查看调度周期



2 管理中心

2.1 DataArts Studio 支持治理哪些数据湖？

DataArts Studio目前支持的数据连接，请参见[DataArts Studio支持的数据源](#)章节。

2.2 创建数据连接需要注意哪些事项？

创建DWS/MRS Hive/RDS/SparkSQL类型的数据连接时，需要绑定由CDM集群提供的代理服务，目前不支持低于1.8.6版本的CDM集群。

2.3 为什么 DWS/Hive/HBase 数据连接突然无法获取数据库或表的信息？

可能是由于CDM集群被关闭或者并发冲突导致，您可以通过切换agent代理来临时规避此问题。

建议您通过以下措施解决此问题：

步骤1 检查CDM集群是否被关机。

- 是，将CDM集群开机后，确认管理中心的数据连接恢复正常。
- 否，跳转至**步骤2**。

步骤2 检查该CDM集群是否同时被用于数据迁移作业和管理中心连接代理。

- 是，您可以错开数据迁移作业和管理中心连接代理的使用时间，或再创建CDM集群，与原有CDM集群分开使用。
- 否，跳转至**步骤3**。

步骤3 直接重启该CDM集群，释放连接池资源。确认管理中心的数据连接恢复正常。

----结束

2.4 为什么在创建数据连接的界面上 MRS Hive/HBase 集群不显示？

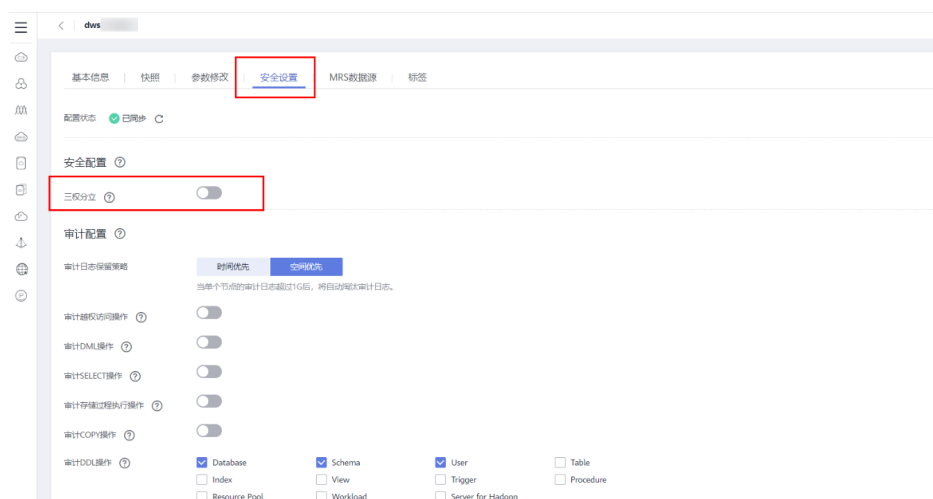
出现该问题的可能原因有：

- 创建MRS集群时未选择Hive/HBase组件。
- 创建MRS数据连接时所选择的CDM集群和MRS集群网络不互通。
CDM集群作为网络代理，与MRS集群需网络互通才可以成功创建基于MRS的数据连接。

2.5 创建 DWS 数据连接，开启 SSL 连接时测试连接失败？

可能是由于DWS集群的三权分立功能导致的。请在DWS控制台，单击进入对应的DWS集群后，选择“安全设置”，然后关闭三权分立功能。

图 2-1 关闭 DWS 集群三权分立功能



2.6 通过代理方式创建数据连接，一个空间可以创建多个连接吗？

同一个工作空间可以创建多个不同类型或相同类型的连接，但是连接的名字不能相同。

2.7 创建 DWS 连接的时候，连接方式是直接连还是通过代理连比较好？

连接方式一般选择代理连接即可。

2.8 如何将一个空间的数据开发作业和数据连接迁移到另一空间？

您可以在数据开发中将作业导出，随后在新空间数据开发中再导入作业。具体操作请参考[导出导入作业](#)。

您可以在管理中心中资源迁移进行数据连接的导入导出。具体操作请参考[资源迁移](#)。

3 数据集成

3.1 CDM 与其他数据迁移服务有什么区别，如何选择？

华为云上涉及数据迁移的服务有以下几种：

- [云数据迁移服务 CDM](#)
- [对象存储迁移服务 OMS](#)
- [数据复制服务 DRS](#)
- [主机迁移服务 SMS](#)
- [数据库和应用迁移 UGO](#)
- [数据快递服务 DES](#)

上述数据迁移服务的区别请参见[各个数据迁移服务区别](#)。

什么是云数据迁移服务(CDM)？

云数据迁移（Cloud Data Migration，简称CDM）是一种高效、易用的数据集成服务。CDM围绕大数据迁移上云和智能数据湖解决方案，提供了简单易用的迁移能力和多种数据源到数据湖的集成能力，降低了客户数据源迁移和集成的复杂性，有效的提高您数据迁移和集成的效率。更多详情请参见[云数据迁移服务](#)。

CDM进行数据迁移时，目标端为数据湖或其他大数据系统；源端可以是数据库也可以是对象存储。

CDM与DRS的区别：

- 目的端是大数据系统时，推荐使用CDM。
- 目的端是OLTP数据库或DWS时，推荐使用DRS迁移。

CDM与OMS的区别：

- OMS用于入云迁移，支持以下源端云服务商：亚马逊云、阿里云、微软云、百度云、青云、七牛云、腾讯云。
- CDM主要用于OBS数据迁移到数据湖或其他大数据系统，以便对数据进行开发、清洗、治理等。同时，整桶迁移建议使用OMS。

什么是对象存储迁移服务(OMS)?

对象存储迁移服务 (Object Storage Migration Service, 简称OMS) 是一种线上数据迁移服务, 帮助您将其他云服务商对象存储服务中的数据在线迁移至华为云的对象存储服务 (Object Storage Service, OBS) 中。简言之, 入云迁移、对象存储迁移。更多详情请参见[对象存储迁移服务](#)。

OMS主要功能有以下两个:

- 线上数据迁移服务: 帮助用户把对象存储数据从其他云服务商的公有云轻松、平滑地迁移上云。
- 跨区域的复制: 指的是华为云各个Region之间的数据复制和备份。

目前支持以下他云对象存储数据的入云迁移: 亚马逊云、阿里云、微软云、百度云、华为云、金山云、青云、七牛云、腾讯云。

云数据迁移CDM服务也同样支持对象存储数据迁移, 两者的区别为:

- OMS用于他云到华为云的数据迁移。
- CDM主要用于OBS数据迁移到数据湖或其他大数据系统, 以便对数据进行开发、清洗、治理等。

什么是数据复制服务(DRS)?

数据复制服务 (Data Replication Service, 简称DRS) 是一种易用、稳定、高效、用于数据库实时迁移和数据库实时同步的云服务。DRS适合迁移OLTP->OLTP、OLTP->DWS的场景都可以由DRS来完成数据迁移。即主流数据库到数据库 (含第三方数据库) 的场景, 使用DRS进行迁移。更多详情请参见[数据复制服务](#)。

目前支持的数据库链路有:

自建/他云MySQL->RDS for MySQL

自建/他云PostgreSQL->RDS for PostgreSQL

自建/他云MongoDB->DDS

Oracle->RDS for MySQL

.....

DRS与CDM的区别:

- DRS的目的端为数据库系统, 例如MySQL、MongoDB等。
- CDM的目的端主要为数据湖或其他大数据系统, 例如MRS HDFS、FusionInsight HDFS。

DRS和UGO的区别:

- DRS是针对数据的全量/增量迁移或数据同步。
- UGO用于异构数据库迁移前的评估、结构迁移和语法转化。

什么是主机迁移服务(SMS)?

主机迁移服务 (Server Migration Service, 简称SMS) 是一种P2V/V2V迁移服务, 可以帮您把X86物理服务器或者私有云、公有云平台上的虚拟机迁移到华为云弹性云服务器云主机上, 从而帮助您轻松地把服务器上的应用和数据迁移到华为云。更多详情请参见[主机迁移服务](#)。

主机迁移服务 SMS 是一种P2V/V2V迁移服务，可以把X86物理服务器、私有云或公有云平台上的虚拟机迁移到华为ECS上。

什么是数据库和应用迁移(UGO)?

数据库和应用迁移 UGO (Database and Application Migration UGO, 简称UGO) 是专注于异构数据库结构迁移的专业服务。可将数据库中的DDL、业务程序中封装的数据库SQL一键自动将语法转换为华为云GaussDB/RDS的SQL语法，通过预迁移评估、结构迁移两大核心功能和自动化语法转换，提前识别可能存在的改造工作、提高转化率、最大化降低用户数据库迁移成本。更多详情请参见[数据库和应用迁移](#)。

简言之，UGO用于异构数据库迁移前的数据库评估、结构迁移、语法转化。

什么是数据快递服务(DES)?

数据快递服务 (Data Express Service, 简称DES) 是一种海量数据传输解决方案，支持TB到PB级数据上云，通过Teleport设备或硬盘（外置USB接口、SATA接口、SAS接口类型）向华为云传输大量数据，致力于解决海量数据传输网络成本高、传输时间长等难题。更多详情请参见[数据快递服务](#)。

各个数据迁移服务区别

表 3-1 各个数据迁移服务区别

服务名	主要功能	与其他服务的区别
云数据迁移 CDM	<ul style="list-style-type: none"> 大数据迁移上云 多种数据源到数据湖的迁移 	与DRS的区别： 数据库迁移使用DRS；到大数据系统的迁移使用CDM。
对象存储迁移服务 OMS	对象存储迁移 <ul style="list-style-type: none"> 他云对象存储数据迁移到华为云 华为云各Region间的数据迁移 	与CDM的区别： OMS用于他云到华为云的数据迁移；CDM主要用于OBS数据迁移到数据湖或其他大数据系统，以便对数据进行开发、清洗、治理等。
数据复制服务 DRS	支持主流数据库到华为云的入云和出云迁移 <ul style="list-style-type: none"> 数据库在线迁移 数据库实时同步 	<ul style="list-style-type: none"> 与CDM的区别： 数据库迁移使用DRS；到大数据系统的迁移使用CDM。 与UGO的区别： DRS支持同构和异构的数据库迁移/同步；UGO用于异构数据库的结构迁移、数据库迁移前评估、语法迁移等。
主机迁移服务 SMS	主机迁移 含物理机到华为云、其他自建或他云虚拟机到华为云	-

服务名	主要功能	与其他服务的区别
数据库和应用迁移 UGO	<ul style="list-style-type: none"> 数据库结构迁移 数据库迁移前评估 语法迁移 	与DRS的区别： DRS支持同构和异构的数据库迁移/同步；UGO用于异构数据库的结构迁移、数据库迁移前评估、语法迁移等
数据快递服务 DES	<ul style="list-style-type: none"> 海量数据，支持TB级到PB级数据上云 使用物理介质 	-

3.2 CDM 有哪些优势？

云数据迁移（Cloud Data Migration，简称CDM）服务基于分布式计算框架，利用并行化处理技术，使用CDM迁移数据的优势如表3-2所示。

表 3-2 CDM 优势

优势项	用户自行开发	CDM
易使用	自行准备服务器资源，安装配置必要的软件并进行配置，等待时间长。 程序在读写两端会根据数据源类型，使用不同的访问接口，一般是数据源提供的对外接口，例如JDBC、原生API等，因此在开发脚本时需要依赖大量的库、SDK等，开发管理成本较高。	CDM提供了Web化的管理控制台，通过Web页实时开通服务。 用户只需要通过可视化界面对数据源和迁移任务进行配置，服务会对数据源和任务进行全面的管理和维护，用户只需关注数据迁移的具体逻辑，而不用关心环境等问题，极大降低了开发维护成本。 CDM还提供了REST API，支持第三方系统调用和集成。
实时监控	需要自行选型开发。	您可以使用云监控服务监控您的CDM集群，执行自动实时监控、告警和通知操作，帮助您更好地了解CDM集群的各项性能指标。
免运维	需要自行开发完善运维功能，自行保证系统可用性，尤其是告警及通知功能，否则只能人工值守。	使用CDM服务，用户不需要维护服务器、虚拟机等资源。CDM的日志，监控和告警功能，有异常可以及时通知相关人员，避免7*24小时人工值守。
高效率	在迁移过程中，数据读写过程都是由一个单一任务完成的，受限于资源，整体性能较低，对于海量数据场景通常不能满足要求。	CDM任务基于分布式计算框架，自动将任务切分为独立的子任务并行执行，能够极大提高数据迁移的效率。针对Hive、HBase、MySQL、DWS（数据仓库服务）数据源，使用高效的数据导入接口导入数据。

优势项	用户自行开发	CDM
多种数据源支持	数据源类型繁杂，针对不同数据源开发不同的任务，脚本数量成千上万。	支持数据库、Hadoop、NoSQL、数据仓库、文件等多种类型的数据源。
多种网络环境支持	随着云计算技术的发展，用户数据可能存在于各种环境中，例如公有云、自建/托管IDC、混合场景等。在异构环境中进行数据迁移需要考虑网络连通性等因素，给开发和维护都带来较大难度。	无论数据是在用户本地自建的IDC中（Internet Data Center，互联网数据中心）、云服务中、第三方云中，或者使用ECS自建的数据库或文件系统中，CDM均可帮助用户轻松应对各种数据迁移场景，包括数据上云，云上数据交换，以及云上数据回流本地业务系统。

3.3 CDM 有哪些安全防护？

CDM是一个完全托管的服务，提供了以下安全防护能力保护用户数据安全。

- 实例隔离：CDM服务的用户只能使用自己创建的实例，实例和实例之间是相互隔离的，不可相互访问。
- 系统加固：CDM实例的操作系统进行了特别的安全加固，攻击者无法从Internet访问CDM实例的操作系统。
- 密钥加密：用户在CDM上创建连接输入的各种数据源的密钥，CDM均采用高强度加密算法保存在CDM数据库。
- 无中间存储：数据在迁移的过程中，CDM只处理数据映射和转换，而不会存储任何用户数据或片段。

3.4 如何降低 CDM 使用成本？

如果是迁移公网的数据上云，可以使用NAT网关服务，实现CDM服务与子网中的其他弹性云服务器共享弹性IP，可以更经济、更方便的通过Internet迁移本地数据中心或第三方云上的数据。

具体操作如下：

1. 假设已经创建好了CDM集群（无需为CDM集群绑定专用弹性IP），记录下CDM集群所在的VPC和子网。
2. 创建NAT网关，注意选择和CDM集群相同的VPC、子网。
3. 创建完NAT网关后，回到NAT网关控制台列表，单击创建好的网关名称，然后选择“添加SNAT规则”。

图 3-1 添加 SNAT 规则

添加SNAT规则

* 子网名称: ↻ ?
192.168.2.0/24

* 弹性IP: ↻ ?

4. 选择子网和弹性IP，如果没有弹性IP，需要先申请一个。
完成之后，就可以到CDM控制台，通过Internet迁移公网的数据上云了。例如：迁移本地数据中心FTP服务器上的文件到OBS、迁移第三方云上关系型数据库到云服务RDS。

3.5 CDM 未使用数据传输功能时，是否会计费？

CDM集群运行状态下，即便未使用也是正常计费的，如果长期不使用建议删除集群，需要的时候再创建集群。CDM集群计费详情请参考[价格详情](#)。

3.6 已购买包年包月的 CDM 套餐包，为什么还会产生按需计费的费用？

请您先确认套餐包和实际的CDM集群是否具有相同区域和规格，如果非相同区域和规格，则无法使用套餐包。CDM集群规格和区域可以通过进入CDM主界面，进入“集群管理”，单击集群列表中的集群名称查看。

如果套餐包和实际的CDM集群具有相同区域和规格，则以下情况也会产生按需费用：

如果您先购买按需计费增量包，再购买套餐包，则在购买套餐包之前已经产生的费用以按需计费结算，购买套餐包之后的费用按套餐包计时。

3.7 如何查看套餐包的剩余时长？

您可以进入华为云官网，在用户名下拉列表中选择“费用中心”，然后进入“订单管理-续费管理”查看对应套餐包的剩余时长。

3.8 CDM 可以跨账户使用吗？

CDM不支持跨账户使用，可以同一账户IAM子用户使用。

3.9 CDM 集群是否支持升级操作？

CDM集群目前不支持升级操作，如果需要使用高版本集群则需要重新创建。

3.10 CDM 迁移性能如何？

单个cdm.large规格实例理论上可以支持1TB~8TB/天的数据迁移，实际传输速率受公网带宽、集群规格、文件读写速度、作业并发数设置、磁盘读写性能等因素影响。更多详情请参见[性能白皮书](#)。

3.11 CDM 不同集群规格对应并发的作业数是多少？

CDM通过数据迁移作业，将源端数据迁移到目的端数据源中。其中，主要运行逻辑如下：

1. 数据迁移作业提交运行后，CDM会根据作业配置中的“抽取并发数”参数，将每个作业拆分为多个Task，即作业分片。

📖 说明

- 不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。
2. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。

如何调整抽取并发数

1. 集群最大抽取并发数的设置与CDM集群规格有关，并发数上限建议配置为vCPU核数*2，如[表3-3](#)所示。

表 3-3 集群最大抽取并发数配置建议

规格名称	vCPUs/内存	集群并发数上限参考
cdm.large	8核 16GB	16
cdm.xlarge	16核 32GB	32
cdm.4xlarge	64核 128GB	128

图 3-2 集群最大抽取并发数配置



2. 作业抽取并发数的配置原则如下：

- a. 迁移的目的端为文件时，CDM不支持多并发，此时应配置为单进程抽取数据。
- b. 表中每行数据大小为1MB以下的可以设置多并发抽取，超过1MB的建议单线程抽取数据。
- c. 作业抽取并发数可参考集群最大抽取并发数配置，但不建议超过集群最大抽取并发数上限。
- d. 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。

图 3-3 作业抽取并发数配置

任务配置

作业失败重试 ?

作业分组 ? 添加 编辑 删除

是否定时执行

[隐藏高级属性](#)

抽取并发数 ?

分片重试次数 ?

是否写入脏数据 ?

开启限速 ?

3.12 是否支持增量迁移？

CDM支持增量数据迁移。利用定时任务配置和时间宏变量函数等参数，可支持以下场景的增量数据迁移：

- 文件增量迁移
- 关系数据库增量迁移
- HBase/CloudTable增量迁移

详情请参见[增量迁移](#)。

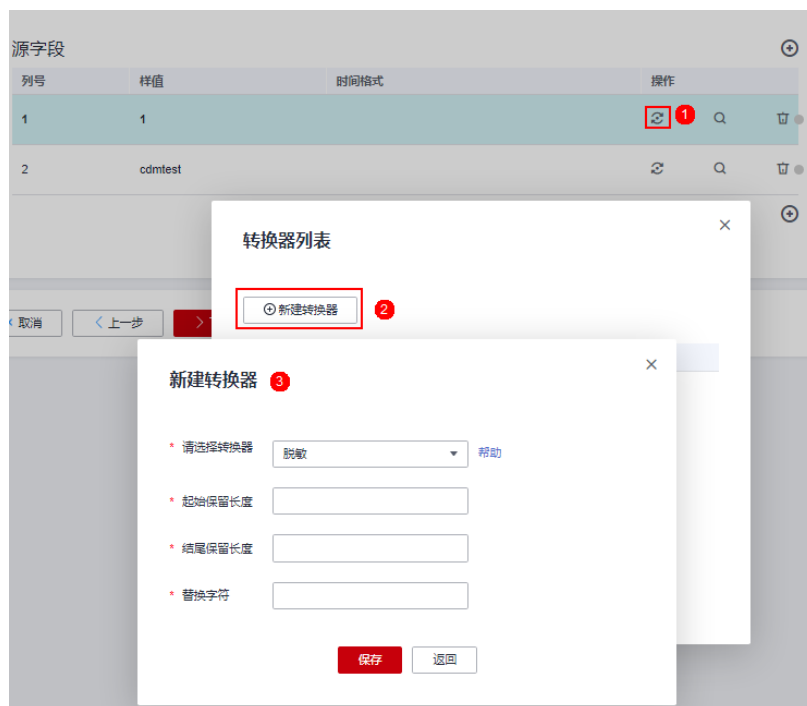
3.13 是否支持字段转换？

支持，CDM支持以下字段转换器：

- [脱敏](#)
- [去前后空格](#)
- [字符串反转](#)
- [字符串替换](#)
- [表达式转换](#)

在创建表/文件迁移作业的字段的映射界面，可新建字段转换器，如[图3-4](#)所示。

图 3-4 新建字段转换器



脱敏

隐藏字符串中的关键信息，例如要将“12345678910”转换为“123****8910”，则配置如下：

- “起始保留长度”为“3”。
- “结尾保留长度”为“4”。
- “替换字符”为“*”。

去前后空格

自动去字符串前后的空值，不需要配置参数。

字符串反转

自动反转字符串，例如将“ABC”转换为“CBA”，不需要配置参数。

字符串替换

替换字符串，需要用户配置被替换的对象，以及替换后的值。

表达式转换

使用JSP表达式语言（Expression Language）对当前字段或整行数据进行转换。JSP表达式语言可以用来创建算术和逻辑表达式。在表达式内可以使用整型数，浮点数，字符串，常量true、false和null。

- 表达式支持以下两个环境变量：
 - value：当前字段值。

- row: 当前行，数组类型。
- 表达式支持的工具类用法罗列如下，未列出即表示不支持：
 - a. 如果当前字段为字符串类型，将字符串全部转换为小写，例如将“aBC”转换为“abc”。
表达式: `StringUtils.toLowerCase(value)`
 - b. 将当前字段的字符串全部转为大写。
表达式: `StringUtils.toUpperCase(value)`
 - c. 如果想将第1个日期字段格式从“2018-01-05 15:15:05”转换为“20180105”。
表达式: `DateUtils.format(DateUtils.parseDate(row[0],"yyyy-MM-dd HH:mm:ss"),"yyyyMMdd")`
 - d. 如果想将时间戳转换成“yyyy-MM-dd hh:mm:ss”格式的日期字符串的类型，例如字段值为“1701312046588”，转化后为“2023-11-30 10:40:46”。
表达式: `DateUtils.format(NumberUtils.toLong(value),"yyyy-MM-dd HH:mm:ss")`
 - e. 如果想将“yyyy-MM-dd hh:mm:ss”格式的日期字符串转换成时间戳的类型。
表达式: `DateUtils.getTime(DateUtils.parseDate(value,"yyyy-MM-dd hh:mm:ss"))`
 - f. 如果当前字段值为“yyyy-MM-dd”格式的日期字符串，需要截取年，例如字段值为“2017-12-01”，转换后为“2017”。
表达式: `StringUtils.substringBefore(value,"-")`
 - g. 如果当前字段值为数值类型，转换后值为当前值的两倍。
表达式: `value*2`
 - h. 如果当前字段值为“true”，转换后为“Y”，其它值则转换后为“N”。
表达式: `value=="true"? "Y": "N"`
 - i. 如果当前字段值为字符串类型，当为空时，转换为“Default”，否则不转换。
表达式: `empty value? "Default":value`
 - j. 如果想将日期字段格式从“2018/01/05 15:15:05”转换为“2018-01-05 15:15:05”。
表达式: `DateUtils.format(DateUtils.parseDate(value,"yyyy/MM/dd HH:mm:ss"),"yyyy-MM-dd HH:mm:ss")`
 - k. 获取一个36位的UUID（Universally Unique Identifier，通用唯一识别码）。
表达式: `CommonUtils.randomUUID()`
 - l. 如果当前字段值为字符串类型，将首字母转换为大写，例如将“cat”转换为“Cat”。
表达式: `StringUtils.capitalize(value)`
 - m. 如果当前字段值为字符串类型，将首字母转换为小写，例如将“Cat”转换为“cat”。
表达式: `StringUtils.uncapitalize(value)`
 - n. 如果当前字段值为字符串类型，使用空格填充为指定长度，并且将字符串居中，当字符串长度不小于指定长度时不转换，例如将“ab”转换为长度为4的“ab”。

- 表达式: `StringUtils.center(value,4)`
- o. 删除字符串末尾的一个换行符 (包括“`\n`”、“`\r`”或者“`\r\n`”), 例如将“`abc\r\n\r\n`”转换为“`abc\r\n`”。
- 表达式: `StringUtils.chomp(value)`
- p. 如果字符串中包含指定的字符串, 则返回布尔值`true`, 否则返回`false`。例如“`abc`”中包含“`a`”, 则返回`true`。
- 表达式: `StringUtils.contains(value,"a")`
- q. 如果字符串中包含指定字符串的任一字符, 则返回布尔值`true`, 否则返回`false`。例如“`zzabyycdxx`”中包含“`z`”或“`a`”任意一个, 则返回`true`。
- 表达式: `StringUtils.containsAny(value,"za")`
- r. 如果字符串中不包含指定的所有字符, 则返回布尔值`true`, 包含任意一个字符则返回`false`。例如“`abz`”中包含“`xyz`”里的任意一个字符, 则返回`false`。
- 表达式: `StringUtils.containsNone(value,"xyz")`
- s. 如果当前字符串只包含指定字符串中的字符, 则返回布尔值`true`, 包含任意一个其它字符则返回`false`。例如“`abab`”只包含“`abc`”中的字符, 则返回`true`。
- 表达式: `StringUtils.containsOnly(value,"abc")`
- t. 如果字符串为空或`null`, 则转换为指定的字符串, 否则不转换。例如将空字符转换为`null`。
- 表达式: `StringUtils.defaultIfEmpty(value,null)`
- u. 如果字符串以指定的后缀结尾 (包括大小写), 则返回布尔值`true`, 否则返回`false`。例如“`abcdef`”后缀不为`null`, 则返回`false`。
- 表达式: `StringUtils.endsWith(value,null)`
- v. 如果字符串和指定的字符串完全一样 (包括大小写), 则返回布尔值`true`, 否则返回`false`。例如比较字符串“`abc`”和“`ABC`”, 则返回`false`。
- 表达式: `StringUtils.equals(value,"ABC")`
- w. 从字符串中获取指定字符串的第一个索引, 没有则返回整数-1。例如从“`aabaabaa`”中获取“`ab`”的第一个索引1。
- 表达式: `StringUtils.indexOf(value,"ab")`
- x. 从字符串中获取指定字符串的最后一个索引, 没有则返回整数-1。例如从“`aFkyk`”中获取“`k`”的最后一个索引4。
- 表达式: `StringUtils.lastIndexOf(value,"k")`
- y. 从字符串中指定的位置往后查找, 获取指定字符串的第一个索引, 没有则转换为“-1”。例如“`aabaabaa`”中索引3的后面, 第一个“`b`”的索引是5。
- 表达式: `StringUtils.indexOf(value,"b",3)`
- z. 从字符串获取指定字符串中任一字符的第一个索引, 没有则返回整数-1。例如从“`zzabyycdxx`”中获取“`z`”或“`a`”的第一个索引0。
- 表达式: `StringUtils.indexOfAny(value,"za")`
- aa. 如果字符串仅包含Unicode字符, 返回布尔值`true`, 否则返回`false`。例如“`ab2c`”中包含非Unicode字符, 返回`false`。
- 表达式: `StringUtils.isAlpha(value)`
- ab. 如果字符串仅包含Unicode字符或数字, 返回布尔值`true`, 否则返回`false`。例如“`ab2c`”中仅包含Unicode字符和数字, 返回`true`。
- 表达式: `StringUtils.isAlphanumeric(value)`

- ac. 如果字符串仅包含Unicode字符、数字或空格，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。
表达式: `StringUtils.isAlphanumericSpace(value)`
- ad. 如果字符串仅包含Unicode字符或空格，返回布尔值true，否则返回false。例如“ab2c”中包含Unicode字符和数字，返回false。
表达式: `StringUtils.isAlphaSpace(value)`
- ae. 如果字符串仅包含ASCII可打印字符，返回布尔值true，否则返回false。例如“!ab-c~”返回true。
表达式: `StringUtils.isAsciiPrintable(value)`
- af. 如果字符串为空或null，返回布尔值true，否则返回false。
表达式: `StringUtils.isEmpty(value)`
- ag. 如果字符串中仅包含Unicode数字，返回布尔值true，否则返回false。
表达式: `StringUtils.isNumeric(value)`
- ah. 获取字符串最左端的指定长度的字符，例如获取“abc”最左端的2位字符“ab”。
表达式: `StringUtils.left(value,2)`
- ai. 获取字符串最右端的指定长度的字符，例如获取“abc”最右端的2位字符“bc”。
表达式: `StringUtils.right(value,2)`
- aj. 将指定字符串拼接至当前字符串的左侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”左侧，拼接后长度为8，则转换后为“zyzybat”。
表达式: `StringUtils.leftPad(value,8,"yz")`
- ak. 将指定字符串拼接至当前字符串的右侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”右侧，拼接后长度为8，则转换后为“batzyzy”。
表达式: `StringUtils.rightPad(value,8,"yz")`
- al. 如果当前字段为字符串类型，获取当前字符串的长度，如果该字符串为null，则返回0。
表达式: `StringUtils.length(value)`
- am. 如果当前字段为字符串类型，删除其中所有的指定字符串，例如从“queued”中删除“ue”，转换后为“qd”。
表达式: `StringUtils.remove(value,"ue")`
- an. 如果当前字段为字符串类型，移除当前字段末尾指定的子字符串。指定的子字符串若不在当前字段的末尾，则不转换，例如移除当前字段“www.domain.com”后的“.com”。
表达式: `StringUtils.removeEnd(value,".com")`
- ao. 如果当前字段为字符串类型，移除当前字段开头指定的子字符串。指定的子字符串若不在当前字段的开头，则不转换，例如移除当前字段“www.domain.com”前的“www.”。
表达式: `StringUtils.removeStart(value,"www.")`
- ap. 如果当前字段为字符串类型，替换当前字段中所有的指定字符串，例如将“aba”中的“a”用“z”替换，转换后为“zba”。
表达式: `StringUtils.replace(value,"a","z")`

- aq. 如果当前字段为字符串类型，一次替换字符串中的多个字符，例如将字符串“hello”中的“h”用“j”替换，“o”用“y”替换，转换后为“jelly”。
表达式：`StringUtils.replaceChars(value,"ho","jy")`
- ar. 如果字符串以指定的前缀开头（区分大小写），则返回布尔值true，否则返回false，例如当前字符串“abcdef”以“abc”开头，则返回true。
表达式：`StringUtils.startsWith(value,"abc")`
- as. 如果当前字段为字符串类型，去除字段中首、尾处所有指定的字符，例如去除“abcyx”中首尾所有的“x”、“y”、“z”和“b”，转换后为“abc”。
表达式：`StringUtils.strip(value,"xyzb")`
- at. 如果当前字段为字符串类型，去除字段末尾所有指定的字符，例如去除当前字段末尾的“abc”字符串。
表达式：`StringUtils.stripEnd(value,"abc")`
- au. 如果当前字段为字符串类型，去除字段开头所有指定的字符，例如去除当前字段开头的空格。
表达式：`StringUtils.stripStart(value,null)`
- av. 如果当前字段为字符串类型，获取字符串指定位置后（索引从0开始，包括指定位置的字符）的子字符串，指定位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后的字符串，则转换后为“cde”。
表达式：`StringUtils.substring(value,2)`
- aw. 如果当前字段为字符串类型，获取字符串指定区间（索引从0开始，区间起点包括指定位置的字符，区间终点不包含指定位置的字符）的子字符串，区间位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后、第4个字符（即e）之前的字符串，则转换后为“cd”。
表达式：`StringUtils.substring(value,2,4)`
- ax. 如果当前字段为字符串类型，获取当前字段里第一个指定字符后的子字符串。例如获取“abcba”中第一个“b”之后的子字符串，转换后为“cba”。
表达式：`StringUtils.substringAfter(value,"b")`
- ay. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符后的子字符串。例如获取“abcba”中最后一个“b”之后的子字符串，转换后为“a”。
表达式：`StringUtils.substringAfterLast(value,"b")`
- az. 如果当前字段为字符串类型，获取当前字段里第一个指定字符前的子字符串。例如获取“abcba”中第一个“b”之前的子字符串，转换后为“a”。
表达式：`StringUtils.substringBefore(value,"b")`
- ba. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符前的子字符串。例如获取“abcba”中最后一个“b”之前的子字符串，转换后为“abc”。
表达式：`StringUtils.substringBeforeLast(value,"b")`
- bb. 如果当前字段为字符串类型，获取嵌套在指定字符串之间的子字符串，没有匹配的则返回null。例如获取“tagabctag”中“tag”之间的子字符串，转换后为“abc”。
表达式：`StringUtils.substringBetween(value,"tag")`

- bc. 如果当前字段为字符串类型，删除当前字符串两端的控制字符（ $\text{char} \leq 32$ ），例如删除字符串前后的空格。
表达式：`StringUtils.trim(value)`
- bd. 将当前字符串转换为字节，如果转换失败，则返回0。
表达式：`NumberUtils.toByte(value)`
- be. 将当前字符串转换为字节，如果转换失败，则返回指定值，例如指定值配置为1。
表达式：`NumberUtils.toByte(value, 1)`
- bf. 将当前字符串转换为Double数值，如果转换失败，则返回0.0d。
表达式：`NumberUtils.toDouble(value)`
- bg. 将当前字符串转换为Double数值，如果转换失败，则返回指定值，例如指定值配置为1.1d。
表达式：`NumberUtils.toDouble(value, 1.1d)`
- bh. 将当前字符串转换为Float数值，如果转换失败，则返回0.0f。
表达式：`NumberUtils.toFloat(value)`
- bi. 将当前字符串转换为Float数值，如果转换失败，则返回指定值，例如配置指定值为1.1f。
表达式：`NumberUtils.toFloat(value, 1.1f)`
- bj. 将当前字符串转换为Int数值，如果转换失败，则返回0。
表达式：`NumberUtils.toInt(value)`
- bk. 将当前字符串转换为Int数值，如果转换失败，则返回指定值，例如配置指定值为1。
表达式：`NumberUtils.toInt(value, 1)`
- bl. 将字符串转换为Long数值，如果转换失败，则返回0。
表达式：`NumberUtils.toLong(value)`
- bm. 将当前字符串转换为Long数值，如果转换失败，则返回指定值，例如配置指定值为1L。
表达式：`NumberUtils.toLong(value, 1L)`
- bn. 将字符串转换为Short数值，如果转换失败，则返回0。
表达式：`NumberUtils.toShort(value)`
- bo. 将当前字符串转换为Short数值，如果转换失败，则返回指定值，例如配置指定值为1。
表达式：`NumberUtils.toShort(value, 1)`
- bp. 将当前IP字符串转换为Long数值，例如将“10.78.124.0”转换为LONG数值是“172915712”。
表达式：`CommonUtils.ipToLong(value)`
- bq. 从网络读取一个IP与物理地址映射文件，并存放到Map集合，这里的URL是IP与地址映射文件存放地址，例如“`http://10.114.205.45:21203/sqoop/IpList.csv`”。
表达式：`HttpsUtils.downloadMap("url")`
- br. 将IP与地址映射对象缓存起来并指定一个key值用于检索，例如“ipList”。
表达式：`CommonUtils.setCache("ipList", HttpsUtils.downloadMap("url"))`
- bs. 取出缓存的IP与地址映射对象。

- 表达式: `CommonUtils.getCache("ipList")`
- bt. 判断是否有IP与地址映射缓存。
表达式: `CommonUtils.cacheExists("ipList")`
- bu. 根据指定的偏移类型 (month/day/hour/minute/second) 及偏移量 (正数表示增加, 负数表示减少), 将指定格式的时间转换为一个新时间, 例如将 “2019-05-21 12:00:00” 增加8个小时。
表达式: `DateUtils.getCurrentTimeByZone("yyyy-MM-dd HH:mm:ss",value, "hour", 8)`
- bv. 如果value值为空或者null时, 则返回字符串"aaa", 否则返回value。
表达式: `StringUtils.defaultIfEmpty(value,"aaa")`

3.14 Hadoop 类型的数据源进行数据迁移时, 建议使用的组件版本有哪些?

建议使用的组件版本既可以作为目的端使用, 也可以作为源端使用。

表 3-4 建议使用的组件版本

Hadoop类型	组件	说明
MRS/Apache/ FusionInsight HD	Hive	暂不支持2.x版本, 建议使用的版本: <ul style="list-style-type: none"> • 1.2.X • 3.1.X
	HDFS	建议使用的版本: <ul style="list-style-type: none"> • 2.8.X • 3.1.X
	Hbase	建议使用的版本: <ul style="list-style-type: none"> • 2.1.X • 1.3.X

3.15 数据源为 Hive 时支持哪些数据格式?

云数据迁移服务支持从Hive数据源读写的数据格式包括SequenceFile、TextFile、ORC、Parquet。

3.16 是否支持同步作业到其他集群?

CDM虽然不支持直接在不同集群间迁移作业, 但是通过批量导出、批量导入作业的功能, 可以间接实现集群间的作业迁移, 方法如下:

1. 将CDM集群1中的所有作业批量导出, 将作业的JSON文件保存到本地。
由于安全原因, CDM导出作业时没有导出连接密码, 连接密码全部使用 “Add password here” 替换。

2. 在本地编辑JSON文件，将“Add password here”替换为对应连接的正确密码。
3. 将编辑好的JSON文件批量导入到CDM集群2，实现集群1和集群2之间的作业同步。

3.17 是否支持批量创建作业？

CDM可以通过批量导入的功能，实现批量创建作业，方法如下：

1. 手动创建一个作业。
2. 导出作业，将作业的JSON文件保存到本地。
3. 编辑JSON文件，参考该作业的配置，在JSON文件中批量复制出更多作业。
4. 将JSON文件导入CDM集群，实现批量创建作业。

您也可以参考[通过CDM算子批量创建分表迁移作业](#)，配合For Each算子，实现自动批量创建作业。

3.18 是否支持批量调度作业？

支持。

1. 访问DataArts Studio服务的数据开发模块。
2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”，新建作业。
3. 拖动多个CDM Job节点至画布，然后再编排作业。

3.19 如何备份 CDM 作业？

用户可以先通过CDM的批量导出功能，把所有作业脚本保存到本地，仅在需要的时候再重新创建集群、重新导入作业，实现作业备份。

3.20 如果 HANA 集群只有部分节点和 CDM 集群网络互通，应该如何配置连接？

如果HANA集群只有部分节点和CDM网络互通，为确保CDM正常连接HANA集群，则需要进行如下配置：

1. 关闭HANA集群的Statement Routing开关。但须注意，关闭Statement Routing，会增加配置节点的压力。
2. 新建HANA连接时，在高级属性中添加属性“distribution”，并将值置为“off”。

完成配置后，CDM即可正常连接HANA集群。

3.21 如何使用 Java 调用 CDM 的 Rest API 创建数据迁移作业？

CDM提供了Rest API，可以通过程序调用实现自动化的作业创建或执行控制。

这里以CDM迁移MySQL数据库的表city1的数据到DWS的表city2为例，介绍如何使用Java调用CDM服务的REST API创建、启动、查询、删除该CDM作业。

需要提前准备以下数据：

1. 云账号的用户名、账号名和项目ID。
2. 创建一个CDM集群，并获取集群ID。
获取方法：在集群管理界面，单击CDM集群名称可查看集群ID，例如“c110beff-0f11-4e75-8b10-da7cd882b0ef”。
3. 创建一个MySQL数据库和一个DWS数据库，并创建好表city1和表city2，创表语句如下：
4. 在CDM集群下，创建连接到MySQL的连接，例如连接名称为“mysqltestlink”。创建连接到DWS的连接，例如连接名称为“dwstestlink”。
5. 运行下述代码，依赖HttpClient包，建议使用4.5版本。Maven配置如下：

```
MySQL:
create table city1(code varchar(10),name varchar(32));
insert into city1 values('NY','New York');
DWS:
create table city2(code varchar(10),name varchar(32));
```

```
<project>
<modelVersion>4.0.0</modelVersion>
<groupId>cdm</groupId>
<artifactId>cdm-client</artifactId>
<version>1</version>
<dependencies>
<dependency>
<groupId>org.apache.httpcomponents</groupId>
<artifactId>httpclient</artifactId>
<version>4.5</version>
</dependency>
</dependencies>
</project>
```

代码示例

使用Java调用CDM服务的REST API创建、启动、查询、删除CDM作业的代码示例如下：

```
package cdmclient;
import java.io.IOException;
import org.apache.http.Header;
import org.apache.http.HttpEntity;
import org.apache.http.HttpHost;
import org.apache.http.auth.AuthScope;
import org.apache.http.auth.UsernamePasswordCredentials;
import org.apache.http.client.CredentialsProvider;
import org.apache.http.client.config.RequestConfig;
import org.apache.http.client.methods.CloseableHttpResponse;
import org.apache.http.client.methods.HttpDelete;
import org.apache.http.client.methods.HttpGet;
import org.apache.http.client.methods.HttpPost;
import org.apache.http.client.methods.HttpPut;
import org.apache.http.entity.StringEntity;
import org.apache.http.impl.client.BasicCredentialsProvider;
import org.apache.http.impl.client.CloseableHttpClient;
import org.apache.http.impl.client.HttpClients;
import org.apache.http.util.EntityUtils;
public class CdmClient {
private final static String DOMAIN_NAME="云账号名";
```



```
private final static String USER_NAME="云用户名";
private final static String USER_PASSWORD="云用户密码";
private final static String PROJECT_ID="项目ID";
private final static String CLUSTER_ID="CDM集群ID";
private final static String JOB_NAME="作业名称";
private final static String FROM_LINKNAME="源连接名称";
private final static String TO_LINKNAME="目的连接名称";
private final static String IAM_ENDPOINT="IAM的Endpoint";
private final static String CDM_ENDPOINT="CDM的Endpoint";
private CloseableHttpClient httpClient;
private String token;

public CdmClient() {
this.httpClient = createHttpClient();
this.token = login();
}

private CloseableHttpClient createHttpClient() {
CloseableHttpClient httpClient =HttpClients.createDefault();
return httpClient;
}

private String login(){
HttpPost httpPost = new HttpPost("https://" +IAM_ENDPOINT+"/v3/auth/tokens");
String json =
"{\r\n"+
  "\"auth\": {\r\n"+
  "\"identity\": {\r\n"+
  "\"methods\": [\"password\"],\r\n"+
  "\"password\": {\r\n"+
  "\"user\": {\r\n"+
  "\"name\": \""+USER_NAME+"\",\r\n"+
  "\"password\": \""+USER_PASSWORD+"\",\r\n"+
  "\"domain\": {\r\n"+
  "\"name\": \""+DOMAIN_NAME+"\"\r\n"+
  "}}\r\n"+
  "}}\r\n"+
  "}}\r\n"+
  "},\r\n"+
  "\"scope\": {\r\n"+
  "\"project\": {\r\n"+
  "\"name\": \"PROJECT_NAME\"\r\n"+
  "}}\r\n"+
  "}}\r\n"+
  "}}\r\n"+
  "}}\r\n";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
httpPost.setEntity(s);
CloseableHttpResponse response = httpClient.execute(httpPost);
Header tokenHeader = response.getFirstHeader("X-Subject-Token");
String token = tokenHeader.getValue();
System.out.println("Login successful");
return token;
} catch (Exception e) {
throw new RuntimeException("login failed.", e);
}
}
/*创建作业*/
```

```
public void createJob(){
    HttpPost httpPost = new HttpPost("https://" + CDM_ENDPOINT + "/cdm/v1.0/" + PROJECT_ID + "/clusters/" + CLUSTER_ID + "/cdm/job");

    /**此处JSON信息比较复杂，可以先在作业管理界面上创建一个作业，然后单击作业后的“作业JSON定义”，复制其中的JSON内容，格式化为Java字符串语法，然后粘贴到此处。
    *JSON消息体中一般只需要替换连接名、导入和导出的表名、导入导出表的字段列表、源表中用于分区的字段。*/

    String json =
        "{\r\n"+
        "\r\n"+
        "\"jobs\": [\r\n"+
        "{\r\n"+
        "\"from-connector-name\": \"generic-jdbc-connector\",\r\n"+
        "\"name\": \""+JOB_NAME+"\",\r\n"+
        "\"to-connector-name\": \"generic-jdbc-connector\",\r\n"+
        "\"driver-config-values\": {\r\n"+
        "\"configs\": [\r\n"+
        "{\r\n"+
        "\"inputs\": [\r\n"+
        "{\r\n"+
        "\"name\": \"throttlingConfig.numExtractors\",\r\n"+
        "\"value\": \"1\"\r\n"+
        "},\r\n"+
        "],\r\n"+
        "\"validators\": [],\r\n"+
        "\"type\": \"JOB\",\r\n"+
        "\"id\": 30,\r\n"+
        "\"name\": \"throttlingConfig\"\r\n"+
        "},\r\n"+
        "],\r\n"+
        "},\r\n"+
        "\"from-link-name\": \""+FROM_LINKNAME+"\",\r\n"+
        "\"from-config-values\": {\r\n"+
        "\"configs\": [\r\n"+
        "{\r\n"+
        "\"inputs\": [\r\n"+
        "{\r\n"+
        "\"name\": \"fromJobConfig.schemaName\",\r\n"+
        "\"value\": \"sqoop\"\r\n"+
        "},\r\n"+
        "},\r\n"+
        "{\r\n"+
        "\"name\": \"fromJobConfig.tableName\",\r\n"+
        "\"value\": \"city1\"\r\n"+
        "},\r\n"+
        "},\r\n"+
        "{\r\n"+
        "\"name\": \"fromJobConfig.columnList\",\r\n"+
        "\"value\": \"code&name\"\r\n"+
        "},\r\n"+
        "},\r\n"+
        "{\r\n"+
        "\"name\": \"fromJobConfig.partitionColumn\",\r\n"+
        "\"value\": \"code\"\r\n"+
        "},\r\n"+
        "],\r\n"+
        "\"validators\": [],\r\n"+
        "\"type\": \"JOB\",\r\n"+
        "\"id\": 7,\r\n"+
        "\"name\": \"fromJobConfig\"\r\n"+
        "},\r\n"+
        "],\r\n"+
        "},\r\n"+
        "\"to-link-name\": \""+TO_LINKNAME+"\",,\r\n"+
```

```
"\to-config-values\": {\r\n"+
"\configs\": [\r\n"+
"{\r\n"+
"inputs\": [\r\n"+
{\r\n"+
"name\": \"toJobConfig.schemaName\", \r\n"+
"value\": \"sqoop\" \r\n"+
}, \r\n"+
{\r\n"+
"name\": \"toJobConfig.tableName\", \r\n"+
"value\": \"city2\" \r\n"+
}, \r\n"+
{\r\n"+
"name\": \"toJobConfig.columnList\", \r\n"+
"value\": \"code&name\" \r\n"+
}, \r\n"+
{\r\n"+
"name\": \"toJobConfig.shouldClearTable\", \r\n"+
"value\": \"true\" \r\n"+
} \r\n"+
], \r\n"+
"validators\": [], \r\n"+
"type\": \"JOB\", \r\n"+
"id\": 9, \r\n"+
"name\": \"toJobConfig\" \r\n"+
} \r\n"+
] \r\n"+
} \r\n"+
} \r\n"+
} \r\n"+
} \r\n"+
} \r\n"+
} \r\n"+
} \r\n";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
httpPost.setEntity(s);
httpPost.addHeader("X-Auth-Token", this.token);
httpPost.addHeader("X-Language", "en-us");
CloseableHttpResponse response = httpClient.execute(httpPost);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
System.out.println("Create job successful.");
}else{
System.out.println("Create job failed.");
HttpEntity entity = response.getEntity();
System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
e.printStackTrace();
throw new RuntimeException("Create job failed.", e);
}
}
}
/*启动作业*/

public void startJob(){
HttpPut httpPut = new HttpPut("https://"+CDM_ENDPOINT+"/cdm/v1.0/"+PROJECT_ID+"/clusters/"+CLUSTER_ID+"/cdm/job/"+JOB_NAME+"/start");
String json = "";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
```

```
httpPut.setEntity(s);
httpPut.addHeader("X-Auth-Token", this.token);
httpPut.addHeader("X-Language", "en-us");
CloseableHttpResponse response = httpClient.execute(httpPut);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
    System.out.println("Start job successful.");
}else{
    System.out.println("Start job failed.");
    HttpEntity entity = response.getEntity();
    System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
    e.printStackTrace();
    throw new RuntimeException("Start job failed.", e);
}
}
/*循环查询作业运行状态，直到作业运行结束。*/

public void getJobStatus(){
    HttpGet httpGet = new HttpGet("https://" + CDM_ENDPOINT + "/cdm/v1.0/" + PROJECT_ID + "/clusters/" + CLUSTER_ID + "/cdm/job/" + JOB_NAME + "/status");
    try {
        httpGet.addHeader("X-Auth-Token", this.token);
        httpGet.addHeader("X-Language", "en-us");
        boolean flag = true;
        while(flag){
            CloseableHttpResponse response = httpClient.execute(httpGet);
            int status = response.getStatusLine().getStatusCode();
            if(status == 200){
                HttpEntity entity = response.getEntity();
                String msg = EntityUtils.toString(entity);
                if(msg.contains("\"status\": \"SUCCEEDED\"")){
                    System.out.println("Job succeeded");
                    break;
                }else if (msg.contains("\"status\": \"FAILED\"")){
                    System.out.println("Job failed.");
                    break;
                }else{
                    Thread.sleep(1000);
                }
            }else{
                System.out.println("Get job status failed.");
                HttpEntity entity = response.getEntity();
                System.out.println(EntityUtils.toString(entity));
                break;
            }
        }
    } catch (Exception e) {
        e.printStackTrace();
        throw new RuntimeException("Get job status failed.", e);
    }
}
/*删除作业*/

public void deleteJob(){
    HttpDelete httpDelte = new HttpDelete("https://" + CDM_ENDPOINT + "/cdm/v1.0/" + PROJECT_ID + "/clusters/" + CLUSTER_ID + "/cdm/job/" + JOB_NAME);
    try {
        httpDelte.addHeader("X-Auth-Token", this.token);
        httpDelte.addHeader("X-Language", "en-us");
```

```
CloseableHttpResponse response = httpClient.execute(httpDelete);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
    System.out.println("Delete job successful.");
}else{
    System.out.println("Delete job failed.");
    HttpEntity entity = response.getEntity();
    System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
    e.printStackTrace();
    throw new RuntimeException("Delete job failed.", e);
}
}
/*关闭*/

public void close(){
    try {
        httpClient.close();
    } catch (IOException e) {
        throw new RuntimeException("Close failed.", e);
    }
}

public static void main(String[] args){
    CdmClient cdmClient = new CdmClient();
    cdmClient.createJob();
    cdmClient.startJob();
    cdmClient.getJobStatus();
    cdmClient.deleteJob();
    cdmClient.close();
}
}
```

3.22 如何将云下内网或第三方云上的私网与 CDM 连通？

很多企业会把关键数据源建设在内网，例如数据库、文件服务器等。由于CDM运行在云上，如果要通过CDM迁移内网数据到云上的话，可以通过以下几种方式连通内网和CDM的网络：

- 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
- 在本地数据中心和云服务VPC之间建立VPN通道。
- 通过NAT（网络地址转换，Network Address Translation）或端口转发，以代理的方式访问。

这里重点介绍如何通过端口转发工具来实现访问内部数据，流程如下：

1. 找一台windows机器作为网关，该机器必须可以直接访问Internet，同时可以访问内网。
2. 在该机器上安装端口映射工具（IPOP）。
3. 通过端口映射工具（IPOP）配置端口映射。

须知

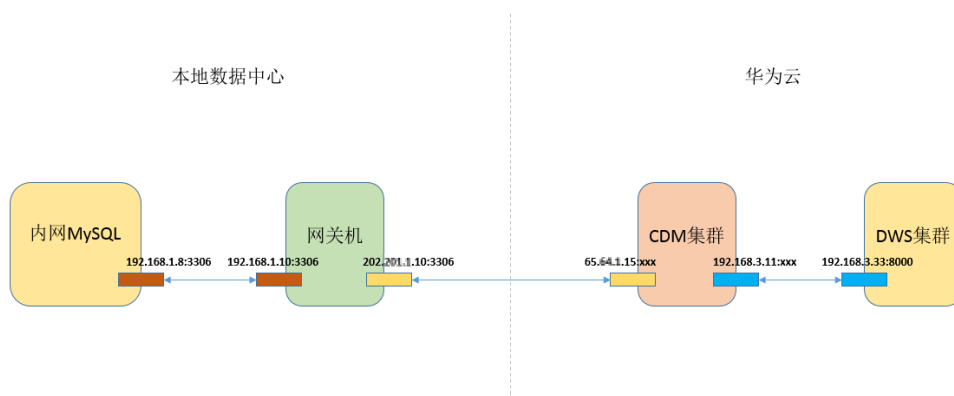
长时间将内网数据库暴露在公网会有安全风险，迁移数据完成后，请及时停止端口映射。

场景描述

这里假设是将内网MySQL迁移到云服务DWS，网络拓扑样例如图3-5所示。

图中的内网既可以是企业自己的数据中心，也可以是在第三方云的虚拟数据中心私网。

图 3-5 网络拓扑样例

**操作步骤**

步骤1 找一台Windows机器作为网关机，该机器同时配置内网和外网IP。通过以下测试来确保网关机器的服务要求：

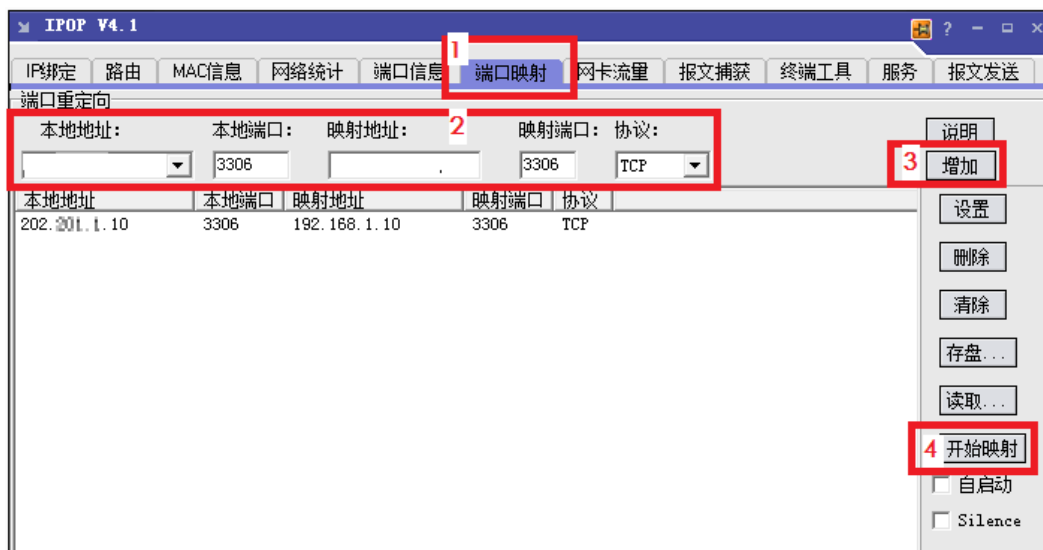
1. 在该机器上ping内网MySQL地址可以ping通，例如：**ping 192.168.1.8**。
2. 在另外一台可上网的机器上ping网关机的公网地址可以ping通，例如**ping 202.xx.xx.10**。

步骤2 下载端口映射工具IPOP，在网关机上安装IPOP。

步骤3 运行端口映射工具，选择“端口映射”，如图3-6所示。

- 本地地址、本地端口：配置为网关机的公网地址和端口（后续在CDM上创建MySQL连接时输入这个地址和端口）。
- 映射地址、映射端口：配置为内网MySQL的地址和端口。

图 3-6 配置端口映射



步骤4 单击“增加”，添加端口映射关系。

步骤5 单击“开始映射”，这时才会真正开始映射，接收数据包。

至此，就可以在CDM上通过弹性IP读取本地内网MySQL的数据，然后导入到云服务DWS中。

说明

1. CDM要访问本地数据源，也必须给CDM集群配置EIP。
2. 一般云服务DWS默认也是只允许VPC内部访问，创建CDM集群时，必须将CDM的VPC与DWS配置一致，且推荐在同一个内网和安全组，如果不同，还需要配置允许两个安全组之间的数据访问。
3. 端口映射不仅可以用于迁移内网数据库的数据，还可以迁移例如SFTP服务器上的数据。
4. Linux机器也可以通过IPTABLE实现端口映射。
5. 内网中的FTP通过端口映射到公网时，需要检查是否启用了PASV模式。这种情况下客户端和服务端建立连接的时候是走的随机端口，所以除了配置21端口映射外，还需要配置PASV模式的端口范围映射，例如vsftp通过配置pasv_min_port和pasv_max_port指定端口范围。

----结束

3.23 CDM 是否支持参数或者变量？

如果CDM作业使用了在数据开发时配置的**作业参数**或者**变量**，则后续在DataArts Studio数据开发模块调度此节点，可以间接实现CDM作业根据参数变量进行数据迁移。

3.24 CDM 迁移作业的抽取并发数应该如何设置？

CDM通过数据迁移作业，将源端数据迁移到目的端数据源中。其中，主要运行逻辑如下：

1. 数据迁移作业提交运行后，CDM会根据作业配置中的“抽取并发数”参数，将每个作业拆分为多个Task，即作业分片。

说明

- 不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。
- CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。

如何调整抽取并发数

- 集群最大抽取并发数的设置与CDM集群规格有关，并发数上限建议配置为vCPU核数*2，如表3-5所示。

表 3-5 集群最大抽取并发数配置建议

规格名称	vCPUs/内存	集群并发数上限参考
cdm.large	8核 16GB	16
cdm.xlarge	16核 32GB	32
cdm.4xlarge	64核 128GB	128

图 3-7 集群最大抽取并发数配置



- 作业抽取并发数的配置原则如下：
 - 迁移的目的端为文件时，CDM不支持多并发，此时应配置为单进程抽取数据。
 - 表中每行数据大小为1MB以下的可以设置多并发抽取，超过1MB的建议单线程抽取数据。
 - 作业抽取并发数可参考集群最大抽取并发数配置，但不建议超过集群最大抽取并发数上限。

- d. 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。

图 3-8 作业抽取并发数配置

任务配置

作业失败重试 ?

作业分组 ? 添加 编辑 删除

是否定时执行

隐藏高级属性

抽取并发数 ?

分片重试次数 ?

是否写入脏数据 ?

开启限速 ?

3.25 CDM 是否支持动态数据实时迁移功能？

不支持。如果源端在迁移过程中写数据，可能会出现报错。

3.26 CDM 是否支持集群关机功能？

从2022年4月开始，CDM已不再支持集群关机功能。当集群关机时，其底层资源可能会被占用，导致集群可能无法正常开机使用。

3.27 如何使用表达式方式获取当前时间？

您可以在字段映射界面使用DateUtils.format(\${timestamp()},"yyyy-MM-dd HH:mm:ss")表达式获取当前时间，更多表达式设置方式可以参考[表达式转换](#)。

3.28 日志提示解析日期格式失败时怎么处理？

问题描述

在使用CDM迁移其他数据源到云搜索服务（Cloud Search Service）的时候，作业执行失败，日志提示“Unparseable date”，如图3-9所示。

图 3-9 日志提示信息

```
java.text.ParseException: Unparseable date: "2018/01/05 15:15:46"
    at java.text.DateFormat.parse(DateFormat.java:366) ~[na:1.8.0_112]
    at org.apache.sqoop.connector.common.DataTypeUtil.convertDateFormat
    at org.apache.sqoop.connector.elasticsearch.ElasticSearchLoader.toJ
    at org.apache.sqoop.connector.elasticsearch.ElasticSearchLoader.arr
7]
    at org.apache.sqoop.connector.elasticsearch.ElasticSearchLoader.loa
```

原因分析

云搜索服务对于时间类型有一个特殊处理：如果存储的时间数据不带时区信息，在 Kibana 可视化的时候，Kibana 会认为该时间为 GMT 标准时间。

在各个地区会产生日志显示时间与本地时区时间不一致的现象，例如，在东八区某地，日志显示时间比本地时区时间少 8 个小时。因此在 CDM 迁移数据到云搜索服务的时候，如果是通过 CDM 自动创建的索引和类型（例如图 3-10 中，目的端的 “date_test” 和 “test1” 在云搜索服务中不存在时，CDM 会在云搜索服务中自动创建该索引和类型），则 CDM 默认会将时间类型字段的格式设置为 “yyyy-MM-dd HH:mm:ss.SSS Z” 的标准格式，例如 “2018-01-08 08:08:08.666 +0800”。

图 3-10 作业配置

作业配置

* 作业名称: test_date

源连接配置 **目的连接配置**

* 源连接名称: OBS1 [新建连接] * 目的连接名称: ES1 [新建连接]

* 桶名: peiveli-cdm [...] * 索引: date_test [...]

* 源目录或文件: /from/date_test.log [...] * 类型: test1 [...]

* 文件格式: CSV格式

[显示高级属性](#)

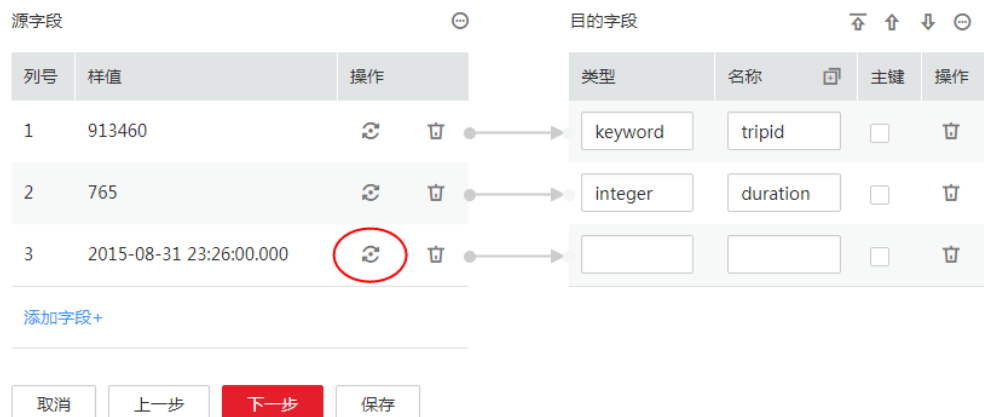
[取消] [下一步]

此时，从其他数据源导入数据到云搜索服务时，如果源端数据中的日期格式不完全满足标准格式，例如 “2018/01/05 15:15:46”，则 CDM 作业会执行失败，日志提示无法解析日期格式。需要通过 CDM 配置字段转换器，将日期字段的格式转换为云搜索服务的目的端格式。

解决方法

1. 编辑作业，进入作业的字段的映射步骤，在源端的时间格式字段后面，选择新建转换器，如图 3-11 所示。

图 3-11 新建转换器



2. 转换器类型选择“表达式转换”，目前表达式转换支持字符串和日期类型的函数，语法和Java的字符串和时间格式函数非常相似，可以查看[表达式转换](#)了解如何编写表达式。
3. 本例中源时间格式是“yyyy/MM/dd HH:mm:ss”，要将其转换成“yyyy-MM-dd HH:mm:ss.SSS Z”，需要经过如下几步：
 - a. 添加时区信息“+0800”到原始日期字符串的尾部，对应的表达式为：**value + " +0800"**。
 - b. 使用原始日期格式来解析字符串，将字符串解析为一个日期对象。可以使用DateUtils.parseDate函数来解析，语法是：**DateUtils.parseDate(String value, String format)**。
 - c. 将日期对象格式化为目标格式的字符串，可以使用DateUtils.format函数来格式化，语法是**DateUtils.format(Date date, String format)**。

因此本例中串起来完整的表达式是：

DateUtils.format(DateUtils.parseDate(value+" +0800","yyyy/MM/dd HH:mm:ss Z"),"yyyy-MM-dd HH:mm:ss.SSS Z")，如[图3-12](#)所示。

图 3-12 配置表达式



4. 保存转换器配置，再保存并运行作业，可解决云搜索服务的解析日期格式失败问题。

3.29 字段映射界面无法显示所有列怎么处理？

问题描述

在使用CDM从HBase/CloudTable导出数据时，在字段映射界面HBase/CloudTable表的字段偶尔显示不全，无法与目的端字段一一匹配，造成导入到目的端的数据不完整。

原因分析

由于HBase/CloudTable无Schema，每条数据的列数不固定，在字段映射界面CDM通过获取样值的方式有较大概率无法获得所有列，此时作业执行完后会造成目的端的数据不全。

这个问题，可以通过以下方法解决：

1. 在CDM的字段映射界面增加字段。
2. 在CDM的作业管理界面直接编辑作业的JSON（修改“fromJobConfig.columns”、“toJobConfig.columnList”这两个参数）。
3. 导出作业的JSON文件到本地，在本地手动修改JSON文件中的参数后（原理同2相同），再导回CDM。

推荐使用方法1，下面以HBase导出到DWS为例进行说明。

解决方法一：CDM 的字段映射界面增加字段

1. 获取源端HBase待迁移的表中所有的字段，列族与列之间用“:”分隔，例如：

```
rowkey:rowkey
g:DAY_COUNT
g:CATEGORY_ID
g:CATEGORY_NAME
g:FIND_TIME
g:UPLOAD_PEOPLE
g:ID
g:INFOMATION_ID
g:TITLE
g:COORDINATE_X
g:COORDINATE_Y
g:COORDINATE_Z
g:CONTENT
g:IMAGES
g:STATE
```

2. 在CDM的作业管理界面，找到HBase导出数据到DWS的作业，单击作业后面的“编辑”，进入字段映射界面，如图3-13所示。

图 3-13 字段映射 03

源字段					+	目的字段
列族	列号	样值	时间格式	操作		名称
rowkey	rowkey	1		🔄 🔍 🗑️	→	rowkey
g	DAY_COUNT	3		🔄 🔍 🗑️	→	day_count
g	CATEGORY_ID	4		🔄 🔍 🗑️	→	category
g	CATEGORY_NAME	3		🔄 🔍 🗑️	→	category_name

取消 上一步 下一步


- 单击  添加字段，在弹出框中选择“添加新字段”，如图3-14所示。

图 3-14 添加字段 04

添加字段 ×

添加被移除的字段
添加新字段

列名

列族

确认
取消

📖 说明

- 添加完字段后，新增的字段在界面不显示样值，这个不影响字段值的传输，CDM会将字段值直接写入目的端。
 - 这里“添加新字段”的功能，要求源端数据源为：MongoDB、HBase、关系型数据库或Redis，其中Redis必须为Hash数据格式。
- 全部字段添加完之后，检查源端和目的端的字段映射关系是否正确，如果不正确可以拖拽字段调整字段位置。
 - 单击“下一步”后保存作业。

解决方法二：修改 JSON 文件

1. 获取源端HBase待迁移的表中所有的字段，列族与列之间用“:”分隔，例如：

```
rowkey:rowkey
g:DAY_COUNT
g:CATEGORY_ID
g:CATEGORY_NAME
g:FIND_TIME
g:UPLOAD_PEOPLE
g:ID
g:INFOMATION_ID
g:TITLE
g:COORDINATE_X
g:COORDINATE_Y
g:COORDINATE_Z
g:CONTENT
g:IMAGES
g:STATE
```

2. 在DWS目的表中，获取与HBase表对应的字段。

如果DWS目的表中没有HBase对应的字段名，需在DWS表定义中加上，假设DWS表中的字段齐全且如下：

```
rowkey
day_count
category
category_name
find_time
upload_people
id
infomation_id
title
coordinate_x
coordinate_y
coordinate_z
content
images
state
```

3. 在CDM的作业管理界面，找到HBase到DWS的作业，选择作业后面的“更多 > 编辑作业JSON”。
4. 在CDM界面编辑作业的JSON文件。

- a. 修改源端的“fromJobConfig.columns”参数，配置为1获取的HBase的字段，列号之间使用“&”分隔，列族与列之间用“:”分隔，如下：

```
"from-config-values": {
  "configs": [
    {
      "inputs": [
        {
          "name": "fromJobConfig.table",
          "value": "HBase"
        },
        {
          "name": "fromJobConfig.columns",
          "value":
"rowkey:rowkey&g:DAY_COUNT&g:CATEGORY_ID&g:CATEGORY_NAME&g:FIND_TIME&g:UP  
LOAD_PEOPLE&g:ID&g:INFOMATION_ID&g:TITLE&g:COORDINATE_X&g:COORDINATE_Y&g:  
COORDINATE_Z&g:CONTENT&g:IMAGES&g:STATE"
        }
      ],
      "name": "fromJobConfig.formats",
      "value": {
        "2": "yyyy-MM-dd",
        "undefined": "yyyy-MM-dd"
      }
    }
  ],
  "name": "fromJobConfig"
```

```
    }  
  ]  
}
```

- b. 修改目的端的“toJobConfig.columnList”参数，配置为2中DWS的字段列表。

这里的顺序必须与HBase保持一致，才能保证正确的字段映射关系，字段名之间使用“&”分隔，如下：

```
"to-config-values": {  
  "configs": [  
    {  
      "inputs": [  
        {  
          "name": "toJobConfig.schemaName",  
          "value": "dbadmin"  
        },  
        {  
          "name": "toJobConfig.tablePreparation",  
          "value": "DO_NOTHING"  
        },  
        {  
          "name": "toJobConfig.tableName",  
          "value": "DWS "  
        },  
        {  
          "name": "toJobConfig.columnList",  
          "value":  
"rowkey&day_count&category&category_name&find_time&upload_people&id&infomation_  
id&title&coordinate_x&coordinate_y&coordinate_z&content&iimages&state"  
        },  
        {  
          "name": "toJobConfig.shouldClearTable",  
          "value": "true"  
        }  
      ],  
      "name": "toJobConfig"  
    }  
  ]  
}
```

- c. 其他参数保持不变，单击“保存并运行”。
5. 作业完成后，查询DWS表中的数据是否和HBase中的数据匹配。如果不匹配，请检查JSON文件中HBase和DWS字段的顺序是否一致。

3.30 CDM 迁移数据到 DWS 时如何选取分布列？

在使用CDM迁移数据到数据仓库服务（DWS）或者FusionInsight LibrA，且CDM在DWS端自动创建一个新表时，在创建作业的字段的映射界面，需要选择分布列，如图3-15所示。

图 3-15 选取分布列

源字段				目的字段			
名称	样值	类型	操作	名称	类型	分布列	操作
COLUMN1	1	VARCHAR(50)		COLUMN1	VARCHAR(50)	<input type="checkbox"/>	
COLUMN2	LU	VARCHAR(50)		COLUMN2	VARCHAR(50)	<input type="checkbox"/>	
COLUMN3	15	VARCHAR(50)		COLUMN3	VARCHAR(50)	<input type="checkbox"/>	

由于分布列的选取，对于DWS/FusionInsight LibrA的运行非常重要，在CDM数据迁移到DWS/FusionInsight LibrA过程中，建议按如下顺序选取分布列：

1. 有主键可以使用主键作为分布列。
2. 多个数据段联合做主键的场景，建议设置所有主键作为分布列。
3. 在没有主键的场景下，如果没有选择分布列，DWS会默认第一列作为分布列，可能会有数据倾斜风险。

因此，在单表或整库导入到DWS/FusionInsight LibrA时，建议您在此处手动选择分布列，如果您没有选择，CDM会自动选择一个分布列。关于分布列的更多信息，请参见[数据仓库服务](#)。

DWS主键或表只有一个字段时，要求字段类型必须是如下常用的字符串、数值、日期类型。从其他数据库迁移到DWS时，如果选择自动建表，主键必须为以下类型，未设置主键的情况下至少要有有一个字段是以下类型，否则会无法创建表导致CDM作业失败。

- INTEGER TYPES: TINYINT, SMALLINT, INT, BIGINT, NUMERIC/DECIMAL
- CHARACTER TYPES: CHAR, BPCHAR, VARCHAR, VARCHAR2, NVARCHAR2, TEXT
- DATA/TIME TYPES: DATE, TIME, TIMETZ, TIMESTAMP, TIMESTAMPTZ, INTERVAL, SMALLDATETIME

3.31 迁移到 DWS 时出现 value too long for type character varying 怎么处理？

问题描述

在使用CDM迁移数据到数据仓库服务（DWS）或者FusionInsight LibrA时，如果迁移作业失败，且执行日志中出现“value too long for type character varying”错误提示，如图3-16所示。

图 3-16 日志信息

```
Caused by: org.postgresql.util.PSQLException: ERROR: value too long for type character varying(50)
  Where: COPY fl_behavior_module, line 72, column MODULE_NAME: "..."
  at org.postgresql.core.v3.QueryExecutorImpl.receiveErrorResponse(QueryExecutorImpl.java:2477)
  at org.postgresql.core.v3.QueryExecutorImpl.processCopyResults(QueryExecutorImpl.java:1107)
  at org.postgresql.core.v3.QueryExecutorImpl.writeToCopy(QueryExecutorImpl.java:989)
  at org.postgresql.core.v3.CopyInImpl.writeToCopy(CopyInImpl.java:35)
  ... 16 common frames omitted
```


原因分析

这种情况一般是在迁移到DWS时数据有中文，且创建作业时选择了目的端自动建表的情况下。原因是DWS的varchar类型是按字节计算长度，一个中文字符在UTF-8编码下可能要占3个字节。当中文字符的字节超过DWS的varchar的长度时，就会出现错误：value too long for type character varying。

解决方法

这个问题，可以通过将目的端作业参数“扩大字符字段长度”选择“是”来解决，选择此选项后，再创建目的表时会自动将varchar类型的字段长度扩大3倍。

编辑CDM的表/文件迁移作业，目的端作业配置下“自动创表”选择“不存在时创建”，则高级属性下面会出现参数“扩大字符字段长度”，配置该参数为“是”即可，如图3-17所示。

图 3-17 扩大字符字段长度



目的端作业配置

* 目的连接名称

* 模式或表空间

自动创表

* 表名

是否压缩 是 否

存储模式

导入前清空数据 是 否

隐藏高级属性

先导入阶段表 是 否

扩大字符字段长度 是 否

3.32 OBS 导入数据到 SQL Server 时出现 Unable to execute the SQL statement 怎么处理？

问题描述

使用CDM从OBS导入数据到SQL Server时，作业运行失败，错误提示为：Unable to execute the SQL statement. Cause : 将截断字符串或二进制数据。

原因分析

用户OBS中的数据超出了SQL Server数据库的字段长度限制。

解决方法

在SQL Server数据库中建表时，将数据库字段改大，长度不能小于源端OBS中的数据长度。

3.33 获取集群列表为空/没有权限访问/操作时报当前策略不允许执行？

问题描述

在使用CDM时，可能遇到如下权限相关的问题：

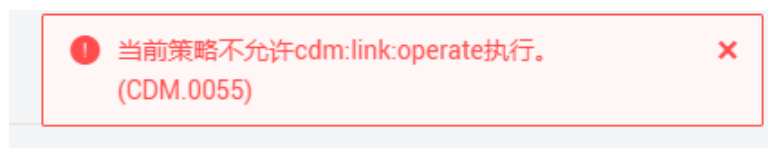
- 跳转到CDM首页，获取到的集群列表为空。
- 提示没有权限访问，如[图3-18](#)所示。
- 执行启动作业/重启集群等操作时，报错当前策略不允许执行，如[图3-19](#)所示。

图 3-18 没有权限访问



很抱歉，您没有访问权限。
请联系您的账号管理员开通权限。

图 3-19 不允许创建连接



原因分析

以上所列的问题均属于权限配置问题。

解决方法

- 如果是作为DataArts Studio服务CDM组件使用：
 - a. 检查用户是否添加DAYU Administrator或DAYU User角色，参考[DataArts Studio权限管理](#)。
 - b. 是否有对应工作空间的权限，如开发者、访客等，参考[DataArts Studio权限列表](#)。
- 如果是独立CDM服务使用：
 - a. 检查是否开启IAM细粒度鉴权
 - 如果未开启，检查用户组是否添加CDM Administrator角色。
 - 如果已开启，请继续执行[步骤2](#)继续检查。
 - b. 检查用户是否添加cdm访问策略，包含自定义策略或预设策略，如CDM FullAccess、CDM ReadOnlyAccess等，参考[CDM权限管理](#)。
 - c. 检查对应企业项目是否添加拒绝访问策略。

3.34 Oracle 迁移到 DWS 报错 ORA-01555

问题现象

使用CDM迁移Oracle数据至DWS，报错[图3-20](#)所示。

图 3-20 报错现象

```

665 2020-09-21 22:51:02,991 ERROR LocalJobRunner Map Task #3 [org.apache.sqoop.common.SqoopException:111] SqoopException
666 java.sql.SQLException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYS$SMU3_2097677531$" too small
667
668   at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:494)
669   at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:446)
670   at oracle.jdbc.driver.T4C8oall.processERROR(T4C8oall.java:1054)
671   at oracle.jdbc.driver.T4CTTIfun.receive(T4CTTIfun.java:623)
672   at oracle.jdbc.driver.T4CTTIfun.doRPC(T4CTTIfun.java:252)
673   at oracle.jdbc.driver.T4C8oall.doOALL(T4C8oall.java:612)
674   at oracle.jdbc.driver.T4CPreparedStatement.doOall8(T4CPreparedStatement.java:226)
675   at oracle.jdbc.driver.T4CPreparedStatement.fetch(T4CPreparedStatement.java:1023)
676   at oracle.jdbc.driver.OracleStatement.fetchMoreRows(OracleStatement.java:1353)
677   at oracle.jdbc.driver.InsensitiveScrollableResultSet.fetchMoreRows(InsensitiveScrollableResultSet.java:736)
678   at oracle.jdbc.driver.InsensitiveScrollableResultSet.absoluteInternal(InsensitiveScrollableResultSet.java:692)
679   at oracle.jdbc.driver.InsensitiveScrollableResultSet.next(InsensitiveScrollableResultSet.java:406)
680   at org.apache.sqoop.connector.jdbc.sql.impl.WrapResultSet.next(WrapResultSet.java:36)
681   at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extractObjectRecord(GenericJdbcExtractor.java:151)
682   at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:129)
683   at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:59)
684   at org.apache.sqoop.job.map.SqoopMapper.runInternal(SqoopMapper.java:184)
685   at org.apache.sqoop.job.map.SqoopMapper.run(SqoopMapper.java:81)
686   at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:799)
687   at org.apache.hadoop.mapred.MapTask.run(MapTask.java)
688   at org.apache.hadoop.mapred.LocalJobRunner$Job$MapTaskRunnable.run(LocalJobRunner.java:271)
689   at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
690   at java.util.concurrent.FutureTask.run(FutureTask.java:266)
691   at org.apache.sqoop.submission.mapreduce.MapperExecutorGroup$1.lambda$execute$0(MapperExecutorGroup.java:222)
692   at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
693   at java.util.concurrent.FutureTask.run(FutureTask.java:266)
694   at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
695   at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
696   at java.lang.Thread.run(Thread.java:748)
697   Caused by: oracle.jdbc.OracleDatabaseException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYS$SMU3_2097677531$" too small
698
699   at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:498)
700   ... 28 common frames omitted
    
```

原因分析

1. 数据迁移，整表查询且该表数据量大，那么查询时间较长。
2. 查询过程中，其他用户频繁进行commit操作。
3. Oracle的RBS(rollback space 回滚时使用的表空间)较小，造成迁移任务没有完成，源库已更新，回滚超时。

建议与总结

1. 调小每次查询的数据量。
2. 通过修改数据库配置调大Oracle的RBS。

3.35 MongoDB 连接迁移失败时如何处理？

在默认情况下，userAdmin角色只具备对角色和用户的管理，不具备对库的读和写权限。

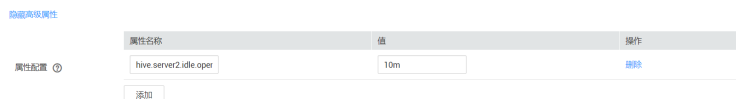
当用户选择MongoDB连接迁移失败时，用户需查看MongoDB连接中用户的权限信息，确保对指定库具备ReadWrite权限。

3.36 Hive 迁移作业长时间卡住怎么办？

为避免Hive迁移作业长时间卡住，可手动停止迁移作业后，通过编辑Hive连接增加如下属性设置：

- 属性名称：hive.server2.idle.operation.timeout
- 值：10m

如图所示：



3.37 使用 CDM 迁移数据由于字段类型映射不匹配导致报错怎么处理？

问题描述

在使用CDM迁移数据到数据仓库服务（DWS）时，迁移作业失败，且执行日志中出现“value too long for type character varying”错误提示。

原因分析

这种情况一般是源表与目标表类型不匹配导致，例如源端dli字段为string类型，目标端dws字段为varchar(50)类型，导致精度缺省，就会报：value too long for type character varying。类似的问题还有string转bigint，bigint转int。

解决方案

- 根据报错信息找到哪个字段映射有问题，找DBA修改表结构。
- 如果只有极少数据有问题，可以配置脏数据策略解决。

3.38 MySQL 迁移时报错“JDBC 连接超时”怎么办？

问题描述

MySQL迁移时报错：Unable to connect to the database server. Cause: connect timed out.

原因分析

这种情况是由于表数据量较大，并且源端通过where语句过滤，但并非索引列，或列值不离散，查询会全表扫描，导致JDBC连接超时。例如图3-21所示c_date字段为非索引列。

图 3-21 非索引列

解决方案

1. 优先联系DBA修改表结构，将需要过滤的列配置为索引列，然后重试。
如果由于数据不离散，导致还是失败请参考2~4，通过增大JDBC超时时间解决。
2. 根据作业找到对应的MySQL连接名称，查找连接信息。

图 3-22 连接信息

名称	连接信息
mysql2dli	mysql--dli

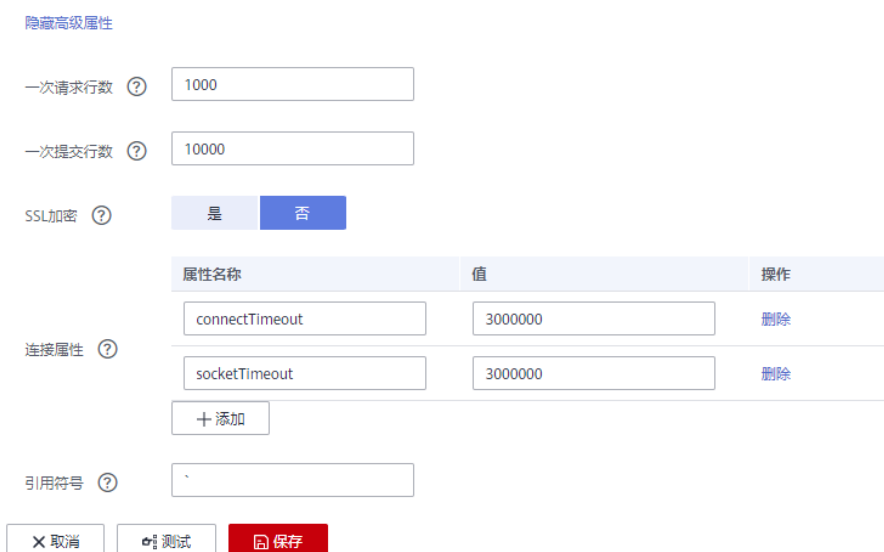
3. 单击“连接管理”，在“操作”列中，单击“连接”进行编辑。

图 3-23 连接



4. 打开高级属性，在“连接属性”中建议新增“connectTimeout”与“socketTimeout”参数及参数值，单击“保存”。

图 3-24 编辑高级属性



3.39 创建了 Hive 到 DWS 类型的连接，进行 CDM 传输任务失败时如何处理？

建议清空历史数据后再次尝试该任务。在使用 CDM 迁移作业的时候需要配置清空历史数据，然后再做迁移，可大大降低任务失败的概率。

3.40 如何使用 CDM 服务将 MySQL 的数据导出成 SQL 文件，然后上传到 OBS 桶？

CDM 服务暂不支持该操作，建议通过手动导出 MySQL 的数据文件，然后在服务器上开启 SFTP 服务，然后新建 CDM 作业，源端是 SFTP 协议，目的端是 OBS，将文件传过去。

3.41 如何处理 CDM 从 OBS 迁移数据到 DLI 出现迁移中断失败的问题？

此类作业问题表现为配置了脏数据写入，但并无脏数据。这种情况下需要调低并发任务数，即可避免此类问题。

3.42 如何处理 CDM 连接器报错“配置项 [linkConfig.iamAuth] 不存在”？

客户证书过期，需要完成更新证书操作，完成后重新配置连接器即可。

3.43 创建数据连接时报错“配置项 [linkConfig.createBackendLinks]不存在”或创建作业时报错“配置项 [throttlingConfig.concurrentSubJobs] 不存在”怎么办？

当同时存在多个不同版本的集群，先在低版本CDM集群创建数据连接或保存作业后，再进入高版本CDM集群时，会偶现此类故障。

需手动清理浏览器缓存，即可避免此类问题。

3.44 新建 MRS Hive 连接时，提示：CORE_0031:Connect time out. (Cdm.0523) 怎么解决？

新建MRS Hive连接时，提示无法下载配置文件，实际是用户权限不足。建议您新建一个业务用户，给对应的权限后重试即可。

如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。

📖 说明

- 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。
- 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。
- 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。

3.45 迁移时已选择表不存在时自动创表，提示“CDM not support auto create empty table with no column”怎么处理？

这是由于数据库表名中含有特殊字符导致识别出语法错误，按数据库对象命名规则重新命名后恢复正常。

例如，DWS数据仓库中的数据表命名需要满足以下约束：长度不超过63个字符，以字母或下划线开头，中间字符可以是字母、数字、下划线、\$、#。

3.46 创建 Oracle 关系型数据库迁移作业时，无法获取模式名怎么处理？

这是由于可能上传了暂不支持的最新ORACLE_8驱动（如Oracle Database 21c (21.3) drivers），推荐使用Oracle Database 12c中的ojdbc8.jar驱动（下载地址：<https://www.oracle.com/database/technologies/jdbc-ucp-122-downloads.html>）。

3.47 MySQL 迁移时报错：invalid input syntax for integer: "true"

问题描述

数据库中存储的是1或0，但没有true和false的数据，但MySQL迁移时读取到的是true或false，提示报错信息：Unable to execute the SQL statement. Cause: ERROR: invalid input syntax for integer: "true" Where: COPY sd_mask_ext, line 1, column mask_type.

原因分析

MySQL默认开启配置tinyInt1isBit=true，会将TINYINT(1)当作BIT也就是Types.BOOLEAN来处理，将1或0读取为true或false。

解决方案

在MySQL数据连接高级属性中，连接属性新增如下参数之一即可，这样就可以在目的端正常建表。

- “tinyInt1isBit”参数，参数值设为“false”。
- “mysql.bool.type.transform”参数，参数值设为“false”。

图 3-25 添加连接属性

隐藏高级属性

一次请求行数 ①

一次提交行数 ②

SSL加密 ③ 是 否

属性名称	值	操作
<input type="text"/>	<input type="text"/>	删除

连接属性 ④

+ 添加

4 数据架构

4.1 码表和数据标准有什么关系？

码表由多条表字段的名称+编码+数据类型组成，码表的表字段可以关联到数据标准上，数据标准会应用到某张模型表的字段上。

4.2 关系建模和维度建模的区别？

- 关系建模为事务性模型，对应三范式建模。
- 维度建模为分析性模型，主要包括事实表、维度表的设计，多用于实现多角度、多层次的数据查询和分析。

DataArts Studio是基于数据湖的数据运营平台，维度建模使用的场景比较多。

4.3 数据架构支持哪些数据建模方法？

DataArts Studio数据架构支持的建模方法有以下两种：

- **关系建模**

关系建模是用实体关系（Entity Relationship, ER）模型描述企业业务，它在范式理论上符合3NF，出发点是整合数据，将各个系统中的数据以整个企业角度按主题进行相似性组合和合并，并进行一致性处理，为数据分析决策服务，但是并不能直接用于分析决策。

用户在关系建模过程中，可以从以下三个层次去设计关系模型，这三个层次是逐层递进的，先设计概念模型，再进一步细化设计出逻辑模型，最后设计物理模型。

- **概念模型**：是从用户的视角，主要从业务流程、活动中涉及的主要业务数据出发，抽象出关键的业务实体，并描述这些实体间的关系。
- **逻辑模型**：是概念模型的进一步细化，通过实体、属性和关系勾勒出企业的业务信息蓝图，是IT和业务人员沟通的桥梁。逻辑数据模型是一组规范化的逻辑表结构，逻辑数据模型是根据业务规则确定的，关于业务对象、业务对象的数据项及业务对象之间关系的基本蓝图。

- **物理模型**：是在逻辑数据模型的基础上，考虑各种具体的技术实现因素，进行数据库体系结构设计，真正实现数据在数据库中的存放，例如：所选的数据仓库是DWS或DLI。
- **维度建模**

维度建模是从分析决策的需求出发构建模型，它主要是为分析需求服务，因此它重点关注用户如何更快速地完成需求分析，同时具有较好的大规模复杂查询的响应性能。

多维模型是由数字型度量值组成的一张事实表连接到一组包含描述属性的多张维度表，事实表与维度表通过主/外键实现关联。

典型的维度模型有星形模型，以及在一些特殊场景下使用的雪花模型。

在DataArts Studio数据架构中，维度建模是以维度建模理论为基础，构建总线矩阵、抽象出事实和维度，构建维度模型和事实模型，同时对报表需求进行抽象整理出相关指标体系，构建出汇总模型。

4.4 规范化的数据如何使用？

规范化的数据可以作为BI的基本信息，也可以作为上层应用的源数据，也可以接入各类数据可视化报表等。

4.5 数据架构支持逆向数据库吗？

数据架构支持逆向数据库，目前支持基于数据仓库服务（DWS）、数据湖探索（DLI）、MapReduce服务（MRS Hive）的数据库逆向。

4.6 数据架构中的指标与数据质量的指标的区别？

数据架构中指标侧重业务维度，用来衡量目标总体特征的统计数值；数据质量中指标侧重监控维度，用来管理所有业务指标，包括指标的来源、定义等。

注意，数据质量模块的指标与数据架构模块的业务指标、技术指标当前是相互独立的，不支持交互。

4.7 为什么关系建模或维度建模修改字段后，数据库中表无变化？

关系建模或维度建模修改字段更新表后，但实际上数据库中物理表并无变化，这是因为未对数据表更新方式做配置，此选项默认为“不更新”。

配置数据表更新方式操作如下：

1. 单击“数据架构 > 配置中心”。
2. 单击“功能配置”页签。
3. 配置“数据表更新方式”选择为“依据DDL更新模板”或“重建数据表”。
 - **不更新**：不更新数据库中的表。
 - **依据DDL更新模板**：依据DDL模板管理中配置的DDL更新模板，更新数据库中的表，但能否更新成功是由底层数仓引擎的支持情况决定的。由于不同类型的数仓支持的更新表的能力不同，在数据架构中所做的表更新操作，如果

数仓不支持，则无法确保数据库中的表和数据架构中的表是一致的。例如，DLI类型的表更新操作不支持删除表字段，如果在数据架构的表中删除了表字段，则无法在数据库中相应的删除表字段。

如果线下数据库支持更新表结构语法，可以在DDL模板配置对应语法，之后更新操作就可以通过DataArts Studio管控；如果线下数据库不支持更新，那只有通过重建这种方式更新。

- **重建数据表**：先删除数据库中已有的表，再重新创建表。选择该选项可以确保数据库中的表和数据架构中的表是一致的，但是由于会先删除表，因此一般建议只在开发设计阶段或测试阶段使用该选项，产品上线后不推荐使用该选项。

4. 单击“确定”，完成配置。

4.8 表是否可配置生命周期管理？

目前暂不支持表生命周期管理的配置。

5 数据开发

5.1 数据开发可以创建多少个作业，作业中的节点数是否有限制？

目前默认每个用户最多可以创建10000个作业，每个作业建议最多包含200个节点。

另外，系统支持用户根据实际需求调整最大配额。如有需求，请提交工单进行申请。

5.2 DataArts Studio 支持自定义的 Python 脚本吗？

支持。

5.3 作业关联的 CDM 集群删除后，如何快速修复？

CDM集群被删除后，作业中的关联信息会保留原配置。用户只需在CDM中新建同名集群和作业，作业将使用新的同名CDM集群和作业，同时提示用户原CDM集群和作业将被替代。

5.4 作业的计划时间和开始时间相差大，是什么原因？

如图所示，在作业监控页面查看作业运行记录时，发现作业的计划时间和开始时间相差较大。其中计划时间是作业预期开始执行的时间，即用户为作业配置的调度计划。开始时间是作业实际开始执行的时间。

图 5-1 问题示例图

状态	调度方式	计划时间	开始时间	结束时间	运行时间 (min)	操作
运行成功	正常调度	2018/10/09 16:50:00 GMT +08...	2018/10/09 17:50:08 GMT +08...	2018/10/09 17:59:28 GMT +08...	9.3	停止 继续执行 强制成功

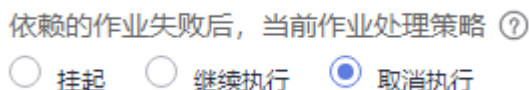
这是因为在数据开发中，单个作业最多允许5个实例并行执行，如果作业实际执行时间大于作业配置的调度周期，会导致后面批次的作业实例堆积，从而出现上述问题。

出现上述问题时，请检查作业配置的调度周期是否小于作业实际执行所需要的时间，根据实际情况调整作业的调度计划。

5.5 相互依赖的几个作业，调度过程中某个作业执行失败，是否会影响后续作业？这时该如何处理？

这种情况会影响后续作业，后续作业可能会挂起，继续执行或取消执行。

图 5-2 作业依赖关系



这时请勿停止作业，您可以将失败的作业实例进行重跑，或者将异常的实例停止再重跑。失败实例成功后，后续作业会继续正常运行。如果不通过数据开发，手动将作业实例中的业务场景处理后，可以强制成功作业实例，后续作业也会继续正常运行。

5.6 通过 DataArts Studio 调度大数据服务时需要注意什么？

DLI和MRS作为大数据服务，不具备锁管理的能力。因此如果同时对表进行读和写操作时，会导致数据冲突、操作失败。

如果您需要对大数据服务数据表进行读表和写表操作，建议参考以下方式之一进行串行操作处理：

- 将读表和写表操作拆分为同一作业的不同节点，两个节点通过连线建立先后执行关系，避免同时执行冲突。
- 将读表和写表操作拆分为两个不同的作业，两个作业之间设置依赖关系，避免同时执行冲突。

5.7 环境变量、作业参数、脚本参数有什么区别和联系？

环境变量、作业参数、脚本参数均可以配置参数，但作用范围不同；另外如果环境变量、作业参数、脚本参数同名冲突，调用的优先级顺序为：**作业参数 > 环境变量参数 > 脚本参数**。

环境变量、作业参数、脚本参数的介绍和使用方式如下：

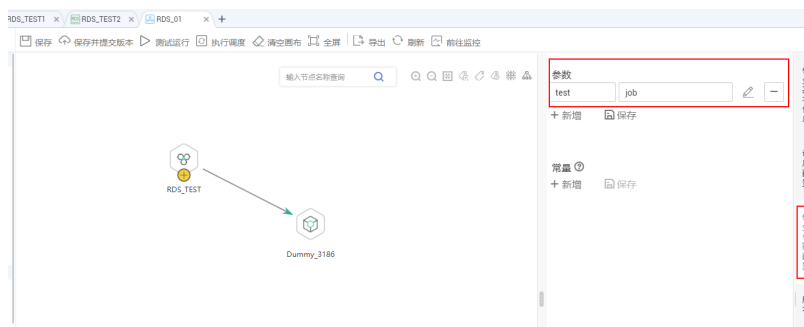
- 环境变量中支持定义变量和常量，环境变量的作用范围为当前工作空间。
 - 变量是指不同的空间下取值不同，需要重新配置值，比如“工作空间名称”变量，这个值在不同的空间下配置不一样，导出导入后需要重新进行配置。
 - 常量是指在不同的空间下都是一样的，导入的时候，不需要重新配置值。

图 5-3 环境变量



- 作业参数中支持定义参数和常量，作业参数的作用范围为当前作业。
 - 参数是指不同的作业下取值不同，需要重新配置值，导出导入后需要重新进行配置。
 - 常量是指在不同的作业下都是一样的，导入的时候，不需要重新配置值。

图 5-4 作业参数



- 脚本参数支持如下使用方式，脚本参数的作用范围为当前脚本。
 - SQL脚本支持在脚本编辑器中直接输入参数（Flink SQL不支持），脚本独立执行时可通过编辑器下方配置，如图5-5所示；通过作业调度时可通过节点属性赋值，如图5-6所示。
 - Shell脚本可以配置参数和交互式参数以实现参数传递功能。
 - Python脚本可以配置参数和交互式参数以实现支持参数传递功能。

图 5-5 独立执行时的脚本参数

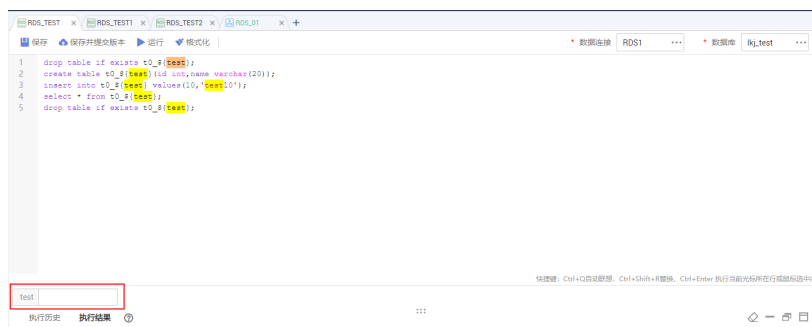
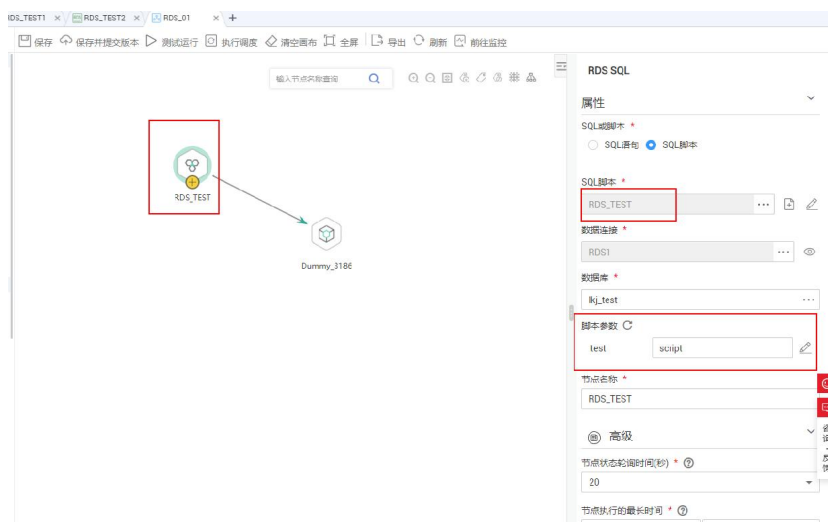


图 5-6 作业调度时的脚本参数



5.8 打不开作业日志，返回 404 报错?

作业日志在OBS桶中存储，您需要先在工作空间中配置作业日志的桶目录，然后确认当前账户是否具有OBS读权限（可以通过检查IAM中OBS权限、OBS桶策略来确认）。

📖 说明

OBS路径仅支持OBS桶，不支持并行文件系统。

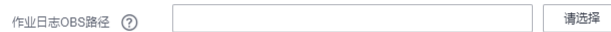
修改工作空间的用户账号，需要满足如下任一条件：

- DAYU Administrator或Tenant Administrator账号。
- DAYU User账号，但为当前工作空间的管理员。

配置作业日志的桶目录的步骤操作如下：

1. 登录DataArts Studio管理控制台。
2. 在“空间管理”页签，单击列表中相应工作空间后的“编辑”，弹出“空间信息”弹窗。
3. 在“空间信息”中，单击“作业日志OBS路径”后的“请选择”按钮，选择日志存储路径，可选择某个具体的目录。

图 5-7 修改日志存储路径



- 修改完成后，单击“确定”，即完成作业日志存储路径的修改。

说明

用户在创建作业时，会默认创建dlf-log-{projectID}命名的桶，此桶若存在，会跳过创建。

5.9 配置委托时获取委托列表失败如何处理？

当配置工作空间级或者作业级委托，查看委托列表时，报如下错误：

Policy doesn't allow iam:agencies:listAgencies to be performed.

则需要使用账号给当前用户添加“查看委托列表”的权限。

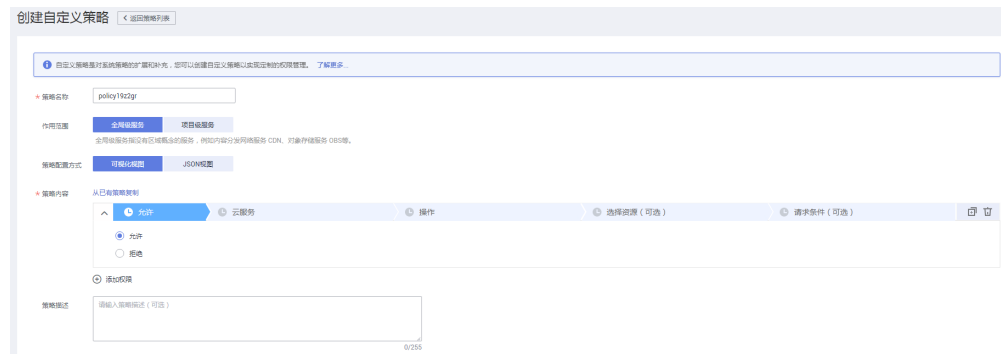
先创建自定义策略（查询指定条件下的委托列表），再通过给用户组授予自定义策略来进行精细的访问控制。

步骤1 登录华为云控制台。

步骤2 在控制台页面，鼠标移动至右上方的账号名，在下拉列表中选择“统一身份认证”。

步骤3 在左侧导航窗格中，单击“角色授权”>“创建自定义策略”。

步骤4 输入“策略名称”。



步骤5 选择“作用范围”，即自定义策略的生效范围，根据服务的部署区域选择，这里我们要授予的是IAM查询指定条件下的委托列表的权限。因IAM是全局级服务，所以作用范围选择“全局级服务”。

步骤6 “策略配置方式”选择“可视化视图”。

步骤7 在“策略内容”下配置策略。

- 选择“允许”。
- 选择“云服务”为“统一身份认证服务”。
- 选择“操作”，勾选产品权限（iam:agencies:listAgencies）。

步骤8 单击“确定”，自定义策略创建完成。

步骤9 参见[创建用户组并授权](#)，给当前用户所在的组添加**步骤7**中定义的策略。

步骤10 在左侧导航窗格中，单击“委托”，选择对应的委托，单击“授权”，将创建的自定义策略添加到该委托，单击“确定”。

当前用户退出系统，重新登录后，即可正常获取委托列表。

----结束

5.10 数据开发创建数据连接，为什么选不到指定的周边资源？

请确认当前DataArts Studio实例与周边资源在同一个Region且在同一个IAM项目下。如果账户开通企业项目，则还需在同一个企业项目下。

5.11 配置了 SMN 通知，却收不到作业失败告警通知？

如图，在“运维调度 > 通知管理”中配置了作业异常/失败的SMN通知，但却收不到作业失败的告警通知。

图 5-8 通知管理



此时可按以下步骤依次排查：

- 步骤1** 确认失败作业为调度中的作业。测试运行的作业是不发通知的，只有调度中的作业才会发SMN通知。
- 步骤2** 在“运维调度 > 通知管理”中查看此作业的通知配置开发是否为打开状态。
- 步骤3** 登录SMN页面，排查对应的SMN主题是否有被订阅。
- 步骤4** 排查对应SMN主题的订阅终端中是否有自己的终端名，还需确认订阅的状态是“已确认”。
- 步骤5** SMN通道异常。可通过在SMN界面中给自己的主题直接发送消息，判断能否收到SMN的通知。

----结束

5.12 作业配置了周期调度，但是实例监控没有作业运行调度记录？

1. 在“运维调度 > 作业监控”界面确认作业的调度状态是否是调度中，只有调度中的作业到了调度周期后才会调度。

图 5-9 查看作业调度状态



2. 如果作业有依赖于其他作业，在“运维调度 > 实例监控”界面，查看依赖作业的运行状态。如果作业有自依赖，扩大搜索时间窗口，查看是否当前作业历史实例失败，导致作业在等待运行，而没有生成新作业实例。

5.13 Hive SQL 和 Spark SQL 脚本执行失败，界面只显示执行失败，没有显示具体的错误原因？

请确认当前Hive SQL和Spark SQL脚本使用的数据连接为“直接连接”还是“通过代理连接”。

“直接连接”模式下DataArts Studio通过API把脚本提交给MRS，然后查询是否执行完成；而MRS不会将具体的错误原因反馈到DataArts Studio，因此导致数据开发脚本执行界面只能显示执行成功还是失败。

如果需要查看具体的错误原因，则需要到MRS的作业管理界面进行查看。

5.14 数据开发节点运行中报 TOKEN 不合法？

请确认当前用户在IAM的权限管理中权限是否有变更、是否退出用户组，或者用户所在的用户组权限策略是否有变更？

如果有变更，请重新登录即可解决。

5.15 作业开发时，测试运行后如何查看运行日志？

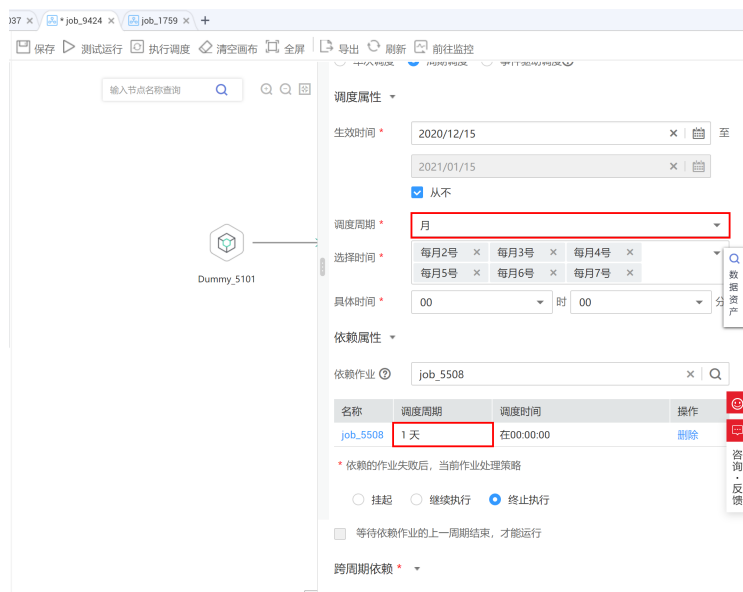
方式1：待节点测试运行完成后，在当前节点鼠标右键选择查看日志。

方式2：通过画布上方的“前往监控”，在实例监控中展开作业实例，查看节点日志。

5.16 月周期的作业依赖天周期的作业，为什么天周期作业还未跑完，月周期的作业已经开始运行？

如下图，月周期的作业依赖天周期的作业。为什么在天周期的作业还未跑完，月周期的作业已经开始运行？

图 5-10 查看作业调度周期及依赖属性



事实上，月周期的作业依赖天周期作业指的是当月的月周期作业是否运行取决于上月的天周期作业是否全部运行完成，而不是由当月的天周期作业决定。

例如在11月中，11月的月周期作业是否运行取决于10月的天周期作业是否全部运行完成。

5.17 执行 DLI 脚本，报 Invalid authentication 怎么办？

请确认当前用户在IAM中是否具有DLI Service User 或者 DLI Service Admin权限。

5.18 创建数据连接时，在代理模式下为什么选不到需要的 CDM 集群？

请确认CDM集群是否被关机。如果关机，请重新启动。

5.19 作业配置了每日调度，但是实例没有作业运行调度记录？

问题描述

作业配置了每日调度，但是实例没有作业运行调度记录。

原因分析

原因1：确认作业是否启动调度，如果没有启动，不会进行调度。

原因2：实例查询时间区间过大，如果配置有依赖作业或者自依赖，查看历史作业实例是否因为依赖失败，导致等待运行，没有生成新作业实例。

解决方案

配置作业失败异常告警通知，以及实例超时时间，当等待时间超过实例超时时间，系统将发送告警通知。

5.20 查看作业日志，但是日志中没有内容？

问题描述

查看作业日志，日志中没有内容。

原因分析

已在工作空间中配置作业日志的桶目录的前提下，确认用户在IAM中的OBS权限是否具有对象存储服务（OBS）的全局权限，保证用户能够创建桶和操作桶。

解决方案

方式1：用户在对象存储OBS中创建以“dlf-log-{projectID}”命名的桶，并将操作权限赋予调度用户。

说明

OBS路径仅支持OBS桶，不支持并行文件系统。

方式2：在IAM用户权限中增加全局OBS管理员权限。

5.21 创建了 2 个作业，但是为什么无法建立依赖关系？

问题描述

创建2个作业，但是无法建立依赖关系。

原因分析

查看所创建的2个作业的调度周期，确认这2个作业是否均为周调度作业或者月调度作业。目前不支持同周期调度，即周依赖周或者月依赖月的作业，不支持建立依赖关系。

解决方案

如果这2个作业是周依赖周或者月依赖月的作业，可以把这2个作业放到同一个画布中再运行。

5.22 DataArts Studio 执行调度时报错：提示作业没有可以提交的版本怎么办？

问题描述

DataArts Studio执行调度时报错：作业没有已提交的版本，请先提交作业版本。

原因分析

该作业还没有提交版本，就开始执行调度，导致执行调度报错。作业执行调度前必须保证作业存在一个版本。

解决方案

1. 提交作业（不是脚本）版本。
2. 执行作业调度。

图 5-11 提交版本



5.23 DataArts Studio 执行调度时报错：作业中节点 XXX 关联的脚本没有提交的版本？

问题描述

DataArts Studio执行调度时报错：作业中节点XXX关联的脚本没有提交的版本。

原因分析

该作业内的脚本还没有提交版本，就开始执行调度，导致执行调度报错。作业调度前必须保证作业内脚本都存在一个版本。

解决方案

1. 切换到脚本开发，找到对应脚本。
2. 提交脚本版本。
3. 执行作业调度。

5.24 提交调度后的作业执行失败，报 depend job [XXX] is not running or pause 怎么办？

问题描述

提交调度后的作业执行失败，报depend job [XXX] is not running or pause。

原因分析

该问题是由于上游依赖作业不在运行状态而造成。

解决方案

查看上游依赖作业，如果上游依赖的作业不在运行状态中，将这些作业重新执行调度即可。

5.25 如何创建数据库和数据表，数据库对应的是不是数据连接？

数据库和数据表可以在DLI服务中创建。

数据库对应的不是数据连接，数据连接是创建DataArts Studio和其他数据服务的连接通道。

5.26 为什么执行完 HIVE 任务什么结果都不显示？

解决方案：清理缓存数据，采用直连方式，数据就可以显示出来了。

5.27 在作业监控页面里的“上次实例状态”只有运行成功、运行失败，这是为什么？

上次实例状态是作业已经执行完成，只有成功、失败；实例监控里面状态有取消、暂停等好几种，是因为展示了作业的所有状态，另外作业运行异常和错误都会是作业失败的状态。

5.28 如何创建通知配置对全量作业都进行结果监控？

1. 在“运维调度->作业监控”中，选择“批作业监控”页签。
2. 勾选需要配置的作业，单击“通知配置”。

图 5-12 创建通知配置



3. 设置通知配置参数，单击“确定”完成作业的通知配置。

5.29 数据开发的并行执行节点数是多少？

DataArts Studio的并行执行节点数与作业节点调度次数/天配额有关，对应关系如下表所示。

其中的作业节点调度次数/天配额可通过DataArts Studio实例卡片上的“更多 > 配额使用量”入口查看，其中的“作业节点调度次数/天”总量即为当前实例配额。

表 5-1 DataArts Studio 实例并行节点数上限

DataArts Studio实例作业节点调度次数/天配额	DataArts Studio实例并行节点数上限
<=500	10
<=5000	50
<=20000	100
<=40000	200
<=80000	300
> 80000	400

当前工作空间级别的节点并发数支持配置，方法如下：

配置方法

- 步骤1** 参考[访问DataArts Studio实例控制台](#)登录DataArts Studio管理控制台。
- 步骤2** 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
- 步骤3** 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。
- 步骤4** 选择“节点并发数”。
- 步骤5** 配置工作空间的节点并发数，工作空间的节点并发数不能大于DataArts Studio实例的并行节点并发数上限。

DataArts Studio实例的节点并发数上限可通过[表5-2](#)获取。其中的作业节点调度次数/天配额可通过DataArts Studio实例卡片上的“更多 > 配额使用量”入口查看，其中的“作业节点调度次数/天”总量即为当前实例配额。

表 5-2 DataArts Studio 实例并行节点数上限

DataArts Studio实例作业节点调度次数/天配额	DataArts Studio实例并行节点数上限
<=500	10
<=5000	50
<=20000	100
<=40000	200
<=80000	300
> 80000	400

图 5-13 配置节点并发数



步骤6 单击“保存”，完成配置。

----结束

查看历史节点并发数

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“节点并发数”。

步骤3 在历史节点并发数界面，选择历史时间段。

步骤4 单击“确定”。

📖 说明

查看历史节点并发数的时间区间最大为24小时。

----结束

5.30 DataArts Studio 是否支持修改时区？

DataArts Studio实例暂不支持修改时区。

数据开发作业调度时可通过EL表达式适配当地时间，例如：

```
#{DateUtil.format(DateUtil.addHours(Job.planTime,-7),"yyyy-MM-dd")}
```

5.31 CDM 作业改名后，在数据开发中如何同步？

CDM作业改名后，需要在数据开发作业的CDM节点属性中，重新选择改名后的CDM作业名称。

5.32 执行 RDS SQL，报错 hll 不存在，在 DataArts Studio 可以执行成功？

hll插件默认创建在public schema，SQL需要带上hll所属的schema。

5.33 创建 DWS 数据连接时报错提示：The account has been locked?

连接DWS集群输入密码错误的次数达到集群参数failed_login_attempts所设置的值（默认10）时，账户将会被自动锁定。解锁方式参考[账号被锁住了，如何解锁？](#)

5.34 作业实例取消了，日志提示：The node start execute failed, so the current node status is set to cancel.

依赖的作业有失败的，实例监控的状态取消右边有个问号，单击查看依赖作业的失败实例。

5.35 调用数据开发接口报错，Workspace does not exists?

代码的request请求的header要添加项目Id，即header.add("X-Project-Id",项目Id)。

5.36 Postman 调用接口返回结果正常，为什么测试环境调用接口的 URL 参数不生效？

URL的参数连接符&需要转义。

5.37 执行 Python 脚本报错：Agent need to be updated?

创建的主机连接需要使用2.8.6版本及以上的CDM集群。

5.38 节点状态为成功，为什么日志显示运行失败？

强制成功操作会更新作业实例（和节点）状态为成功。

5.39 调用数据开发 API 报错 Unknown Exception?

DataArts Studio是项目级服务，获取Token的scope要选择project级别。

5.40 调用创建资源的 API 报错“资源名不合法”是什么原因？

资源名称只能包含英文字母、数字、中文字符、下划线或中划线，且长度为1-32个字符。

5.41 补数据的作业实例都是成功的，为什么补数据任务失败了？

补数据任务包含了其他工作空间的作业，可以在别的工作空间查看同名补数据任务的作业实例执行（失败）情况。

5.42 DWS 数据连接可视化建表，报错提示“表已存在”，但是展开数据连接看不到该表？

DWS数据连接的用户没有该表的查看与编辑权限，实际上该表已经存在。

5.43 调度 MRS spark 作业报错 The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute.

DataArts Studio作业调度MRS spark作业报错：“The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute”，如下图所示。

图 5-14 报错信息

```
    },
    "cluster_id":"2a4e57fe-f45a-48dd-9be6-dbca7e5963c3"
  }
[2020/12/31 04:45:11 GMT+0800] [DEBUG] =====
[2020/12/31 04:45:12 GMT+0800] [DEBUG] =====
[2020/12/31 04:45:12 GMT+0800] [INFO] mrs response of submitting job
[2020/12/31 04:45:12 GMT+0800] [INFO]
{
  "error_msg":"The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute",
  "error_code":"APIGW.0308",
  "request_id":"3c525a73feda9e1ff5b0b753b9a5409b"
}

[2020/12/31 04:45:12 GMT+0800] [DEBUG] =====
[2020/12/31 04:45:12 GMT+0800] [INFO] Submit MRS job failed, Retry times: 1
[2020/12/31 04:45:17 GMT+0800] [DEBUG] =====
[2020/12/31 04:45:17 GMT+0800] [INFO] try to submit job to mrs, request body is
[2020/12/31 04:45:17 GMT+0800] [INFO]
{
  "job_name":"mrs_dwsapp_fcgdb_fict",
```

由于MRS服务的接口限制了单个用户每分钟最多调用60次，因此只能通过降低调用频率来解决该问题。

5.44 执行 Python 脚本，报错 UnicodeEncodeError : 'ascii' codec cant encode characters in position 63-64 : ordinal not in range (128)

在DataArts Studio的python脚本中，设置参数json.dumps(json_data, ensure_ascii=False)时，执行报错UnicodeEncodeError : 'ascii' codec cant encode characters in position 63-64 : ordinal not in range (128)，如下图所示。

图 5-15 报错信息



```
1
2 import sys
3 import json
4
5
6 dict01 = {"id":1,"name":"张三"}
7 a = json.dumps(dict01,ensure_ascii=False)
8 print(a)
9 print("张三")
10 print(sys.getdefaultencoding())
foreach
31
```

File "/tmp/ea8e4463-6023-4407-b4bf-fabef7d622e9.py", line 8, in <module>
print(json.dumps(json_data, ensure_ascii=False))
UnicodeEncodeError: 'ascii' codec can't encode characters in position 63-64: ordinal not in range(128)

原因分析

DataArts Studio默认用的python2的解释器，python2默认是编码格式是ASCII编码，因ASCII编码不能编码汉字所以报错。因此需要将编码格式转化为“utf8”。

解决方法

1. 用python3解释器，在主机上做一个软连接，如下图所示。

图 5-16 主机上做软连接

```
[root@ecs-dws ~]# rm /bin/python
rm: remove symbolic link '/bin/python'? y
[root@ecs-dws ~]# ln -s /bin/python3.6 /bin/python
[root@ecs-dws ~]# ll /bin/python*
lrwxrwxrwx 1 root root 14 Oct 26 11:34 /bin/python -> /bin/python3.6
lrwxrwxrwx 1 root root 18 Sep 20 07:14 /bin/python2 -> /usr/bin/python2.7
-rwxr-xr-x 1 root root 7144 Nov 16 2020 /bin/python2.7
-rwxr-xr-x 2 root root 11328 Nov 16 2020 /bin/python3.6
-rwxr-xr-x 2 root root 11328 Nov 16 2020 /bin/python3.6m
lrwxrwxrwx 1 root root 7 Feb 26 2021 /bin/python.backup -> python2
[root@ecs-dws ~]#
```

2. 在文件开始标准编码方式：
-*- coding: utf-8 -*-; 或者设置主机的编码格式：在python安装目录的Lib\site-packages文件夹下新建一个sitecustomize.py文件，在文件中写入：
encoding=utf8
#import sys

- ```
#reload(sys)
#sys.setdefaultencoding('utf8')
```
3. 重启python，通过sys.getdefaultencoding()查看默认编码，这时为'utf8'。

## 5.45 查看日志时，系统提示“OBS 日志文件不存在，请检查文件是否被删除或者没有 OBS 写入权限。”怎么办？

### 问题现象

查看数据开发的节点日志时，系统提示“OBS日志文件不存在，请检查文件是否被删除或者没有OBS写入权限”，如下图所示：

图 5-17 提示信息



### 原因分析

数据开发的日志存储在OBS桶中，您所在的用户组没有OBS的操作权限，导致在查看节点日志时系统提示报错，或者OBS日志文件不存在时系统提示报错。

### 解决方法

1. 管理员登录IAM控制台。
2. 在统一身份认证服务的左侧导航窗格中，选择“用户”，单击用户名进入用户信息界面。
3. 查看用户所属的用户组。

图 5-18 用户所属的用户组



4. 在左侧导航窗格中，选择“用户组”，单击用户所属的用户组后面“操作”列的“授权”。
5. 在授权界面，选择需要给用户组添加的权限，搜索需要的权限名称，请配置为OBS OperateAccess或OBS Administrator。

图 5-19 给用户组授权



6. 单击“下一步”，选择最小授权范围，系统默认“所有资源”。
7. 单击“确定”。

如果权限没有问题，请检查OBS日志文件是否存在。

### 运行作业后查看日志时系统提示“OBS 日志文件不存在，请检查文件是否被删除或者没有 OBS 写入权限”的处理方法


1. 管理员登录IAM控制台。
2. 在统一身份认证服务的左侧导航窗格中，选择“用户”，单击用户名进入用户信息界面。
3. 单击“访问方式”后面的 ，修改访问方式。
4. 勾选“编程访问”和“管理控制台访问”。

图 5-20 配置访问方式



5. 单击“确定”。

#### 须知

- 在管理控制台创建工作空间时，作业日志OBS路径只支持OBS对象桶，不支持并行文件系统。如果不配置作业日志OBS路径，DataArts Studio数据开发默认会把日志写到dlf-log-{projectId}桶中，DataArts Studio数据服务默认会把日志写到dlm-log-{projectId}桶中。
- 如果“作业日志OBS路径”没有选择已有的OBS桶，首次运行作业时，默认的DLF桶创建不出来，无法写入日志。为了确保作业日志正常写入OBS桶中，当创建工作空间时，请选择已有的OBS路径。

## 5.46 Shell/Python 节点执行失败，后台报错 session is down

本指导以Shell算子为例。

### 问题背景与现象

Shell节点运行失败了，实际上Shell脚本运行成功了。

| 名称                                | 类型           | 状态   | 运行时间 (min) | 开始时间                          | 失败重试次数(次) | 错误信息 |
|-----------------------------------|--------------|------|------------|-------------------------------|-----------|------|
| query_domain_id                   | DLJ SQL      | 运行成功 | 0.45       | 2021/11/17 02:00:08 GMT+08:00 | 0         | --   |
| ods_ops_cloudroc_vision_require_d | DLJ SQL      | 取消   | 0.01       | 2021/11/17 02:05:56 GMT+08:00 | 0         | --   |
| get_vision_data                   | Shell Script | 失败   | 5.33       | 2021/11/17 02:00:36 GMT+08:00 | 0         | --   |

### 原因分析

1. 获取Shell节点的运行日志。

```
[2021/11/17 02:00:36 GMT+0800] [INFO] No job-level agency is set, Workspace-level agency is
dlg_agency, Execute job use agency dlg_agency, job id is
07572F197E4642E5BE549C2B656F157Ctm7cHkHd
[2021/11/17 02:00:36 GMT+0800] [DEBUG]
=====
[2021/11/17 02:00:36 GMT+0800] [INFO] Get response from agent when try to submit shell running
job :
[2021/11/17 02:00:36 GMT+0800] [INFO]
{
 "jobResultList":[
 {
 "jobId":"a567f7f5-3c9e-4dfc-a464-bd477ac5b1ea",
 "status":"created",
 "errorCode":0,
 "failCount":0,
 "result":[
]
 }
],
 "agentId":"614853ee-c1c6-456d-9aa6-fc84ad1281ed"
}
[2021/11/17 02:00:36 GMT+0800] [DEBUG]
=====
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
=====
[2021/11/17 02:05:56 GMT+0800] [INFO] Job Run finish , the raw output is :
[2021/11/17 02:05:56 GMT+0800] [INFO]
{
 "jobId":"a567f7f5-3c9e-4dfc-a464-bd477ac5b1ea",
 "status":"failed",
 "errorCode":3427,
 "errorMessage":"Shell script job execute failed.",
 "failCount":0,
 "result":[
 {
 "is_success":false,
 "exeTime":300.609
 }
]
}
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
=====
```

```
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
=====
[2021/11/17 02:05:56 GMT+0800] [INFO] The return code is : [-1].
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
=====
[2021/11/17 02:05:56 GMT+0800] [INFO] Execute shell script job finished.
[2021/11/17 02:05:56 GMT+0800] [ERROR] Shell exit code is not 0
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
=====
[2021/11/17 02:05:56 GMT+0800] [ERROR] Shell script job execute failed. Please contact ECS
Service.
[2021/11/17 02:05:56 GMT+0800] [ERROR] Exception message: RuntimeException: Shell script job
execute failed. Please contact ECS Service.
[2021/11/17 02:05:56 GMT+0800] [ERROR] Root Cause message:RuntimeException: Shell script job
execute failed. Please contact ECS Service.
```

2. 确认其ECS的sshd\_config参数如下。

```
ClientAliveInterval 300
ClientAliveCountMax 0
```

原因分析：由于ssh session超时断开了，因此Shell节点失败。

## 解决办法

1. 编辑ECS的/etc/ssh/sshd\_config文件，添加或者更新如下两个值。

```
ClientAliveInterval 300
ClientAliveCountMax 3
```

### 📖 说明

ClientAliveInterval指定了服务器端向客户端请求消息的时间间隔，默认是0，不发送请求。然而ClientAliveInterval 300表示五分钟发送一次，然后客户端响应，这样就保持长连接了。ClientAliveCountMax的默认值3。ClientAliveCountMax表示服务器发出请求后客户端没有响应的次数达到一定值，就自动断开，正常情况下，客户端会正常响应。

2. 修改后，重启ECS的sshd，执行如下命令：

```
[root@kwephisprc10123 ssh]# service sshd restart
Redirecting to /bin/systemctl restart sshd.service
```

3. 检查sshd是否启动成功（下图为成功）：

```
Redirecting to /bin/systemctl status sshd.service
ssh.service - OpenSSH server daemon
Loaded: loaded (/usr/lib/systemd/system/ssh.service; enabled; vendor preset: enabled)
Active: active (running) since Wed 2021-11-17 17:14:27 CST; 2min 54s ago
Docs: man:sshd(8)
 man:sshd_config(5)
Main PID: 24384 (sshd)
Tasks: 1 (limit: 26213)
Memory: 904.0K
CGroup: /system.slice/ssh.service
└─24384 /usr/sbin/sshd -D

Nov 17 17:14:27 kwephisprc10123 systemd[1]: Starting OpenSSH server daemon...
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: /etc/ssh/sshd_config line 154: Deprecated option RSAAuthentication
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: /etc/ssh/sshd_config line 156: Deprecated option RhostsRSAAuthentication
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: Server listening on 0.0.0.0 port 22.
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: Server listening on :: port 22.
Nov 17 17:14:27 kwephisprc10123 systemd[1]: Started OpenSSH server daemon.
```

## 5.47 请求头中参数值长度超过 512 个字符时，何如处理？

以Rest Client算子为例。

## 问题现象

在配置作业算子参数时，在添加请求头中时，需要输入参数及参数值，如果该参数的参数值长度超过512个字符时，则不能继续输入，如下图所示。

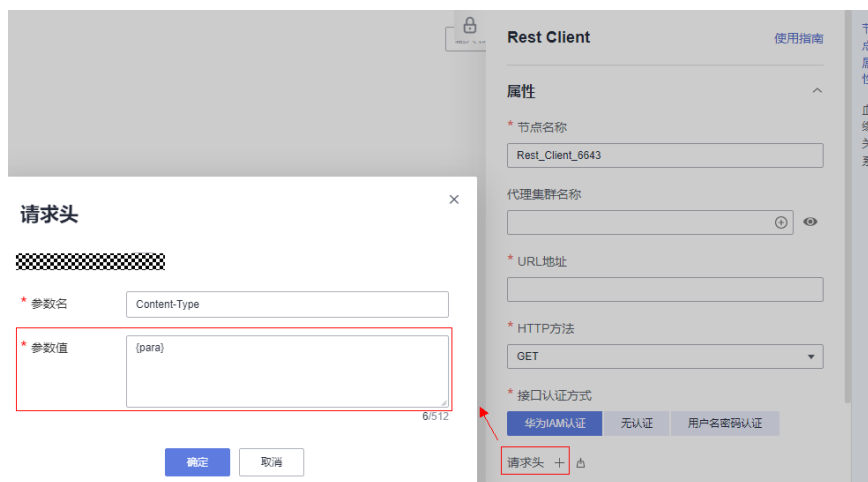
图 5-21 配置请求头参数



## 处理方法

- 配置作业节点的请求头参数。  
在“参数值”里面引入变量名称，例如{para}。

图 5-22 配置请求头参数



- 配置作业参数。
  - 单击“作业参数配置”，进入“作业参数配置”界面。
  - 在“变量”里面输入该变量para和变量值。该值的大小不能超过1024个字符。



图 5-23 配置作业参数



按照上述方法就可以解决请求头参数值输入的长度问题。

## 5.48 执行 DWS SQL 脚本时，提示 id 不存在，如何处理？

在执行DWS SQL脚本时，提示id不存在，原因是由于id的大小写引起的。

DWS执行SQL时，系统默认是小写，如果是大写字段需要加""。

举例：select \* from *table1* order by "ID";

```
select * from table order by "ID";
```

## 5.49 如何查看 CDM 作业被哪些作业进行调用？

1. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
2. 单击“批作业监控”页签，进入批作业的监控页面。
3. 通过条件筛选，查询CDM作业的调度执行信息。

### 📖 说明

通过筛选“CDM作业”可以查询CDM作业的调度执行信息。

通过筛选“节点类型 > CDMJob”可以查询CDMJob节点算子的调度执行信息。



```
def execute_script(ak, sk, endpoint, project_id, script_name, wp_id):
 try:
 print("执行脚本%s开始" %script_name)
 sig = signer.Signer()
 sig.Key = ak
 sig.Secret = sk

 post_url = "%s/v1/%s/scripts/%s/execute" % (endpoint, project_id, script_name)
 print("请求url:%s" %post_url)
 #调用脚本的输入参数
 post_data = """"{"params": {"tableVar": "citys","time": "2019-07-25"}}""

 r = signer.HttpRequest("POST", post_url)
 r.headers = {
 "content-type": "application/json;charset=utf-8",
 "workspace": wp_id
 }
 # r.body = json.dumps(post_data)
 r.body = post_data
 sig.Sign(r)
 print("请求头:%s" %(r.headers))
 print("请求body:%s" %(r.body))

 resp = requests.request(r.method, r.scheme + "://" + r.host + r.uri, headers=r.headers, data=r.body,
 verify=False)
 if resp.status_code == 200:
 instanceId = resp.json().get("instanceId");
```

修改参数Content-Type的参数类型后可以执行成功。

```
执行脚本api_call_test开始
请求url:https://dayu-dlf.cn-southwest-204-dev.myhuaweicloud.com/v1/.../scripts/api_call_test/execute
请求头: {'content-type': 'application/json;charset=utf-8', 'workspace': '6c147b4623bd4317b1d497bffb95fc2d', 'X-Sdk-Date': '20230824T084405Z', 'host': 'dayu-dlf.cn-southwest-204-dev.myhuaweicloud.com', 'Authorization': 'SDK-HMAC-SHA256 Access=Q5HCPDSNZ0CZUYWSN411, SignedHeaders=content-type;host;workspace;x-sdk-date, Signature=dd1d7bffb089c08855867c36dfb3020427e9ccb89e5a1ac263cd48fe23215980', 'content-length': '54'}
请求body: b'{"params": {"tableVar": "citys","time": "2019-07-25"}}'
脚本执行成功,生成实例id: 91c915ce-8da4-4794-858f-dc5e49b3dc50
```

# 6 数据质量

## 6.1 质量作业和对账作业有什么区别？

- 质量作业可将创建的规则应用到建好的表中进行质量监控。
- 对账作业支持跨源数据对账能力，可将创建的规则应用到两张表中进行质量监控，并输出对账结果。

数据对账对于数据开发和数据迁移流程中的数据一致性至关重要，而跨源数据对账的能力是检验数据迁移或数据加工前后是否一致的关键指标。

## 6.2 如何确认质量作业或对账作业已经阻塞？

作业运行状态长时间处于运行中时，选择“运维管理”，单击操作栏中的“结果&日志”并选择查看“运行日志”，当“运行日志”不再更新，表示作业已经阻塞。



```
2021-01-08 11:31:13 start instance execute...
2021-01-08 11:31:14 start auto scan data.
2021-01-08 11:31:14 finish auto scan data.
2021-01-08 11:31:14 generating sql...
2021-01-08 11:31:14 [select count(*) from ops_dw_odsdsdsds;]
2021-01-08 11:31:14 使用DLI引擎运行内置规则运行开始！
2021-01-08 11:31:15 [1385253c-ba94-4f55-8436-7810083896ad@ops_dw_ods.ods_biz_app_t_app_config]submit sql job process:1/1
2021-01-08 11:31:17 sub-rule custom-sql-rule:current 1 jobs need to check status, waiting...
2021-01-08 11:31:17 sub-rule 1385253c-ba94-4f55-8436-7810083896ad run failed!
2021-01-08 11:31:18 for detail:DLI-0005: Table or view not found: ops_dw_odsdsdsds; line 1 pos 21
2021-01-08 11:31:23 dirty data not found, stop dirty data event.
2021-01-08 11:31:24 log info:sub rule custom-sql-rule execute failed:null
2021-01-08 11:31:26 sub-rule 1385253c-ba94-4f55-8436-7810083896ad run failed!
2021-01-08 11:31:26 for detail:DLI-0005: Table or view not found: ops_dw_odsdsdsds; line 1 pos 21
```

## 6.3 如何手工重启阻塞的质量作业或对账作业？

阻塞的作业需要进行手工重启，如不重启1天内也会因作业超时自动结束该作业。

手工重启需要选择“运维管理”，先单击对应作业操作栏中的“取消”，作业运行状态变更为“失败”，此时然后单击操作栏中的“重跑”即可完成作业重启。



| 实例名称 | 类型   | 运行状态 | 操作 | 开始时间                          | 运行时间     | 操作            |
|------|------|------|----|-------------------------------|----------|---------------|
| 质量作业 | 质量作业 | 失败   | 重跑 | 2021/01/08 11:31:13 GMT+08:00 | 00:02:40 | 取消 详情&日志 处理日志 |

## 6.4 怎样查看质量规则模板关联的作业？

**步骤1** 单击待操作规则模板操作列的“发布历史”。

图 6-1 发布历史



**步骤2** 单击历史版本最右侧的“下线”按钮。则可以查看该规则模板对应的关联作业。

图 6-2 查看关联作业



----结束

## 6.5 用户在执行质量作业时提示无 MRS 权限怎么办？

用户在执行质量作业时报错，查看质量作业的日志，提示 “The current user does not exist on MRS Manager. Grant the user sufficient permissions on IAM and then perform IAM user synchronization on the Dashboard tab page. !”

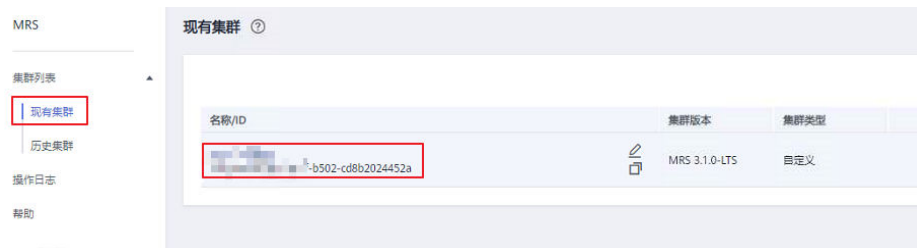
此类问题一般是由于用户不具备MRS集群操作权限导致的。

对于租户下新增的用户，需要在MRS集群列表的界面找到对应的MRS集群实例，手动单击同步。

操作如下：

**步骤1** 进入MRS控制台，查看现有集群，单击对应的集群名称进入概览页。

图 6-3 MRS 集群实例



**步骤2** 在“IAM用户同步”处，单击同步。

图 6-4 单击同步



**步骤3** 在操作日志处查看操作结果。

图 6-5 操作日志

| 操作类型 | 操作IP       | 操作内容                                                                                       |
|------|------------|--------------------------------------------------------------------------------------------|
| 集群操作 | 24.2.0.134 | 添加消息订阅规则, 集群ID为f6baa260-c4e4-47df-b502-cd8b2024452a, 规则名称为mrs, 主题名称为MRS。                   |
| 集群操作 | 25.0.0.50  | 集群f6baa260-c4e4-47df-b502-cd8b2024452a添加服务Flink。                                           |
| 数据操作 | 24.2.0.134 | 执行新增用户操作, 集群ID为f6baa260-c4e4-47df-b502-cd8b2024452a, 操作返回码为200, 操作详情为Operation succeeded。! |

**步骤4** 如果经过上述步骤，账号已同步。但还是提示MRS权限不足的话，则需要登录到Manger管理页面中创建一个与当前主账号同名的账号。

**⚠ 注意**

在步骤4中，需要创建一个与当前主账号同名的账号。

----结束

# 7 数据目录

## 7.1 数据目录组件有什么用？

数据目录的核心是通过元数据采集任务，采集并展示企业的数据资产地图，包括所有的元数据信息和数据血缘关系。

## 7.2 数据目录支持采集哪些对象的资产？

数据目录目前支持采集的资产请参见[支持的数据源](#)。

## 7.3 什么是数据血缘关系？

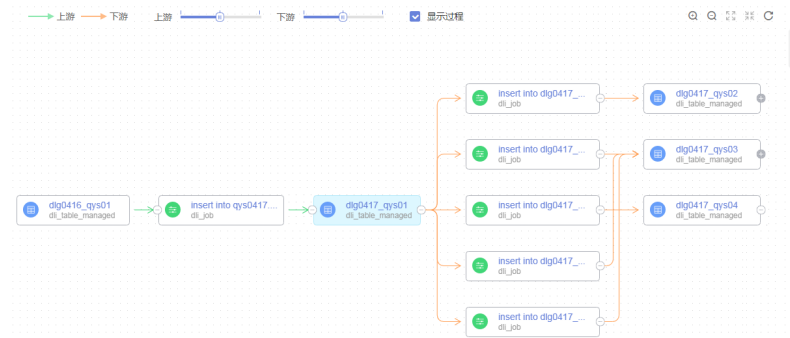
大数据时代，数据爆发性增长，海量的、各种类型的数据在快速产生。这些庞大复杂的数据信息，通过联姻融合、转换变换、流转流通，又生成新的数据，汇聚成数据的海洋。

数据的产生、加工融合、流转流通，到最终消亡，数据之间自然会形成一种关系。我们借鉴人类社会中类似的一种关系来表达数据之间的这种关系，称之为数据的血缘关系。与人类社会中的血缘关系不同，数据的血缘关系还包含了一些特有的特征：

- **归属性**：一般来说，特定的数据归属特定的组织或者个人，数据具有归属性。
- **多源性**：同一个数据可以有多个来源（多个父亲）。一个数据可以是多个数据经过加工而生成的，而且这种加工过程可以是多个。
- **可追溯性**：数据的血缘关系，体现了数据的生命周期，体现了数据从产生到消亡的整个过程，具备可追溯性。
- **层次性**：数据的血缘关系是有层次的。对数据的分类、归纳、总结等对数据进行的描述信息又形成了新的数据，不同程度的描述信息形成了数据的层次。



图 7-1 数据血缘关系示例



## 7.4 数据目录如何可视化展示数据血缘？

数据血缘在数据目录中展示，首先要进行元数据采集，其次需要有相关的作业调度。

数据血缘方案请参见[节点数据血缘](#)。

# 8 数据服务

## 8.1 数据服务 SDK 支持的语言？

数据服务SDK支持的语言有：C#、Python、Go、JavaScript、PHP、C++、C、Android、Java。

## 8.2 创建 API 时提示代理调用失败，怎么办？

需要在空余时间对CDM集群进行重启释放内存。

## 8.3 数据服务 API 接口，访问“测试 APP”，填写了相关参数，但是后台报错要如何处理？

在调用API时配置参数header parameter。

```
header parameter: x-Authorization, nvalid __ parameter: __,
```

## 8.4 使用 API 时报错，请问有什么办法可以解决？

使用API时需注意，每个子域名每天最多可以访问1000次。

## 8.5 API 传参是否支持传递操作符？

不支持传递操作符，传递的只是参数，操作符是固定的，多个参数可使用in({})方式。

## 8.6 数据服务专享版提供的 API 配额已满怎么解决？

如果数据服务专享版提供的API配额已满，无法创建新的API时可修改API配额，具体操作请参考[设置API分配配额](#)。

## 8.7 数据服务专享版发布的 API 如何绑定公网和域名？

- 步骤1** 进入数据服务模块，单击页面左侧“专享版”。
- 步骤2** 单击API管理，找到需要发布的API，选择操作列的“更多>发布”。
- 步骤3** 在弹出的发布页面中选择更多，选择网关类型为APIGW，即API网关服务。并选择分组。
- 步骤4** 发布完成后即可在对应的API网关服务中找到该API分组，在API分组详情页面选择“域名管理标签>绑定独立域名”，输入需要绑定的公网IP和域名。

---结束

## 8.8 如何处理 API 对应的数据表数据量较大时，获取数据总条数比较耗时的问题？

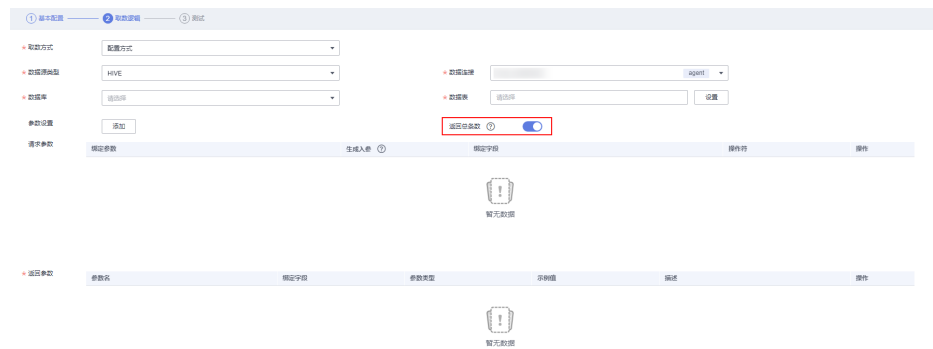
### 使用场景

当API对应的数据表数据量较大时，获取数据总条数比较耗时。在分页查询时，业务可通过参数（参数名use\_total\_num）控制后端是否计算并返回数据总条数。

### 前提条件

业务在创建API时，取数逻辑界面打开“返回总条数”开关。

图 8-1 返回总条数



### 参数说明

表 8-1 参数说明

|      |               |
|------|---------------|
| 参数名  | use_total_num |
| 是否必填 | 否，默认返回数据总条数   |
| 参数位置 | Query         |
| 参数类型 | String        |

|      |                        |
|------|------------------------|
| 参数说明 | 值为1返回数据总条数，值非1不返回数据总条数 |
|------|------------------------|

## 使用建议

分页查询场景下，第一次查询时添加入参use\_total\_num=1获取数据总条数，后续再次请求接口时添加入参use\_total\_num=0不获取数据总条数。

图 8-2 数据总条数

