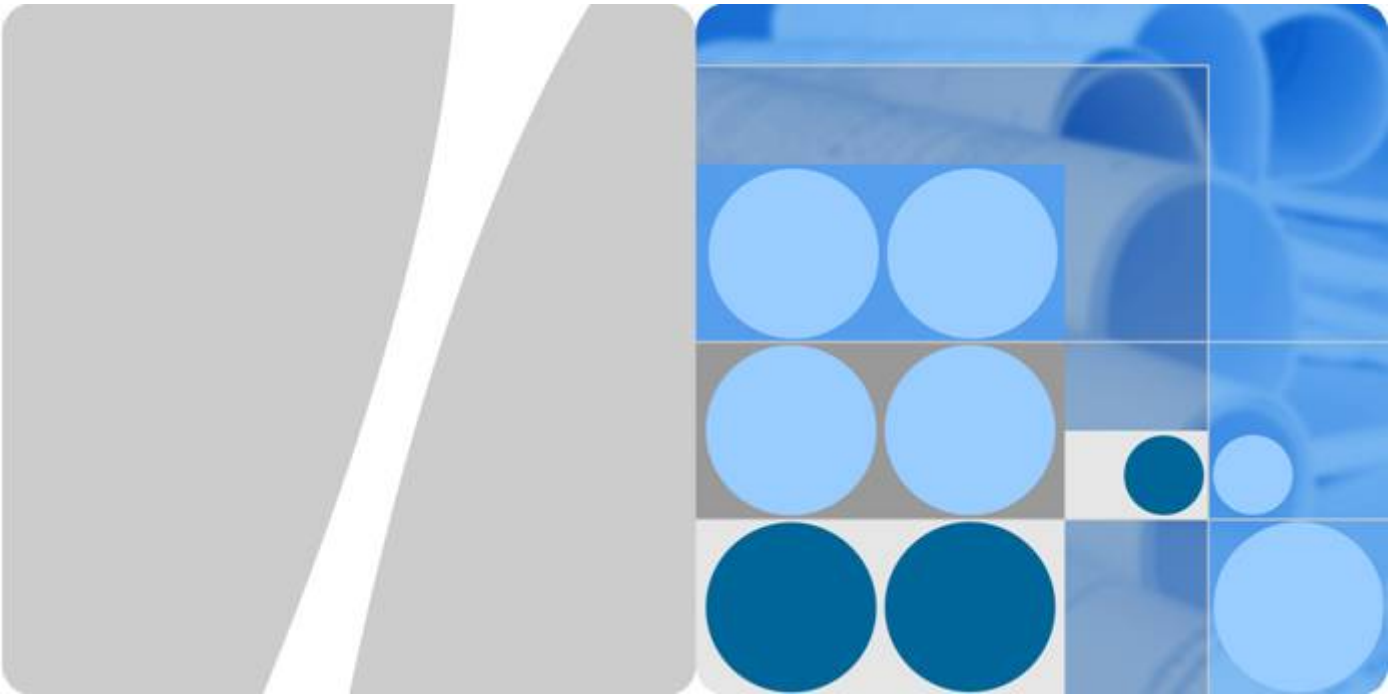


数据治理中心

用户指南

发布日期　2023-06-14

目 录

1 产品介绍

1.1 什么是数据治理中心 DataArts Studio

企业数字化转型面临的挑战

企业在进行数据管理时，通常会遇到下列挑战。

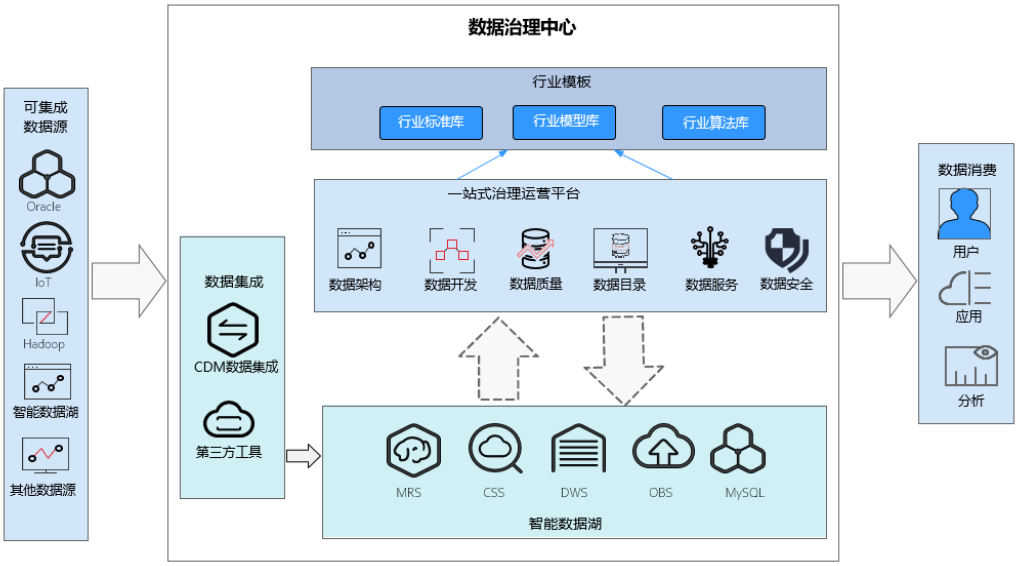
- 数据治理的挑战
  - 缺乏企业数据体系标准和数据规范定义的方法论，数据语言不统一。
  - 缺乏面向普通业务人员的高效、准确的数据搜索工具，数据找不到。
  - 缺乏技术元数据与业务元数据的关联，数据读不懂。
  - 缺乏数据的质量管控和评估手段，数据不可信。
- 数据运营的挑战
  - 数据运营效率低，业务环境的快速变化带来大量多样化的数据分析报表需求，因为缺乏高效的数据运营工具平台，数据开发周期长、效率低，不能满足业务运营决策人员的诉求。
  - 数据运营成本高，数据未服务化，导致数据拷贝多、数据口径不一致，同时数据重复开发，造成资源浪费。
- 数据创新的挑战
  - 企业内部存在大量数据孤岛，导致数据不共享、不流通，无法实现跨领域的数据分析与数据创新。
  - 数据的应用还停留在数据分析报表阶段，缺乏基于数据反哺业务推动业务创新的解决方案。

什么是 DataArts Studio?

数据治理中心DataArts Studio是为了应对上述挑战、针对企业数字化运营诉求提供的数据全生命周期管理、具有智能数据管理能力的一站式治理运营平台，包含数据集成、数据开发等功能，支持行业知识库智能化建设，支持大数据存储、大数据计算分析引擎等数据底座，帮助企业快速构建从数据接入到数据分析的端到端智能数据系统，消除数据孤岛，统一数据标准，加快数据变现，实现数字化转型。

产品架构如图1-1所示。

图 1-1 产品架构



如图所示，DataArts Studio基于数据湖底座，提供数据集成、开发、治理、开放等能力。DataArts Studio支持对接数据湖与数据库云服务作为数据湖底座，例如数据湖探索（Data Lake Insight，简称DLI）、MRS Hive、数据仓库服务DWS等，也支持对接企业传统数据仓库，例如Oracle、Greenplum等。

DataArts Studio包含如下功能组件：

- **管理中心**

  提供DataArts Studio数据连接管理的能力，将DataArts Studio与数据湖底座进行对接，用于数据开发等活动。

- **数据集成**

  数据集成提供20+简单易用的迁移能力和多种数据源到数据湖的集成能力，全向导式配置和管理，支持单表、整库、增量、周期性数据集成。

- **数据开发**

  大数据开发环境，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。支持数据建模、数据集成、脚本开发、工作流编排等操作，轻松完成整个数据的处理分析流程。

1.2 基本概念

DataArts Studio 实例

DataArts Studio实例是数据治理中心给用户提供的最小计算资源单位。数据治理中心以DataArts Studio实例的方式提供给用户，用户可以同时创建多个DataArts Studio实例，并分别管理和访问每个DataArts Studio实例。每个DataArts Studio实例具有用户指定的基础计算资源，包含管理中心、数据架构、数据集成、数据开发、数据质量、数据目录和数据服务七个模块。用户可根据业务需要申请相应规格的DataArts Studio实例。

工作空间

工作空间是从系统层面为管理者提供对使用DataArts Studio的用户（成员）权限、资源、DataArts Studio底层计算引擎配置的管理能力。

工作空间作为成员管理、角色和权限分配的基本单元，每个团队都可具有独立的工作空间。

您只有在加入工作空间并被分配权限后，才可具备管理中心数据开发和数据集成模块的系列操作权限。

成员和角色

成员是被授予工作空间访问或使用权限的。在添加工作空间成员时，您需要同时为添加的成员设置相应的角色。

角色是一组操作权限的集合。不同的角色拥有不同的操作权限，把角色授予成员后，成员即具有了角色的所有权限。每位成员至少要拥有一个角色，并且可以同时拥有多种角色。

数据集成

数据集成给用户提供的最小资源单位，一个数据集成集群运行在一个弹性云服务器之上，用户可以在集群中创建数据迁移作业，在云上和云下的同构/异构数据源之间批量迁移数据。

数据源

即数据的来源，本质是讲存储或处理数据的媒介，比如：关系型数据库、数据仓库、数据湖等。每一种数据源不同，其数据的存储、传输、处理和应用的模式、场景、技术和工具也不相同。

源数据

源数据强调数据状态是“创建”之后的“原始状态”，也就是没有被加工处理的数据。在数据管理的过程中，源数据一般是指直接来自源文件（业务系统数据库、线下文件、IoT等）的数据，或者直接拷贝源文件的“副本数据”。

数据连接

定义访问数据实体存储（计算）空间所需的信息的集合，包括连接类型、名称和登录信息等。

并发数

并发数是数据集成作业中，可以从源端并行读取的最大线程数。

脏数据

脏数据是对于业务没有意义或者格式非法的数据。例如，源端是VARCHAR类型的数据写到INT类型的目标列中，导致因为转换不合理而无法写入的数据。

作业（数据开发）

在数据开发中，作业由一个或多个节点组成，共同执行以完成对数据的一系列操作。

节点

节点用于定义对数据执行的操作。例如，使用“MRS Spark”节点可以实现在MRS中执行预先定义的Spark作业。

解决方案

解决方案定位于为用户提供便捷的、系统的方式管理作业，更好地实现业务需求和目标。每个解决方案可以包含一个或多个业务相关的作业，一个作业可以被多个解决方案复用。

资源

用户可以上传自定义的代码或文本文件作为资源，并在节点运行时调用。

表达式

数据开发作业中的节点参数可以使用表达式语言（Expression Language，简称EL），根据运行环境动态生成参数值。数据开发 EL表达式使用简单的算术和逻辑计算，引用内嵌对象，包括作业对象和一些工具类对象。

环境变量

环境变量是在操作系统中一个具有特定名字的对象，它包含了一个或者多个应用程序所将使用到的信息。

补数据

手工触发周期方式调度的作业任务，生成过去某时间段内的实例。

1.3 产品功能

数据集成：多种方式异构数据源高效接入

数据集成提供20+同构/异构数据源之间数据集成的功能，帮助您实现数据自由流动。支持自建和云上的文件系统，关系数据库，数据仓库，NoSQL，大数据云服务，对象存储等数据源。

数据集成基于分布式计算框架，利用并行化处理技术，支持用户稳定高效地对海量数据进行移动，实现不停服数据迁移，快速构建所需的数据架构。

数据集成提供全向导式任务管理界面，帮助用户在几分钟内完成数据迁移任务的创建，轻松应对复杂迁移场景。数据集成支持的功能主要有：

- **表/文件/整库迁移**

  支持批量迁移表或者文件，还支持同构/异构数据库之间整库迁移，一个作业即可迁移几百张表。

- **增量数据迁移**

  支持文件增量迁移、关系型数据库增量迁移、HBase增量迁移，以及使用Where条件配合时间变量函数实现增量数据迁移。

- **事务模式迁移**

支持当迁移作业执行失败时，将数据回滚到作业开始之前的状态，自动清理目的表中的数据。

- **字段转换**

  支持去隐私、字符串操作、日期操作等常用字段的数据转换功能。

- **文件加密**

  在迁移文件到文件系统时，数据集成支持对写入云端的文件进行加密。

- **MD5校验一致性**

  支持使用MD5校验，检查端到端文件的一致性，并输出校验结果。

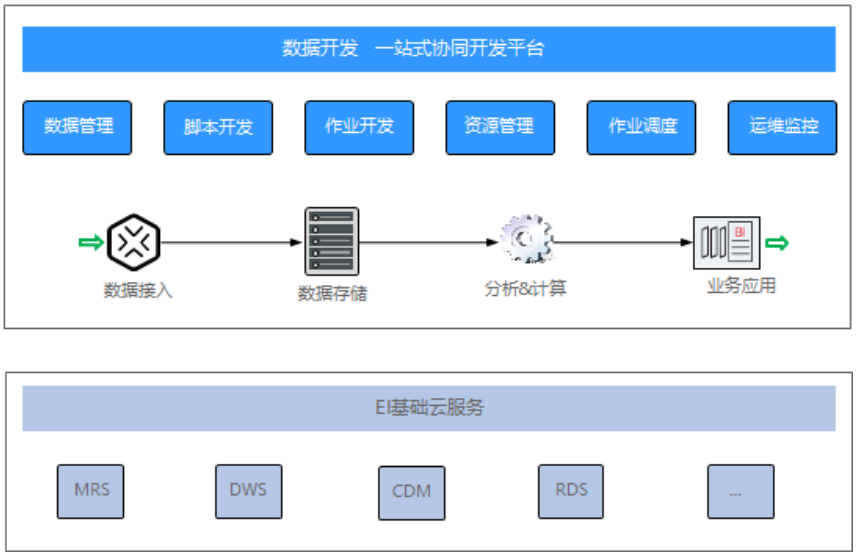
- **脏数据归档**

  支持将迁移过程中处理失败的、被清洗过滤掉的、不符合字段转换或者不符合清洗规则的数据自动归档到脏数据日志中，方便用户分析异常数据。并支持设置脏数据比例阈值，来决定任务是否成功。

数据开发：一站式协同开发平台

DataArts Studio数据开发是一个一站式敏捷大数据开发平台，提供可视化的图形开发界面、丰富的数据开发类型（脚本开发和作业开发）、全托管的作业调度和运维监控能力，内置行业数据处理pipeline，一键式开发，全流程可视化，支持多人在线协同开发，支持管理多种大数据云服务，极大地降低了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。

图 1-2 数据开发模块架构



数据开发支持数据管理、脚本开发、作业开发、资源管理、作业调度、运维监控等操作，帮助用户轻松完成整个数据的处理分析流程。

- **数据管理**
  - 支持管理DWS、DLI、MRS Hive等多种数据仓库。
  - 支持可视化和DDL方式管理数据库表。
- **脚本开发**
  - 提供在线脚本编辑器，支持多人协作进行SQL、Shell、Python脚本在线代码开发和调测。

  - 支持使用变量。
- **作业开发**
  - 提供图形化设计器，支持拖拽式工作流开发，快速构建数据处理业务流水线。
  - 预设数据集成、SQL、Shell等多种任务类型，通过任务间依赖完成复杂数据分析处理。
  - 支持导入和导出作业。
- **资源管理**

  支持统一管理在脚本开发和作业开发使用到的file、jar、archive类型的资源。
- **作业调度**
  - 支持单次调度、周期调度和事件驱动调度，周期调度支持分钟、小时、天、周、月多种调度周期。
  - 作业调度支持多种云服务的多种类型的任务混合编排，高性能的调度引擎已经经过几百个应用的检验。
- **运维监控**
  - 支持对作业进行运行、暂停、恢复、终止等多种操作。
  - 支持查看作业和其内各任务节点的运行详情。
  - 支持配置多种方式报警，作业和任务发生错误时可及时通知相关人，保证业务正常运行。

1.4 产品优势

一站式数据运营平台

贯穿数据全流程的一站式治理运营平台，提供全域数据集成、连接并萃取数据价值等，帮助企业构建完整的数据中台解决方案。

丰富的数据开发类型

支持多人在线协作开发，脚本开发可支持SQL、Shell在线编辑、实时查询；作业开发可支持CDM、SQL、MRS、Shell、MLS、Spark等多种数据处理节点，提供丰富的调度配置策略与海量的作业调度能力。

统一调度和运维

全面托管的调度，支持按时间、事件触发的任务触发机制，支持分钟、小时、天、周和月等多种调度周期。

可视化的任务运维中心，监控所有任务的运行，支持配置各类报警通知，便于责任人实时获取任务的情况，保证业务正常运行。

1.5 应用场景

云上数据平台快速搭建

快速将线下数据迁移上云，将数据集成到云上大数据服务中，并在DataArts Studio的界面中就可以进行快速的数据开发工作，让企业数据体系的建设变得如此简单。

优势

- 数据集成一键式操作

  通过在服务界面配置化操作，可实现线上线下数据快速集成到云数据仓库。
- 支持多种数仓服务类型

  根据需求，可以灵活选择数据服务类型，可以选择DWS服务建数仓，也可以选择MRS服务等数据平台。
- 安全稳定、降低成本

  一站式的服务能力和稳定的数仓服务，让云上数据万无一失；免自建大数据集群、免运维，极大降低企业建设数仓成本。

1.6 DataArts Studio 权限管理

如果您需要对的DataArts Studio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资源的访问。

通过IAM，您可以在帐号中给员工创建IAM用户，并授权来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DataArts Studio的使用权限，但是不希望他们拥有删除工作空间等高危操作的权限，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DataArts Studio服务，但是不允许删除工作空间的权限，控制他们对DataArts Studio资源的使用范围。

DataArts Studio 权限

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。

DataArts Studio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DataArts Studio时，需要先切换至授权区域。

- **IAM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。IAM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。

  DataArts Studio基于IAM角色的权限控制，提供了基于**工作空间角色**授权的能力，这是一种更加灵活的授权方式，可以精确到具体的操作。

  如**表1-1**所示，DataArts Studio的IAM系统角色包括DAYU Administrator和DAYU User；工作空间角色是基于IAM角色DAYU User进一步授予的，**DataArts Studio权限列表**列出了DataArts Studio常用操作与工作空间角色的授权关系，您可以参照这些权限列表选择合适的角色。

表 1-1 DataArts Studio 系统角色

系统角色名称	描述	类别
DAYU Administrator	数据治理中心DataArts Studio管理员权限，拥有对DataArts Studio的所有执行权限。具备对所有工作空间的所有权限。 **说明** Tenant Administrator具有除统一身份认证服务外，其他所有服务的所有执行权限。即Tenant Administrator权限的用户也拥有对DataArts Studio的所有执行权限。	系统角色
DAYU User	数据治理中心DataArts Studio普通用户，拥有被授予的工作空间的指定角色的权限。 赋予DAYU User策略的用户具有什么权限，依赖于该用户在工作空间中被赋予什么角色。工作空间有管理员、开发者、运维者和访客四种角色，每种角色的介绍如下，具体操作权限请参见**DataArts Studio权限列表**。 ● 管理员：具备DataArts Studio管理员权限，拥有工作空间内所有操作的执行权限，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：具备DataArts Studio开发权限，拥有创建、管理工作项的相关权限，但无法对工作空间、集群、审核人等进行操作，建议将任务开发、任务处理的用户设置为开发者。 ● 运维者：具备DataArts Studio运维权限，拥有运维调度等操作的执行权限，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：具备DataArts Studio只读权限，只允许对DataArts Studio进行数据读取，无法操作、更改工作项及配置，建议将只查看空间内容、不进行操作的用户设置为访客。	系统角色

用户通过工作空间角色与权限进行关联，可满足不同的授权需求。DataArts Studio角色的授权方法，请参见《数据治理中心 用户指南》中的“准备工作 > 授权用户使用DataArts Studio”。

1.7 DataArts Studio 权限列表

工作空间成员共有管理员、开发者、运维者和访客四种角色，本文将为您介绍具体角色的权限说明。

- 管理员：具备DataArts Studio管理员权限，拥有工作空间内所有操作的执行权限，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：具备DataArts Studio开发权限，拥有创建、管理工作项的相关权限，但无法对工作空间、集群、审核人等进行操作，建议将任务开发、任务处理的用户设置为开发者。
- 运维者：具备DataArts Studio运维权限，拥有运维调度等操作的执行权限，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：具备DataArts Studio只读权限，只允许对DataArts Studio进行数据读取，无法操作、更改工作项及配置，建议将只查看空间内容、不进行操作的用户设置为访客。

说明

帐号、拥有**DAYU Administrator**或**Tenant Administrator**权限的用户具有DataArts Studio的所有执行权限，包括创建DataArts Studio实例或DataArts Studio增量包的权限。其他用户默认情况下不具备创建DataArts Studio的权限，如需创建，您需要给用户赋予所需的权限。

Tenant Administrator权限具有所有云服务的管理员权限（除IAM管理权限之外），为安全起见，一般不建议给IAM用户授予该权限，请谨慎操作。

工作空间

权限点	管理员	开发者	运维者	访客
创建工作空间	DAYU Administrator或Tenant Administrator权限的用户拥有该功能操作权限。			
修改工作空间	Y	N	N	N
禁用/启用工作空间	Y	N	N	N
查询工作空间	Y	Y	Y	Y
添加工作空间成员	Y	N	N	N
修改工作空间成员	Y	N	N	N
移除工作空间成员	Y	N	N	N
查询工作空间成员	Y	Y	Y	Y

管理中心

权限点	管理员	开发者	运维者	访客
创建数据连接	Y	Y	N	N
更新数据连接	Y	Y	N	N
删除数据连接	Y	Y	N	N
获取数据连接	Y	Y	Y	Y

权限点	管理员	开发者	运维者	访客
测试数据连接	Y	Y	N	N
获取数据源类型列表	Y	Y	Y	Y
获取数据目录可用数据源类型列表	Y	Y	Y	Y
查询hive连接信息	Y	Y	Y	Y
获取数据源目录列表	Y	Y	Y	Y
数据源扩展表信息更新	Y	Y	N	N
创建数据采集任务	Y	Y	N	N
获取obs桶列表	Y	Y	Y	Y
获取obs桶中文件列表	Y	Y	Y	Y
导入数据源	Y	Y	N	N
导出数据源	Y	Y	N	N
获取kms密钥列表	Y	Y	Y	Y
获取cdm集群列表	Y	Y	Y	Y

数据集成

权限点	管理员	开发者	运维者	访客
查询连接	Y	Y	Y	Y
测试连接	Y	Y	Y	N
测试连通性	Y	Y	Y	N
创建连接	Y	Y	Y	N
删除连接	Y	Y	Y	N
查询历史作业	Y	Y	Y	Y
查询整库作业	Y	Y	Y	Y
查询普通作业	Y	Y	Y	Y

权限点	管理员	开发者	运维者	访客
查询作业名称是否存在	Y	Y	Y	Y
查询单个作业的状态	Y	Y	Y	Y
取连接元数据	Y	Y	Y	Y
创建连接元数据	Y	Y	Y	N
修改连接元数据	Y	Y	Y	N
保存作业	Y	Y	Y	N
编辑作业	Y	Y	Y	N
执行作业	Y	Y	Y	N
停止作业	Y	Y	Y	N
查询多个作业的状态	Y	Y	Y	Y
查询作业详情 / 查看作业JSON	Y	Y	Y	Y
查询作业执行的历史记录	Y	Y	Y	Y
查看作业日志	Y	Y	Y	Y
删除作业	Y	Y	Y	N
导入作业	Y	Y	Y	N
导出作业	Y	Y	Y	N
备份作业	Y	Y	Y	N
查询作业分组	Y	Y	Y	Y
创建作业分组	Y	Y	Y	N
修改作业分组	Y	Y	Y	N
删除作业分组	Y	Y	Y	N
查询配置变量	Y	Y	Y	N
设置配置变量	Y	Y	Y	N
用户隔离	Y	Y	Y	N
弹性IP检测授权	Y	N	N	N
重启集群	Y	Y	Y	N
绑定EIP	Y	N	N	N

权限点	管理员	开发者	运维者	访客
解绑EIP	Y	N	N	N
修改集群信息	Y	Y	N	N
删除集群	Y	Y	N	N
创建动态集群	Y	Y	N	N
查询集群列表	Y	Y	Y	Y
查询单个集群详情	Y	Y	Y	Y
查询单个实例详情	Y	Y	Y	Y
集群统计信息	Y	Y	Y	Y
集群agent	Y	Y	Y	N

数据开发

权限点	管理员	开发者	运维者	访客
获取环境变量列表	Y	Y	Y	Y
更新环境变量	Y	Y	N	N
导入环境变量	Y	Y	N	N
导出环境变量	Y	Y	N	N
获取数据表列表	Y	Y	Y	Y
查看表详情	Y	Y	Y	Y
创建数据表	Y	Y	N	N
更新数据表	Y	Y	N	N
删除数据表	Y	Y	N	N
获取数据库列表	Y	Y	Y	Y
查看数据库详情	Y	Y	Y	Y
新建数据库	Y	Y	N	N
更新数据库	Y	Y	N	N
删除数据库	Y	Y	N	N
获取schema列表	Y	Y	Y	Y
查看schema详情	Y	Y	Y	Y
创建schema	Y	Y	N	N

权限点	管理员	开发者	运维者	访客
更新schema	Y	Y	N	N
删除schema	Y	Y	N	N
获取目录树	Y	Y	Y	Y
新建目录	Y	Y	N	N
更新目录	Y	Y	N	N
删除目录	Y	Y	N	N
执行脚本	Y	Y	Y	N
创建脚本	Y	Y	N	N
获取脚本详情	Y	Y	Y	Y
更新脚本	Y	Y	N	N
删除脚本	Y	Y	N	N
脚本列表	Y	Y	Y	Y
取消执行	Y	Y	Y	N
导入脚本	Y	Y	N	N
导出脚本/执行结果	Y	Y	Y	N
创建解决方案	Y	Y	N	N
删除解决方案	Y	Y	N	N
更新解决方案	Y	Y	N	N
查看解决方案详情	Y	Y	Y	Y
获取解决方案列表	Y	Y	Y	Y
导出解决方案	Y	Y	Y	N
导入解决方案	Y	Y	N	N
获取作业列表	Y	Y	Y	Y
查看作业详情	Y	Y	Y	Y
创建作业	Y	Y	N	N
重命名作业	Y	Y	N	N
删除作业	Y	Y	N	N
更新作业	Y	Y	Y	N
导出作业	Y	Y	Y	N

权限点	管理员	开发者	运维者	访客
导入作业	Y	Y	N	N
导入作业校验参数	Y	Y	N	N
测试运行	Y	Y	Y	N
暂停作业运行	Y	Y	Y	N
继续执行作业	Y	Y	Y	N
运行作业	Y	Y	Y	N
停止作业	Y	Y	Y	N
获取实例列表	Y	Y	Y	Y
重跑实例	Y	Y	Y	N
停止实例	Y	Y	Y	N
强制成功	Y	Y	Y	N
继续执行实例	Y	Y	Y	N
实时作业禁用	Y	Y	Y	N
实时作业恢复	Y	Y	Y	N
作业节点手工重试	Y	Y	Y	N
跳过作业节点	Y	Y	Y	N
暂停作业节点	Y	Y	Y	N
恢复作业节点	Y	Y	Y	N
强制成功	Y	Y	Y	N
查看数据连接详情	Y	Y	Y	Y
获取数据连接列表	Y	Y	Y	Y
创建数据连接	Y	Y	N	N
更新数据连接	Y	Y	N	N
删除数据连接	Y	Y	N	N
测试数据连接	Y	Y	N	N
导入数据连接	Y	Y	N	N
导出数据连接	Y	Y	N	N
获取资源列表	Y	Y	Y	Y
查看资源详情	Y	Y	Y	Y
创建资源	Y	Y	N	N

权限点	管理员	开发者	运维者	访客
更新资源	Y	Y	N	N
删除资源	Y	Y	N	N
导入资源	Y	Y	N	N
导出资源	Y	Y	Y	N
启动每日备份	Y	Y	Y	N
停止每日备份	Y	Y	Y	N
获取备份列表	Y	Y	Y	Y
获取通知列表	Y	Y	Y	Y
配置通知	Y	Y	N	N
更新通知	Y	Y	N	N
删除通知	Y	Y	N	N
创建作业监控补数据	Y	Y	N	N
补数据监控列表	Y	Y	Y	Y
停止作业补数据	Y	Y	Y	N

1.8 约束与限制

浏览器限制

您需要使用支持的浏览器版本登录DataArts Studio。

表 1-2 浏览器兼容性

浏览器版本	说明
Google Chrome浏览器93.x及以上	建议优选

使用限制

使用DataArts Studio前，您需要认真阅读并了解以下使用限制。

1. DataArts Studio基于数据湖底座提供数据一站式集成、开发、治理等能力，本身不具备存储和计算的能力，需要配合数据湖底座使用。
2. DataArts Studio各组件对不同数据源的支持程度不一，您需要按照您的业务需求来选择数据湖底座。DataArts Studio平台当前支持的数据湖产品请参见“DataArts Studio用户指南 > 管理中心 > DataArts Studio支持的数据源”。

3. 数据集成的使用限制请参见用户指南中“数据集成-> 约束与限制”章节。

可靠性限制

DataArts Studio在使用过程中，为了达到高可靠性，建议您了解如下限制和对应措施：

1. 数据集成CDM集群为单集群部署，集群故障可能会导致业务、数据损失。建议您使用数据开发作业CDM Job节点调用CDM作业，并选择两个CDM集群以提升可靠性。详情请参见用户指南中“数据开发 > 节点参考 > CDM Job”章节。
2. CDM作业支持自动备份和恢复，将备份数据存储到OBS中，该功能需要您手动开启。详情请参见用户指南中“数据集成 > 管理作业 > 作业配置管理”章节。
3. 数据开发脚本、作业等资产支持备份管理，将备份数据存储到OBS中，该功能需要您手动开启。详情请参见用户指南中“数据开发 > 运维调度 > 备份管理”章节。

1.9 与其他云服务的关系

统一身份认证服务

DataArts Studio使用统一身份认证服务（Identity and Access Management，简称IAM）实现认证和鉴权功能。

云审计服务

DataArts Studio使用云审计服务（Cloud Trace Service，简称CTS）审计用户在管理控制台页面的操作，可用于检视是否存在非法或越权操作，完善服务安全管理。

弹性云服务器服务

DataArts Studio使用弹性云服务器（Elastic Cloud Server，简称ECS）进行CDM集群和数据服务集群的创建，另外DataArts Studio可以通过主机连接在ECS上执行Shell或Python脚本。

虚拟私有云服务

DataArts Studio使用虚拟私有云服务（Virtual Private Cloud，简称VPC）的创建隔离的网络环境。

弹性公网 IP 服务

DataArts Studio使用弹性公网IP服务（Elastic IP，简称EIP）打通与公网间的网络通信。

对象存储服务

DataArts Studio使用对象存储服务（Object Storage Service，简称OBS）的桶存储日志信息。

消息通知服务

DataArts Studio使用消息通知服务（Simple Message Notification，简称SMN）依据用户的订阅需求主动推送通知消息，使用户可以在触发告警（如质量监控）时能立即接收到通知。

云专线服务

DataArts Studio使用云专线服务（Direct Connect，简称DC）打通与第三方数据中心的网络通信。

API 网关服务

DataArts Studio通过API网关服务（API Gateway，简称APIG）对外开放各组件的API接口。

数据湖探索服务

DataArts Studio支持将数据湖探索服务（Data Lake Insight，简称DLI）作为数据湖底座，进行数据集成、开发、治理与开放。

MapReduce 服务

DataArts Studio支持将MapReduce服务（简称MRS）作为数据湖底座，进行数据集成、开发与治理。

云数据仓库服务

DataArts Studio支持将云数据仓库服务（GaussDB(DWS)，简称DWS）作为数据湖底座，进行数据集成、开发、治理与开放。

云数据库服务

DataArts Studio支持将云数据库服务（Relational Database Service，简称RDS）作为作为数据源，进行数据集成、开发与开放。

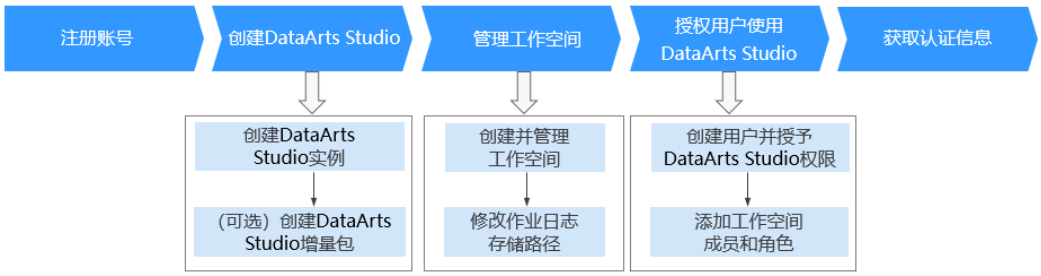
2 准备工作

2.1 准备工作简介

您需要完成创建帐号、创建DataArts Studio实例、授权用户使用DataArts Studio等一系列准备工作，才能开始DataArts Studio的正式使用。

需要进行的准备工作与具体操作请参考后续章节。

图 2-1 DataArts Studio 准备工作流程简介



2.2 创建 DataArts Studio 实例

2.2.1 创建 DataArts Studio 基础包

背景信息

只有帐号、拥有**DAYU Administrator**或**Tenant Administrator**权限的用户才可以创建DataArts Studio实例或DataArts Studio增量包。如需创建，您需要给用户授予所需的权限。

说明

Tenant Administrator策略具有所有云服务的管理员权限（除IAM管理权限之外），为安全起见，一般不建议给IAM用户授予该权限，请谨慎操作。

前提条件

已申请VPC、子网和安全组，您也可以在创建DataArts Studio实例过程中申请VPC、子网和安全组。

登录 DataArts Studio 控制台

1. 登录云控制台。
2. 在控制台左上方，单击“服务列表”按钮 ，选择“数据治理中心”，进入DataArts Studio控制台。

创建 DataArts Studio 基础包

步骤1 在DataArts Studio控制台页面，单击“创建实例”，进入创建DataArts Studio实例界面。

步骤2 配置DataArts Studio实例参数，各参数说明如表2-1所示。

表 2-1 DataArts Studio 实例参数

参数名称	样例	说明
区域	-	选择实例的区域，不同区域的资源之间内网不互通。
企业项目	default	DataArts Studio实例关联的企业项目。 如果已经创建了企业项目，这里才可以选择。当DataArts Studio实例需连接云上服务（如DWS、MRS、RDS等），还必须确保DataArts Studio实例企业项目与该云服务实例的企业项目相同。 • 一个企业项目下只能创建一个DataArts Studio实例。 • 需要与其他云服务互通时，需要确保与其他云服务的企业项目一致。
实例名称	DataArts Studio-test	自定义DataArts Studio实例名称。

步骤3 （可选）如果设置了标签键和标签值，单击右侧的“添加”，即可成功添加一条标签。

说明

- 最多支持20个标签。
- 一个“键”只能添加一个“值”。
- 每个实例中的键名不能重复。

步骤4 查看当前配置，确认无误后单击“立即创建”。

步骤5 返回DataArts Studio控制台首页时，系统会自动弹出“云资源访问授权”的对话框，提示您对所列出的服务进行委托授权。DataArts Studio与这些云服务之间存在业务交互关系，需要与这些云服务协同工作，因此需要您创建云服务委托，将操作权限委托

给DataArts Studio，让DataArts Studio以您的身份使用这些云服务，代替您进行一些任务调度、资源运维等工作。

云服务委托包含DWS、MRS、RDS、OBS、SMN、KMS等服务的相关权限，作用范围可以访问IAM的委托界面查看。 另外子账号以主账号的委托为准，不需要额外申请委托。

勾选所有服务并单击“同意授权”，系统会自动创建委托。

- 完成了委托授权后，下次再进入DataArts Studio控制台首页时，系统不会再弹出访问授权的对话框。
- 如果您只勾选了其中的某几个服务进行委托授权，下次进入DataArts Studio控制台首页时，系统仍会弹出访问授权的对话框，提示您对未授权的云服务进行访问授权。

步骤6 在已创建的实例中单击“进入控制台”，进入DataArts Studio控制台。

----结束

2.2.2 （可选）创建 DataArts Studio 增量包

DataArts Studio采用基础包+增量包的模式。如果创建的基础包无法满足您的使用需求，您可以额外创建增量包。在创建增量包前，请确保您已创建DataArts Studio实例。

您可以选择创建如下增量包：

- 数据集成增量包

  DataArts Studio实例中不包含数据集成集群，如果您需要使用数据集成的功能，需要创建数据集成增量包。

背景信息

创建增量包，系统会按照您所选规格自动创建一个所属服务的集群。

创建数据集成集群

1. 单击已开通实例卡片上的“创建增量包”。
2. 进入创建DataArts Studio增量包页面，参见**表2-2**进行配置。

表 2-2 配置数据集成的增量包

参数	说明
增量包类型	选择数据集成增量包。
可用区	第一次DataArts Studio实例或增量包时，可用区无要求。 再次创建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。

参数	说明
工作空间	选择需要使用数据集成增量包的工作空间。例如在DataArts Studio实例test的A工作空间中创建数据集成的增量包，这里工作空间选择A。创建成功后，即可通过A工作空间查看到已经创建的数据集成集群。
集群名称	自定义数据集成集群名称。
实例类型	目前数据集成集群支持以下部分规格供用户选择： ● cdm.large：8核CPU、16G内存的虚拟机，最大带宽/基准带宽为3/0.8 Gbps，能够并发执行的作业个数为20。 ● cdm.xlarge：16核CPU、32G内存的虚拟机，最大带宽/基准带宽为10/4 Gbps，能够并发执行的作业个数为100，适合使用10GE高速带宽进行TB级以上的数据量迁移。 ● cdm.4xlarge：64核CPU、128G内存的虚拟机，最大带宽/基准带宽为40/36 Gbps，能够并发执行的作业个数为300。
虚拟私有云	DataArts Studio实例中的数据集成CDM集群所属的VPC。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。 如果DataArts Studio实例或CDM集群需连接云上服务（如DWS、MRS、RDS等），则您需要确保CDM集群与该云服务网络互通。同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通，如果同虚拟私有云而子网或安全组不同，还需配置路由规则及安全组规则。 VPC的详细操作，请参见《虚拟私有云用户指南》。 **说明** 目前CDM实例创建完成后不支持切换虚拟私有云，请谨慎选择所属虚拟私有云。
子网	DataArts Studio实例中的数据迁移CDM集群所属的子网。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。 如果DataArts Studio实例或CDM集群需连接云上服务（如DWS、MRS、RDS等），则您需要确保CDM集群与该云服务网络互通。同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通，如果同虚拟私有云而子网或安全组不同，还需配置路由规则及安全组规则。 子网的详细操作，请参见《虚拟私有云用户指南》。 **说明** 目前CDM实例创建完成后不支持切换子网，请谨慎选择所属子网。

参数	说明
安全组	DataArts Studio实例中的数据集成CDM集群所属的安全组。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。 如果DataArts Studio实例或CDM集群需连接云上服务（如DWS、MRS、RDS等），则您需要确保CDM集群与该云服务网络互通。同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通，如果同虚拟私有云而子网或安全组不同，还需配置路由规则及安全组规则。 安全组的详细操作，请参见《虚拟私有云用户指南》。 **说明** 目前CDM实例创建完成后不支持切换安全组，请谨慎选择所属安全组。

须知

集群创建好以后不支持修改规格，如果需要使用更高规格，需要重新创建。

3. 单击“立即创建”，确认规格后单击“创建”。
4. 创建成功后，即可返回对应的工作空间查看已创建的数据集成集群。

2.3 管理工作空间

2.3.1 创建并管理工作空间

创建DataArts Studio实例的用户，系统将默认为其创建一个默认的工作空间“default”，并赋予该用户管理员角色。您可以使用默认的工作空间，也可以参考本章节的内容创建一个新的工作空间。

DataArts Studio实例内的工作空间作为成员管理、角色和权限分配的基本单元，包含了完整的DataArts Studio功能，工作空间的划分通常按照分子公司（如集团、子公司、部门等）、业务领域（如采购、生产、销售等）或者实施环境（如开发、测试、生产等），没有特定的划分要求。

工作空间从系统层面为管理者提供对使用DataArts Studio的用户（成员）权限、资源、DataArts Studio底层计算引擎配置的管理能力。为实现多角色协同开发，管理员可将相关用户加入到工作空间，并赋予DataArts Studio预设的项目管理员、开发者、运维者、访客等角色，其他帐号也只有在加入工作空间并被分配权限后，才可具备管理中心、数据集成、数据开发模块系列的操作权限。

约束限制

存储作业日志和脏数据依赖于OBS服务；如无OBS服务，则不支持作业日志和脏数据存储。

前提条件

请参见**创建DataArts Studio基础包**，确认已创建DataArts Studio实例。

背景说明

- DataArts Studio实例的用户，具有创建工作空间的权限。DataArts Studio将默认为其创建一个default工作空间，并赋予该用户管理员角色。
- 在主帐号创建的DataArts Studio实例中，该帐号下的IAM用户如需创建工作空间，需要由主帐号给IAM用户赋予**DAYU Administrator**或**Tenant Administrator**权限。在子用户创建的DataArts Studio实例中，主帐号默认具有该DataArts Studio实例的所有执行权限。
- 工作空间创建成功后，暂不支持删除空间的操作，您可以将不需要的工作空间禁用，以后仍可以重新启用工作空间。
- 赋予了**DAYU User**权限的用户，只有当其被添加为工作空间的成员后，才可以访问该工作空间。

创建工作空间

1. 使用**DAYU Administrator**帐号进入DataArts Studio控制台。
2. 单击控制台的“空间管理”页签，进入工作空间页面。
3. 单击“新建”，在空间信息页面请根据页面提示配置参数，参数说明如**表2-3**所示，配置完成后，单击“确定”完成工作空间的创建。

表 2-3 新建空间参数说明

参数名	说明
空间名称	空间名称，只能包含字母、数字、下划线、中划线、中文字符，且长度不超过32个字符。在当前的DataArts Studio实例中，工作空间名称必须唯一。
空间描述	空间的描述信息。
企业项目	DataArts Studio实例关联的企业项目。 如果已经创建了企业项目，这里才可以选择。当DataArts Studio实例需连接云上服务（如DWS、MRS、RDS等），还必须确保DataArts Studio实例企业项目与该云服务实例的企业项目相同。 ● 一个企业项目下只能创建一个DataArts Studio实例。 ● 需要与其他云服务互通时，需要确保与其他云服务的企业项目一致。
作业日志OBS路径	用于指定DataArts Studio数据开发作业的日志存储的OBS桶。工作空间成员如需使用DataArts Studio数据开发，必须具备“作业日志OBS桶”的读、写权限，否则，在使用过程中，系统将无法正常读、写数据开发的作业日志。 ● 单击“请选择”按钮，您可以选择一个已创建的OBS桶和对象，系统将基于工作空间全局配置作业日志OBS桶。 ● 如果不配置该参数，DataArts Studio数据开发的作业日志默认存储在以“dlf-log-{projectId}”命名的OBS桶中。{projectId}即项目ID，您可以参考**获取项目ID和帐号ID**进行获取。

参数名	说明
DLI脏数据OBS路径	用于指定DataArts Studio数据开发中DLI SQL执行过程中的脏数据存储的OBS桶。工作空间成员如需使用DataArts Studio数据开发执行DLI SQL，必须具备“DLI脏数据OBS桶”的读、写权限，否则，在使用过程中，系统将无法正常读、写DLI SQL执行过程中的脏数据。 ● 单击“请选择”按钮，您可以选择一个已创建的OBS桶和对象，系统将基于工作空间全局配置DLI脏数据OBS桶。 ● 如果不配置该参数，DataArts Studio数据开发的DLI SQL脏数据默认存储在以“dlf-log-{projectId}”命名的OBS桶中。

编辑工作空间

1. 登录DataArts Studio控制台。
2. 找到所需要的DataArts Studio实例，在DataArts Studio实例上单击“进入控制台”。然后，选择“空间管理”页签。
3. 在“空间管理”页面，找到所需编辑的工作空间，单击其所在行的“编辑”，此时显示“空间信息”页面。
4. 在“空间信息”页面的最上方，单击编辑按钮，您就可以编辑空间信息以及管理空间成员，请根据页面提示进行配置。
5. 配置完成后，在“空间信息”页面的最上方单击保存按通过成功以保存配置。

禁用工作空间

工作空间创建成功后，默认为启用状态。如果您不再需要某个工作空间，DataArts Studio暂不支持删除空间的操作，您可以将工作空间禁用，以后仍可以将其重新启用。

说明

工作空间被禁用后，您将无法再访问工作空间，无法编辑工作空间内的工作项，工作空间内调度作业将停止运行。

1. 登录DataArts Studio控制台。
2. 找到所需要的DataArts Studio实例，在DataArts Studio实例上单击“进入控制台”。然后，选择“空间管理”页签。
3. 在“空间管理”页面，找到所需禁用的工作空间，单击其所在行的状态按钮  。
4. 在“禁用”对话框中，了解禁用空间的影响后，如果确认要禁用空间，请单击“确定”。

启用工作空间

1. 登录DataArts Studio控制台。
2. 找到所需要的DataArts Studio实例，在DataArts Studio实例上单击“进入控制台”。然后，选择“空间管理”页签。
3. 在“空间管理”页面，找到所需启用的工作空间，单击其所在行的状态按钮  按钮。

4. 在“启用”对话框中，如果确认启用，请单击“确定”。

2.3.2 （可选）修改作业日志存储路径

作业日志和DLI脏数据默认存储在以dlf-log-{Project id}命名的OBS桶中，您也可以自定义日志存储路径，数据开发模块支持您基于工作区全局配置OBS桶。

约束限制

该功能依赖于OBS服务。

前提条件

修改作业日志存储路径的用户，需要满足如下任一条件：

- 帐号为拥有管理员权限的用户。
- **DAYU User**权限的用户，但需是当前工作空间的管理员。

修改方法

1. 使用**DAYU Administrator**或管理员帐号进入DataArts Studio控制台。
2. 单击控制台的“空间管理”页签，进入工作空间页面。
3. 单击待修改工作空间对应的“编辑”按钮。
4. 在空间信息页面中，单击空间信息后 的“编辑”，该空间信息置于可编辑状态。
   单击作业日志OBS路径后的“请选择”按钮，重新选择日志存储路径，可选择某个具体的目录。

图 2-2 修改日志路径

5. 修改完成后，单击“保存”，即完成作业日志存储路径的自定义修改。

2.4 授权用户使用 DataArts Studio

2.4.1 创建 IAM 用户并授予 DataArts Studio 权限

如果您需要对您所拥有的DataArts Studio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，在您的帐号中，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DataArts Studio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DataArts Studio资源委托给更专业、高效的其他帐号或者云服务，这些帐号或者云服务可以根据权限进行代运维。

如果帐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DataArts Studio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**创建IAM用户并授予DataArts Studio权限**所示。

背景信息

- 给用户组授权之前，请您了解用户组可以添加的DataArts Studio工作空间角色权限，并结合实际需求进行选择。

创建 IAM 用户并授予 DataArts Studio 权限

1. 创建用户组并授权。使用帐号登录IAM控制台，创建用户组，并授予DataArts Studio的普通用户权限，如“DAYU User”。

   创建用户组并授权的具体操作，请参见《统一身份认证服务IAM用户指南》中的“用户组及授权> 创建用户组并授权”。

   说明

   - 配置用户组的DataArts Studio权限时，无需进行筛选，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”。
   - 如果您需要给IAM用户创建工作空间的权限，则需要给IAM用户授予“DAYU Administrator”权限，“DAYU Administrator”权限具有DataArts Studio服务的所有执行权限。
   - 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。

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《统一身份认证服务IAM用户指南》中的“IAM用户> 创建IAM用户”。

2.4.2 添加工作空间成员和角色

如果您需要添加其他IAM用户协同使用DataArts Studio实例，请参考**创建IAM用户并授予DataArts Studio权限**的操作准备必要的IAM用户，然后参考本章节将该用户添加为工作空间成员并配置工作空间角色。

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产品介绍中的“DataArts Studio权限列表”章节。

- 管理员：具备DataArts Studio管理员权限，拥有工作空间内所有操作的执行权限，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：具备DataArts Studio开发权限，拥有创建、管理工作项的相关权限，但无法对工作空间、集群、审核人等进行操作，建议将任务开发、任务处理的用户设置为开发者。
- 运维者：具备DataArts Studio运维权限，拥有运维调度等操作的执行权限，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：具备DataArts Studio只读权限，只允许对DataArts Studio进行数据读取，无法操作、更改工作项及配置，建议将只查看空间内容、不进行操作的用户设置为访客。

背景信息

DAYU Administrator帐号或管理员角色可以在工作空间中添加成员。

添加成员和角色

1. 登录DataArts Studio控制台，进入工作空间列表页面。
2. 单击相应工作空间列表后的“编辑”，进入成员空间页面。
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。
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。

移除空间成员

1. 登录DataArts Studio控制台，进入工作空间列表页面。
2. 在“空间管理”页面，找到需要移除成员的工作空间，单击其所在行“操作”列的“编辑”。
3. 进入空间信息页面后，在成员列表中勾选所需移除的成员，单击“移除”按钮。

   说明

   工作空间的所有者不能被删除。
4. 在“移除”对话框中，如果确认要移除成员，请单击“确定”。

2.5 （可选）获取认证信息

DataArts Studio使用过程中，在数据集成创建OBS连接、API调用、使用问题定位时，您可能需要获取访问密钥、项目ID、终端节点等信息，获取方式如下。

获取访问密钥

您可以通过如下方式获取访问密钥。

由于“亚太-吉隆坡-OP6”区域用户属于联邦认证授权访问“亚太-吉隆坡-OP6”云服务系统的虚拟用户，不是“亚太-吉隆坡-OP6”云服务系统中真实存在的用户。因此需要联系管理员在“亚太-吉隆坡-OP6”区域分别获取访问密钥AK/SK。

获取项目 ID 和帐号 ID

项目ID表示租户的资源，帐号ID对应当前帐号。用户可在对应页面下查看不同Region对应的项目ID和帐号ID。

1. 注册并登录管理控制台。
2. 在用户名的下拉列表中单击“我的凭证”。
3. 在“我的凭证”页面，查看帐号名和帐号ID，在项目列表中查看项目ID。

获取 DataArts Studio 实例 ID 和工作空间 ID

DataArts Studio的实例ID和工作空间ID可以从DataArts Studio控制台的URI链接中获取。

1. 在DataArts Studio控制台首页，选择对应工作空间，并点击任一模块，如“管理中心”。

   图 2-3 选择管理中心

2. 进入管理中心页面后，从浏览器地址栏中获取“instanceId”和“workspace”对应的值，即为DataArts Studio的实例ID和工作空间ID。

   如图2-4所示，实例ID为**6b88…2688**，工作空间ID为**1dd3bc…d93f0**。

   图 2-4 获取实例 ID 和工作空间 ID

获取终端节点

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。

表 2-4 DataArts Studio 终端节点信息

区域名称	区域	组件	终端节点（Endpoint）	协议类型
亚太-吉隆坡-OP6	my-kualalumpur-1	数据集成	cdm.my-kualalumpur-1.alphaedge.tmone.com.my	HTTPS/HTTP
		数据开发	dayu-dlf.my-kualalumpur-1.alphaedge.tmone.com.my	

3 用户指南

3.1 使用 DataArts Studio 前的准备

在使用DataArts Studio前，您应首先进行数据与业务调研，选择合适的数据治理模型。

然后参考本章节，预先做好以下准备工作：

- **DataArts Studio准备工作**
- **准备数据源**
- **准备数据湖**

DataArts Studio 准备工作

如果您是第一次使用DataArts Studio，请参考用户指南中的“准备工作”章节，完成创建DataArts Studio实例、创建工作空间等一系列操作。然后找到对应的工作空间，即可开始数据开发与运营。

准备数据源

在实际业务中，源端数据源大多为云下的MySQL、PostgreSQL、HBase、Hive等类型，您需要作如下准备：

- 确保数据源所在的主机可以访问公网。
- 获取数据源的公网连接地址、数据库端口、数据库管理员用户及密码等信息。
- 确保防火墙规则出方向已开放数据库端口，允许数据传输到云上。

准备好数据源之后，后续您可以通过数据集成将数据源迁移到数据湖底座中，然后再通过DataArts Studio进行数据开发、治理和运营等活动。

准备数据湖

在使用DataArts Studio前，您需要根据业务场景选择符合需求的云服务作为DataArts Studio的数据湖底座，用于存储原始数据和数据开发过程中的数据，并进行后续的数据开发、治理和运营等活动。DataArts Studio平台当前支持的数据湖产品请参见**DataArts Studio支持的数据源**。

准备好数据湖之后，您可以通过**创建数据连接**将DataArts Studio与数据湖底座连接起来，然后进行**1**和**2**的操作。**1**和**2**的操作样例可参考快速入门中的“步骤2：准备工作”章节。

1. **创建数据库**

   在使用DataArts Studio数据集成将数据迁移上云之前，我们需要在目的端数据湖中创建目标数据库。根据数据湖治理落地流程，建议您在数据湖中为SDI层、DWI层、DWR层和DM层分别创建一个数据库，从而对数据进行分层分库。数据分层是后面在数据架构中将涉及到的概念，此处可先简单了解，在数据架构时将深入了解与操作。

   您可以参考以下任一一种方式在数据湖中创建数据库。

   - 您可以在DataArts Studio数据开发模块中，可视化方式创建数据库，具体操作请参见“数据开发 > 数据管理 > 新建数据库”章节。
   - 您可以通过在DataArts Studio数据开发模块或数据湖产品的SQL编辑器上，开发并执行用于创建数据库的SQL脚本，从而创建数据库。在DataArts Studio数据开发模块开发脚本的具体操作请参见“数据开发 > 脚本开发 > 开发脚本> 开发SQL脚本”章节；数据湖产品的SQL编辑器上的具体操作请参见对应数据湖产品的帮助文档。

2. **创建数据表**

   在使用DataArts Studio数据集成将数据迁移上云之前，我们需要在目的端数据湖的SDI层数据库中创建一个目标表，用于存储原始数据。批量数据迁移场景下，关系型数据库之间的迁移和关系型数据库到Hive的迁移支持自动创建目标表，这种情况下可以不预先在目的端数据库中创建目标表。

   您可以参考以下任一一种方式在数据湖中创建原始数据表。如果表字段个数较多，建议使用编写SQL脚本的方式创建表。

   - 您可以在DataArts Studio数据开发模块中，可视化方式创建数据表，具体操作请参见“数据开发 > 数据管理 > 新建数据表”章节。
   - 您可以通过在DataArts Studio数据开发模块或数据湖产品的SQL编辑器上，开发并执行用于创建数据表的SQL脚本，从而创建数据表。在DataArts Studio数据开发模块开发脚本的具体操作请参见“数据开发 > 脚本开发 > 开发脚本> 开发SQL脚本”章节；数据湖产品的SQL编辑器上的具体操作请参见对应数据湖产品的帮助文档。

3.2 管理中心

DataArts Studio管理中心提供了统一的配置和管理入口，可以管理数据连接、资源迁移等，根据需要定制个性化的入口和展示。

3.2.1 DataArts Studio 支持的数据源

在使用DataArts Studio前，您需要根据业务场景选择符合需求的云服务或数据仓库作为数据湖，用于存储原始数据和数据治理过程中的数据，并进行数据开发、服务和运营。DataArts Studio集成了丰富的数据引擎，支持对接如DLI、DWS、MRS Hive等云上数据湖与数据库云服务，也支持对接企业传统数据库，例如MySQL、PostgreSQL等。

DataArts Studio 支持的数据源

DataArts Studio支持的数据源可分为“数据集成组件支持的数据源”和“DataArts Studio其他组件支持的数据源”。

- 数据集成组件支持的数据源。数据集成组件需要集成源数据到数据湖中，因此支持的数据源范围更广。

  数据集成支持的数据源请参见**数据集成支持的数据源**。注意，如需在数据集成中使用这些数据源，请先在数据集成中创建对应的数据连接，这些数据连接仅限于在数据集成模块中使用。

- DataArts Studio其他组件支持的数据源，即为DataArts Studio所支持的数据湖底座。

  其他组件支持的数据源如**表3-1**所示，数据源的介绍请参见**数据源简介**。注意，如需在其他组件中使用这些数据源，请先前往DataArts Studio管理中心控制台创建数据连接，这些数据连接不能在数据集成模块中使用。

表 3-1 DataArts Studio 其他组件支持的数据源

数据源类型	管理中心	数据开发
数据仓库服务（DWS）	√	√
数据湖探索（DLI）	√	√
MapReduce服务（MRS HBase）	√	×
MapReduce服务（MRS Hive）	√	√
MapReduce服务（MRS Kafka）	√	√
MySQL	√	×
MapReduce服务（MRS Spark）	√	√
云数据库 RDS（MySQL）	√	√
云数据库 RDS（PostgreSQL）	√	√
主机连接	√	√
MapReduce服务（MRS Presto）	√	√

数据源简介

表 3-2 数据源简介

数据源类型	简介
数据仓库服务（DWS）	DWS是基于Shared-nothing分布式架构，具备MPP大规模并行处理引擎，兼容标准ANSI SQL 99和SQL 2003，同时兼容PostgreSQL/Oracle数据库生态，为各行业PB级海量大数据分析提供有竞争力的解决方案。
数据湖探索（DLI）	DLI是完全兼容Apache Spark和Apache Flink生态，实现批流一体的Serverless大数据计算分析服务。DLI支持多模引擎，企业仅需使用SQL或程序就可轻松完成异构数据源的批处理、流处理、内存计算、机器学习等，挖掘和探索数据价值。

数据源类型	简介
MapReduce服务（MRS HBase）	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。 使用MRS HBase可实现海量数据存储，并实现毫秒级数据查询。选择MRS HBase可以实现物流数据毫秒级实时入库更新，并支持百万级时序数据查询分析。
MapReduce服务（MRS Hive）	Hive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。 使用MRS Hive可实现TB/PB级的数据分析，快速将线下Hadoop大数据平台（CDH、HDP等）迁移上云，业务迁移 “0”中断，业务代码 “0”改动。
MapReduce服务（MRS Kafka）	MapReduce服务可提供专属MRS Kafka集群。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。
MySQL	MySQL是目前最受欢迎的开源数据库之一，其性能卓越，架构成熟稳定，支持流行应用程序，适用于多领域多行业，支持各种WEB应用，成本低，中小企业首选。
MapReduce服务（MRS Spark）	Spark是一个开源的，并行数据处理框架，能够帮助用户简单的开发快速、统一的大数据应用，对数据进行协处理、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提供类似SQL的Spark SQL语言操作结构化数据。
云数据库 RDS	RDS是一种基于云计算平台的即开即用、稳定可靠、弹性伸缩、便捷管理的在线关系型数据库服务。 注意，DataArts Studio平台目前仅支持RDS中的MySQL和PostgreSQL数据库。
主机连接	通过主机连接，用户可以在DataArts Studio数据开发中连接到指定的主机，通过脚本开发和作业开发在主机上执行Shell或Python脚本。主机连接保存连接某个主机的连接信息，当主机的连接信息有变化时，只需在主机连接管理中编辑修改，而不需要到具体的脚本或作业中逐一修改。

数据源类型	简介
MapReduce服务（MRS 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。

3.2.2 创建数据连接

通过配置数据源信息，可以建立数据连接。DataArts Studio基于管理中心的数据连接对数据湖底座进行数据开发、治理、服务和运营。

约束限制

- RDS数据连接方式依赖于OBS。如果没有与DataArts Studio同区域的OBS，则不支持RDS数据连接。
- 当所连接的数据湖发生变化（如MRS集群扩容等情况）时，您需要重新编辑并保存该连接。

前提条件

- 在创建数据连接前，请确保您已创建所要连接的数据湖（如DataArts Studio所支持的数据库、云服务等）。
  - 在创建DWS类型的数据连接前，您需要先在DWS服务中创建集群，并且具有KMS密钥的查看权限。
  - 在创建MRS HBase、MRS Hive、MRS Kafka、MRS Spark、MRS Presto类型的数据连接前，需确保您已创建MRS集群，并且在创建数据链接时已创建选择所需要的组件。
  - 在创建RDS类型的数据连接前，请确保您已创建RDS数据库实例。DataArts Studio平台目前仅支持RDS中的MySQL和PostgreSQL数据库引擎。
- 在创建数据连接前，请确保待连接的数据湖与DataArts Studio实例之间网络互通。
  - 如果数据湖为云下的数据库，则需要通过公网或者专线打通网络，确保数据源所在的主机可以访问公网，并且防火墙规则已开放连接端口。
  - 如果数据湖为云上服务（如DWS、MRS等），则网络互通需满足如下条件：
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上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。
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。

- 此外，您还必须确保该云服务的实例与DataArts Studio工作空间所属的企业项目必须相同，如果不同，您需要修改工作空间的企业项目。

创建数据连接

1. 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

图 3-1 选择管理中心

2. 在管理中心页面，单击“数据连接”，进入数据连接页面。

图 3-2 创建数据连接

3. 单击“创建数据连接”，在弹出的对话框中，选择“数据连接类型”，并参见**表3-3**配置相关参数。

图 3-3 创建数据连接

表 3-3 数据连接

数据连接类型	参数说明
MRS Hive	请参见**表3-4**。
MRS HBase	请参见**表3-5**。
MRS Kafka	请参见**表3-6**。
DWS	请参见**表3-9**。
ORACLE	请参见**表3-10**
MRS Spark	请参见**表3-7**。
RDS	请参见**表3-8**。 RDS连接类型还支持创建与部分关系型数据库的连接，如MySQL/PostgreSQL/达梦数据库 DM等。
主机连接	请参见**表3-11**。

4. 单击“测试”，测试数据连接的连通性。如果无法连通，数据连接将无法创建。
5. 测试通过后，单击“确定”，完成数据连接的创建。

数据连接参数说明

表 3-4 MRS Hive 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。

参数	是否必选	说明
集群名	是	选择Hive所属的MRS集群。如果在下拉列表中无法显示MRS集群，请检查MRS集群与DataArts Studio实例是否网络互通。 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件： ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。 ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。 ● 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。
连接方式	是	选择所需的连接方式，推荐使用“通过代理连接”。 ● 通过代理连接：通过Agent（即CDM集群）进行代理，以MRS集群的用户名和密码访问MRS集群。代理连接方式支持MRS所有版本的集群。 ● MRS API连接：以MRS API的方式访问MRS集群。MRS API连接仅支持2.X及更高版本的MRS集群。 选择MRS API连接时，有以下约束： 1. 无法查看表和字段。 2. 在SQL编辑器运行SQL时，只能以日志形式显示执行结果。 3. 数据治理（如数据架构、数据质量、数据目录等组件）功能无法使用MRS API连接。

参数	是否必选	说明
用户名	否	MRS集群的用户名，通过代理连接的时候，是必选项。如果使用新建的MRS用户进行连接，您需要先登录Manager页面，并更新初始密码。 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。您可以参考**创建MRS安全集群的kerberos认证用户**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** • MRS 3.1.0及之后版本集群，所创建的用户至少需具备Manager_viewer的角色权限才能在管理中心创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 • MRS 3.1.0版本之前的集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。 • 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。
密码	否	MRS集群的访问密码，通过代理连接的时候，是必选项。
KMS密钥	否	KMS密钥名称。通过代理连接的时候，是必选项。
绑定Agent	否	通过代理连接的时候，是必选项。 MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和MRS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和MRS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

表 3-5 MRS HBase 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。

参数	是否必选	说明
集群名	是	选择HBase所属的MRS集群。如果在下拉列表中无法显示MRS集群，请检查MRS集群与DataArts Studio实例是否网络互通。 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件： • DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。 • DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。 • 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。
用户名	是	MRS集群的用户名。 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。您可以参考**创建MRS安全集群的kerberos认证用户**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** • MRS 3.1.0及之后版本集群，所创建的用户至少需具备Manager_viewer的角色权限才能在管理中心创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 • MRS 3.1.0版本之前的集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。 • 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。
密码	是	MRS集群的访问密码。
KMS密钥	是	KMS密钥名称。

参数	是否必选	说明
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和MRS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和MRS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

表 3-6 MRS Kafka 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。
集群名	是	选择Kafka所属的MRS集群。如果在下拉列表中无法显示MRS集群，请检查MRS集群与DataArts Studio实例是否网络互通。 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件： ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。 ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。 ● 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。

参数	是否必选	说明
用户名	是	MRS集群的用户名。 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。您可以参考**创建MRS安全集群的kerberos认证用户**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** • MRS 3.1.0及之后版本集群，所创建的用户至少需具备Manager_viewer的角色权限才能在管理中心创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 • MRS 3.1.0版本之前的集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。 • 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。
密码	是	MRS集群的访问密码。
KMS密钥	是	KMS密钥名称。
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和MRS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和MRS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

表 3-7 MRS Spark 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。

参数	是否必选	说明
集群名	是	选择Spark所属的MRS集群名称。如果在下拉列表中无法显示MRS集群，请检查MRS集群与DataArts Studio实例是否网络互通。 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件： ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。 ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。 ● 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。
连接方式	是	选择所需的连接方式，推荐使用“通过代理连接”。 ● 通过代理连接：通过Agent（即CDM集群）进行代理，以MRS集群的用户名和密码访问MRS集群。代理连接方式支持MRS所有版本的集群。 ● MRS API连接：以MRS API的方式访问MRS集群。MRS API连接仅支持2.X及更高版本的MRS集群。 选择MRS API连接时，有以下约束： 1. 无法查看表和字段。 2. 在SQL编辑器运行SQL时，只能以日志形式显示执行结果。 3. 数据治理（如数据架构、数据质量、数据目录等组件）功能无法使用MRS API连接。

参数	是否必选	说明
用户名	否	MRS集群的用户名，通过代理连接的时候，是必选项。如果使用新建的MRS用户进行连接，您需要先登录Manager页面，并更新初始密码。 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。您可以参考**创建MRS安全集群的kerberos认证用户**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** • MRS 3.1.0及之后版本集群，所创建的用户至少需具备Manager_viewer的角色权限才能在管理中心创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 • MRS 3.1.0版本之前的集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。 • 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。
密码	否	MRS集群的访问密码，通过代理连接的时候，是必选项。
KMS密钥	否	KMS密钥名称。通过代理连接的时候，是必选项。
绑定Agent	否	通过代理连接的时候，是必选项。 MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和MRS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和MRS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

表 3-8 RDS 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。

参数	是否必选	说明
IP	是	RDS的访问地址。 如果为RDS数据源，可以通过RDS管理控制台获取访问地址： 1. 根据创建的帐号登录管理控制台。 2. 单击“云数据库 RDS”，从左侧列表选择实例管理。 3. 单击某一个实例名称，进入实例基本信息页面。 在连接信息标签中可以获取到内网地址。
端口	是	RDS的访问端口。 如果为RDS数据源，可以通过RDS管理控制台获取访问端口： 1. 根据的帐号登录管理控制台。 2. 单击“云数据库 RDS”，左侧列表选择实例管理。 3. 单击某一个实例名称，进入实例基本信息页面。 在连接信息标签中可以获取到数据库端口。
驱动程序名称	是	驱动程序名称： • com.mysql.jdbc.Driver • org.postgresql.Driver
驱动文件路径	是	驱动文件在OBS上的路径。需要您自行到官网下载.jar格式驱动并上传至OBS中。 • MySQL驱动：获取地址**https://downloads.mysql.com/archives/c-j/**，建议5.1.48版本。 • PostgreSQL驱动：获取地址**https://jdbc.postgresql.org/download**，建议42.1.4版本。 **说明** 如果需要更新驱动文件，则需要先在数据集成页面重启CDM集群，然后通过编辑数据连接的方式重新选择新版本驱动，更新驱动才能生效。
用户名	是	数据库的用户名，创建集群的时候，输入的用户名。
密码	是	数据库的访问密码，创建集群的时候，输入的密码。
KMS密钥	是	KMS密钥名称。 通过KMS管理控制台获取密钥名称： 1. 根据的帐号登录管理控制台。 2. 单击“密钥管理服务”，左侧列表选择密钥管理。 在密钥列表可以获取到密钥名称。

参数	是否必选	说明
绑定Agent	是	RDS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建RDS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和RDS网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和RDS处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

表 3-9 DWS 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。
手动	是	通过单击或来关闭或开启手动开关： ● 当“手动”关闭时候，“IP”和“端口”不需要填写。 ● 当“手动”打开时候，“IP”和“端口”需要填写。
IP	否	“手动”打开时需要填写该项，表示通过内部网络访问集群数据库的IP地址。内网访问IP地址在创建集群时自动生成。
端口	否	“手动”打开时需要填写该项，表示创建DWS集群时指定的数据库端口号。请确保您已在安全组规则中开放此端口，以便DataArts Studio实例可以通过该端口连接DWS集群数据库。
SSL连接	是	DWS支持SSL通道加密和证书认证两种方式进行客户端与服务器端的通信。您可以通过服务器端是否强制使用SSL连接进行设置。开关打开，即只能通过SSL方式连接。开关关闭，即两种方式均可。默认关闭。
集群名	是	选择DWS集群。
用户名	是	数据库的用户名，创建DWS集群时指定的用户名。
密码	是	数据库的访问密码，创建DWS集群时指定的密码。
KMS密钥	是	KMS密钥名称。

参数	是否必选	说明
连接方式	是	选择所需的连接方式，推荐使用“通过代理连接”。 ● 通过代理连接：通过Agent（即CDM集群）进行代理连接访问DWS集群。 ● 直接连接：直接访问DWS集群。
绑定Agent	否	通过代理连接的时候，是必选项。 DWS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建DWS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和DWS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和DWS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

表 3-10 Oracle 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。
IP	是	待连接的数据库IP地址，公网IP和内网IP地址均支持。
端口	是	待连接的数据库端口。

参数	是否必选	说明
用户名	是	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。 **说明** CONNECT权限的用户(只读用户)创建连接时会出现“表或视图不存在”的提示，需要执行如下操作进行授权： 1. 以**root**用户登录oracle节点。 2. 执行如下命令，切换到oracle用户。 **su oracle** 3. 执行如下命令，登录数据库。 **sqlplus /nolog** 4. 执行如下命令，登录**sys**用户 **connect sys as sysdba;** 输入sys用户的密码。 5. 执行如下SQL语句，进行授权。 **GRANT SELECT ON GV_$INSTANCE to *xxx*;** 其中，*xxx*为需要授权的用户名。
密码	是	用户密码。
sid	是	Oracle数据库的唯一标识符。
KMS密钥	是	KMS密钥名称。 通过KMS管理控制台获取密钥名称： 1. 根据创建的帐号登录管理控制台。 2. 单击“密钥管理服务”，左侧列表选择密钥管理。 在密钥列表可以获取到密钥名称。
绑定Agent	是	Oracle为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建Oracle的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和Oracle网络互通才可以成功创建MRS连接。

表 3-11 主机连接

参数	是否必选	说明
数据连接名称	是	主机连接的名称，只能包含字母，数字，中划线或者下划线。
主机地址	是	主机的地址。 请参见《弹性云服务器用户指南》的查看云服务器详细信息页获取。

参数	是否必选	说明
绑定Agent	是	需要选择CDM集群，CDM集群提供Agent。
端口	是	主机的SSH端口号。
用户名	是	主机的登录用户名。
登录方式	是	选择主机的登录方式： ● 密钥对 ● 密码
密钥对	是	主机的登录方式为密钥对时，用户获取并上传其私钥文件至OBS，在此处选择对应的OBS路径。“登录方式”为“密钥对”时，显示该配置项。 **说明** 此处上传的私钥文件需为PEM格式，并且上传的私钥文件和主机上配置的公钥是一个密钥对。
密钥对密码	否	如果密钥对未设置密码，则不需要填写该配置项。
密码	是	主机的登录方式为密码时，填写主机的登录密码。
主机连接描述	否	主机连接的描述信息。

创建 MRS 安全集群的 kerberos 认证用户

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。您可以参考以下步骤创建一个新的MRS用户：

针对MRS 3.x版本集群：

1. 使用admin登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有用户作为kerberos认证用户，并且为这个用户添加用户组和分配角色权限，用户组选择superGroup，角色建议全选，然后根据页面提示完成用户的创建。

   说明

   - MRS 3.1.0及之后版本集群，所创建的用户至少需具备Manager_viewer的角色权限才能在管理中心创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。
   - MRS 3.1.0版本之前的集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。
   - 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。
3. 使用新建的用户登录Manager页面，并更新初始密码，否则会导致创建连接失败。
4. 同步IAM用户。
   a. 登录MRS管理控制台。
   b. 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。

c. 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。

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。

针对MRS 2.x及之前版本集群：

1. 使用admin登录MRS Manager页面。
2. 在MRS Manager页面的“系统设置”中，单击“用户管理”，在用户管理页面，添加用户，添加一个专有用户作为kerberos认证用户，并且为这个用户添加用户组和分配角色权限，用户组选择superGroup，角色建议全选，然后根据页面提示完成用户的创建。

说明

- MRS 2.x及之前版本集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。
- 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。

3. 使用新建的用户登录MRS Manager页面，并更新初始密码，否则会导致创建连接失败。
4. 同步IAM用户。
   a. 登录MRS管理控制台。
   b. 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。
   c. 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。

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。

编辑数据连接

步骤1 登录DataArts Studio管理中心控制台，单击“数据连接”，进入数据连接页面。

步骤2 在数据连接列表中，找到所需编辑的连接，然后单击“编辑”。

步骤3 在“编辑数据连接”对话框中，根据需要修改连接参数，参数描述可参考**数据连接参数说明**。

步骤4 完成修改后，单击“测试”测试数据连接的是否可以正常连接，如果可以正常连接，单击“确定”。

如果测试连接无法连通，数据连接将无法创建，请根据错误提示重新修改连接参数后再进行重试。

----结束

删除数据连接

若删除数据连接，此数据连接下的数据表信息也会被删除，请谨慎操作。删除数据连接时，若待删除的连接已被引用，则不可删除，反之，可删除。

步骤1 登录DataArts Studio管理中心控制台，单击“数据连接”，进入数据连接页面。

步骤2 在数据连接列表中，找到所需删除的连接，然后单击“删除”。

步骤3 在删除确认对话框中，了解删除连接的影响后，若要删除，单击“确定”。

----结束

3.2.3 资源迁移

当您需要将一个工作空间中的资源迁移至另一个工作空间，可使用数据治理中心DataArts Studio的资源迁移功能，对资源进行导入导出。

资源迁移支持迁移的资源包含管理中心数据连接。

前提条件

- 资源导入导出功能依赖于OBS服务。
- 系统中存在可迁移的资源，参见**创建数据连接**创建数据连接。

约束条件

- 导入导出的资源以json格式存储。
- 由于安全原因，导出连接时没有导出连接密码，需要在导入时自行输入。

导出资源

1. 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

图 3-4 选择管理中心

2. 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图 3-5 资源迁移

3. 单击“新建导出”，配置文件的OBS存储位置和文件名称。如无OBS服务，仅需设置导出文件名即可。

图 3-6 选择导出文件

4. 单击“下一步”，勾选导出的模块。
5. 单击“下一步”，等待导出完成，资源包导出到3所设置的OBS存储位置。如无OBS服务，则导出完成后可在资源迁移的对应迁移任务行中，单击“下载”获取导出的资源包。

图 3-7 导出完成

导出资源耗时1分钟仍未显示结果则表示导出失败，请重试。如果仍然无法导出，请联系或技术支持人员协助解决。

导入资源

1. 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

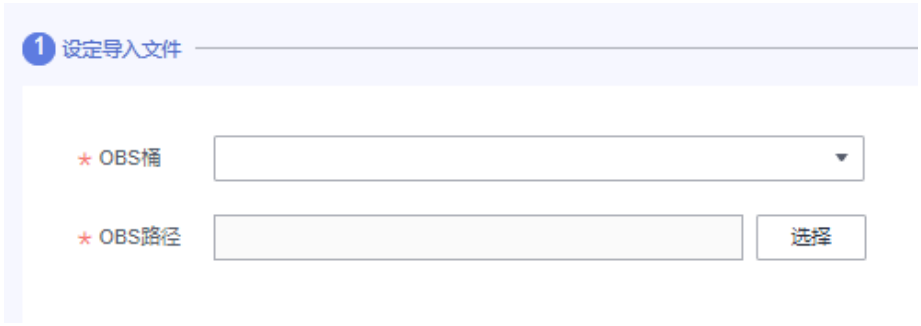
图 3-8 选择管理中心

2. 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图 3-9 资源迁移

3. 单击“新建导入”，配置待导入资源的OBS存储路径。如无OBS服务，需要从本地路径选择待上传的资源包。

图 3-10 配置待导入的资源存储路径

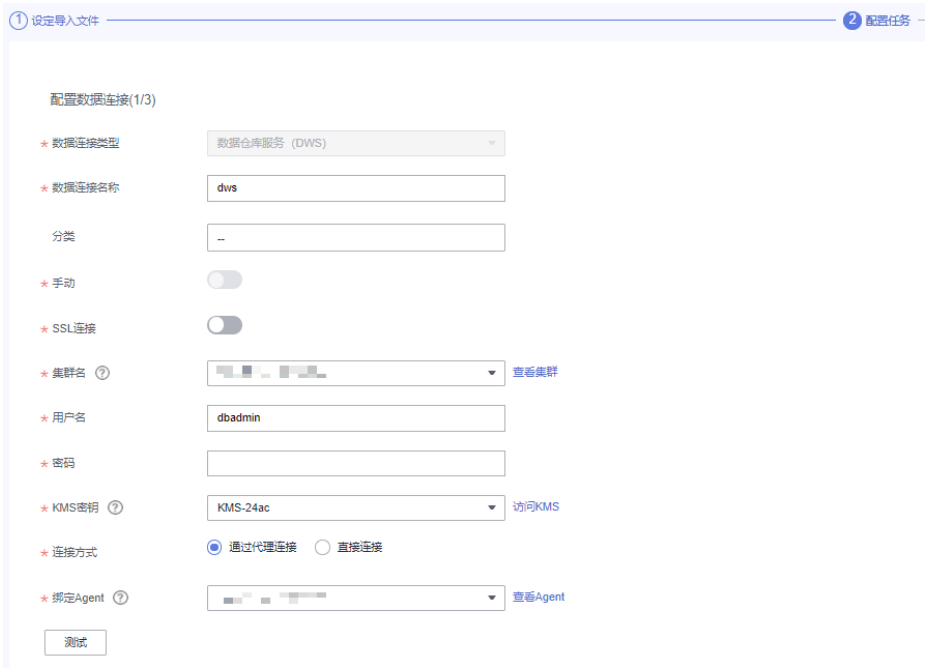


4. 单击“下一步”，勾选导入的资源类型。

图 3-11 勾选导入的资源类型

5. 如果选择导入数据源，则单击“下一步”需要配置数据连接。配置的数据连接个数由数据源的数量决定，每个连接都需要输入密码。

图 3-12 配置数据连接



6. 单击“下一步”，等待导入完成。

图 3-13 导入完成

导入资源耗时1分钟仍未显示结果则表示导入失败，请重试。如果仍然无法导入，请联系或技术支持人员协助解决。

3.2.4 使用教程

3.2.4.1 新建 MRS Hive 连接

本章节以新建MRS Hive连接为例，介绍如何建立DataArts Studio与数据湖底座之间的数据连接。

前提条件

- 在创建数据连接前，请确保您已创建所要连接的数据湖（如DataArts Studio所支持的数据库、云服务等）。
  - 在创建DWS类型的数据连接前，您需要先在DWS服务中创建集群，并且具有KMS密钥的查看权限。
  - 在创建MRS HBase、MRS Hive、MRS Kafka、MRS Spark、MRS Presto类型的数据连接前，需确保您已创建MRS集群，并且在创建数据链接时已创建选择所需要的组件。
  - 在创建RDS类型的数据连接前，请确保您已创建RDS数据库实例。DataArts Studio平台目前仅支持RDS中的MySQL和PostgreSQL数据库引擎。

- 在创建数据连接前，请确保待连接的数据湖与DataArts Studio实例之间网络互通。
  - 如果数据湖为云下的数据库，则需要通过公网或者专线打通网络，确保数据源所在的主机可以访问公网，并且防火墙规则已开放连接端口。
  - 如果数据湖为云上服务（如DWS、MRS等），则网络互通需满足如下条件：
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上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。
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。
    - 此外，您还必须确保该云服务的实例与DataArts Studio工作空间所属的企业项目必须相同，如果不同，您需要修改工作空间的企业项目。

创建数据连接

1. 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

   图 3-14 选择管理中心

2. 在管理中心页面，单击“数据连接”，进入数据连接页面。

图 3-15 创建数据连接

3. 单击“创建数据连接”，在弹出的对话框中，选择“数据连接类型”为“MapReduce服务（MRS Hive）”，并参见**表3-12**配置相关参数。

图 3-16 创建数据连接

图 3-17 MRS Hive 连接配置参数

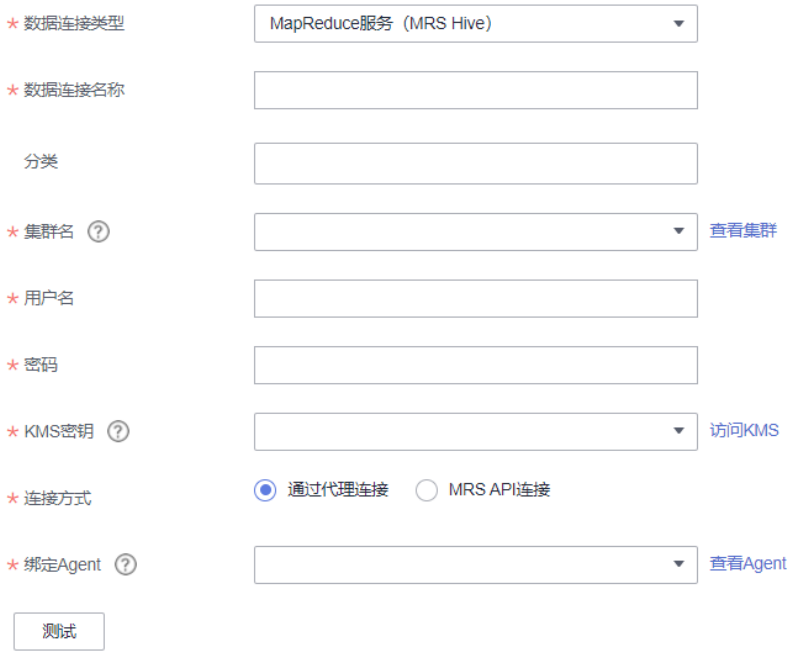


表 3-12 MRS Hive 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。
集群名	是	选择Hive所属的MRS集群。如果在下拉列表中无法显示MRS集群，请检查MRS集群与DataArts Studio实例是否网络互通。 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件： ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。 ● DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。 ● 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。
连接方式	是	选择所需的连接方式，推荐使用“通过代理连接”。 ● 通过代理连接：通过Agent（即CDM集群）进行代理，以MRS集群的用户名和密码访问MRS集群。代理连接方式支持MRS所有版本的集群。 ● MRS API连接：以MRS API的方式访问MRS集群。MRS API连接仅支持2.X及更高版本的MRS集群。选择MRS API连接时，有以下约束： 1. 无法查看表和字段。 2. 在SQL编辑器运行SQL时，只能以日志形式显示执行结果。 3. 数据治理（如数据架构、数据质量、数据目录等组件）功能无法使用MRS API连接。

参数	是否必选	说明
用户名	否	MRS集群的用户名，通过代理连接的时候，是必选项。如果使用新建的MRS用户进行连接，您需要先登录Manager页面，并更新初始密码。 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。您可以参考**创建MRS安全集群的kerberos认证用户**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** • MRS 3.1.0及之后版本集群，所创建的用户至少需具备Manager_viewer的角色权限才能在管理中心创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 • MRS 3.1.0版本之前的集群，所创建的用户需要具备Manager_administrator或System_administrator权限，才能在管理中心创建连接。 • 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。
密码	否	MRS集群的访问密码，通过代理连接的时候，是必选项。
KMS密钥	否	KMS密钥名称。通过代理连接的时候，是必选项。
绑定Agent	否	通过代理连接的时候，是必选项。 MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和MRS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和MRS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

4. 单击“测试”，测试数据连接的连通性。如果无法连通，数据连接将无法创建。
5. 测试通过后，单击“确定”，创建数据连接。

参考

1. 在创建数据连接的界面上MRS Hive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. 为什么Hive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。

3.2.4.2 新建 DWS 连接

本章节以新建DWS连接为例，介绍如何建立DataArts Studio与数据仓库底座之间的数据连接。

前提条件

- 在创建数据连接前，请确保您已创建所要连接的数据湖（如DataArts Studio所支持的数据库、云服务等）。
  - 在创建DWS类型的数据连接前，您需要先在DWS服务中创建集群，并且具有KMS密钥的查看权限。
  - 在创建MRS HBase、MRS Hive、MRS Kafka、MRS Spark、MRS Presto类型的数据连接前，需确保您已创建MRS集群，并且在创建数据链接时已创建选择所需要的组件。
  - 在创建RDS类型的数据连接前，请确保您已创建RDS数据库实例。DataArts Studio平台目前仅支持RDS中的MySQL和PostgreSQL数据库引擎。
- 在创建数据连接前，请确保待连接的数据湖与DataArts Studio实例之间网络互通。
  - 如果数据湖为云下的数据库，则需要通过公网或者专线打通网络，确保数据源所在的主机可以访问公网，并且防火墙规则已开放连接端口。
  - 如果数据湖为云上服务（如DWS、MRS等），则网络互通需满足如下条件:
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上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。
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。
    - 此外，您还必须确保该云服务的实例与DataArts Studio工作空间所属的企业项目必须相同，如果不同，您需要修改工作空间的企业项目。

创建数据连接

1. 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

图 3-18 选择管理中心

2. 在管理中心页面，单击“数据连接”，进入数据连接页面。

图 3-19 创建数据连接

3. 单击“创建数据连接”，在弹出的对话框中，选择“数据连接类型”为“数据仓库服务（DWS）”，并参见表3-13配置相关参数。

图 3-20 创建数据连接

图 3-21 DWS 连接配置参数

表 3-13 DWS 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。
手动	是	通过单击或来关闭或开启手动开关： • 当“手动”关闭时候，“IP”和“端口”不需要填写。 • 当“手动”打开时候，“IP”和“端口”需要填写。
IP	否	“手动”打开时需要填写该项，表示通过内部网络访问集群数据库的IP地址。内网访问IP地址在创建集群时自动生成。

参数	是否必选	说明
端口	否	“手动”打开时需要填写该项，表示创建DWS集群时指定的数据库端口号。请确保您已在安全组规则中开放此端口，以便DataArts Studio实例可以通过该端口连接DWS集群数据库。
SSL连接	是	DWS支持SSL通道加密和证书认证两种方式进行客户端与服务器端的通信。您可以通过服务器端是否强制使用SSL连接进行设置。开关打开，即只能通过SSL方式连接。开关关闭，即两种方式均可。默认关闭。
集群名	是	选择DWS集群。
用户名	是	数据库的用户名，创建DWS集群时指定的用户名。
密码	是	数据库的访问密码，创建DWS集群时指定的密码。
KMS密钥	是	KMS密钥名称。
连接方式	是	选择所需的连接方式，推荐使用“通过代理连接”。 ● 通过代理连接：通过Agent（即CDM集群）进行代理连接访问DWS集群。 ● 直接连接：直接访问DWS集群。
绑定Agent	否	通过代理连接的时候，是必选项。 DWS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建DWS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和DWS集群网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和DWS集群处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

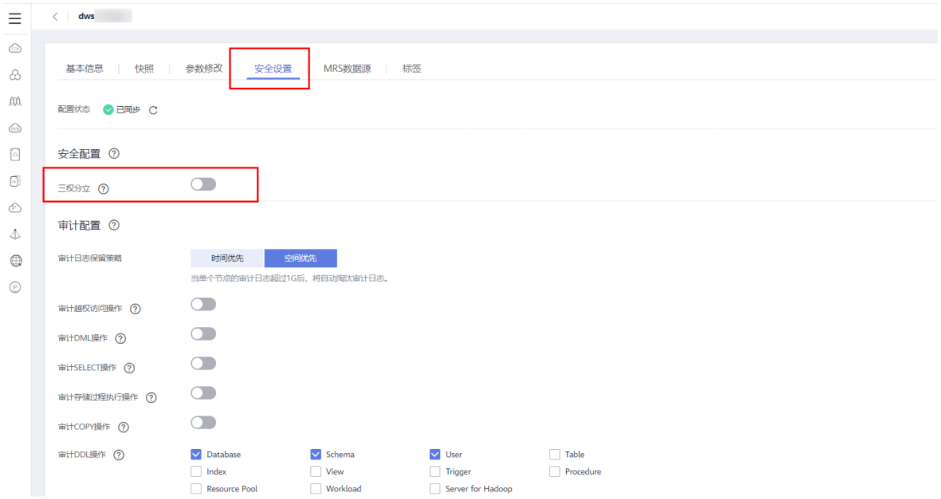
4. 单击“测试”，测试数据连接的连通性。如果无法连通，数据连接将无法创建。
5. 测试通过后，单击“确定”，创建数据连接。

参考

1. 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，点击进入对应的DWS集群后，选择“安全设置”，然后关闭三权分立功能。

图 3-22 关闭 DWS 集群三权分立功能



2. 为什么DWS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。

3.2.4.3 新建 MySQL 连接

本章节以新建MySQL连接为例，介绍如何建立DataArts Studio与数据库底座之间的数据连接。

前提条件

- 在创建数据连接前，请确保您已创建所要连接的数据湖（如DataArts Studio所支持的数据库、云服务等）。
  - 在创建DWS类型的数据连接前，您需要先在DWS服务中创建集群，并且具有KMS密钥的查看权限。
  - 在创建MRS HBase、MRS Hive、MRS Kafka、MRS Spark、MRS Presto类型的数据连接前，需确保您已创建MRS集群，并且在创建数据链接时已创建选择所需要的组件。
  - 在创建RDS类型的数据连接前，请确保您已创建RDS数据库实例。DataArts Studio平台目前仅支持RDS中的MySQL和PostgreSQL数据库引擎。
- 在创建数据连接前，请确保待连接的数据湖与DataArts Studio实例之间网络互通。
  - 如果数据湖为云下的数据库，则需要通过公网或者专线打通网络，确保数据源所在的主机可以访问公网，并且防火墙规则已开放连接端口。
  - 如果数据湖为云上服务（如DWS、MRS等），则网络互通需满足如下条件：
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上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。
    - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的

“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。

- 此外，您还必须确保该云服务的实例与DataArts Studio工作空间所属的企业项目必须相同，如果不同，您需要修改工作空间的企业项目。

创建数据连接

1. 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

图 3-23 选择管理中心

2. 在管理中心页面，单击“数据连接”，进入数据连接页面。

图 3-24 创建数据连接

3. 单击“创建数据连接”，在弹出的对话框中，选择“数据连接类型”为“RDS”，并参见**表3-14**配置相关参数。

图 3-25 创建数据连接

说明

- 不建议使用MySQL(待下线)连接器，推荐使用RDS连接MySQL数据源。
- RDS数据连接方式依赖于OBS。如果没有与DataArts Studio同区域的OBS，则不支持RDS数据连接。

图 3-26 RDS 连接配置参数

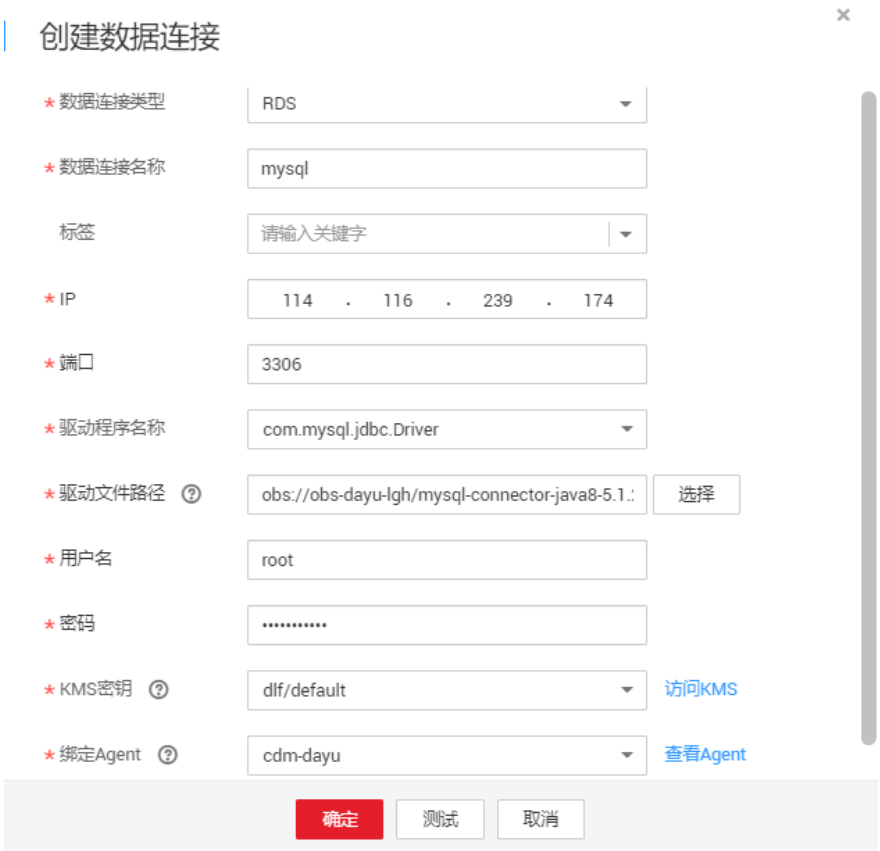


表 3-14 RDS 数据连接

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含英文字母、数字、下划线和中划线，且长度为1~50个字符。
标签	否	标识数据连接的属性。设置标签后，便于统一管理。 **说明** 标签的名称，只能包含中文、英文字母、数字和下划线，不能以下划线开头。且长度不能超过100个字符。
IP	是	RDS的访问地址。 如果为RDS数据源，可以通过RDS管理控制台获取访问地址： 1. 根据创建的帐号登录管理控制台。 2. 单击“云数据库 RDS”，从左侧列表选择实例管理。 3. 单击某一个实例名称，进入实例基本信息页面。 在连接信息标签中可以获取到内网地址。
端口	是	RDS的访问端口。 如果为RDS数据源，可以通过RDS管理控制台获取访问端口： 1. 根据的帐号登录管理控制台。 2. 单击“云数据库 RDS”，左侧列表选择实例管理。 3. 单击某一个实例名称，进入实例基本信息页面。 在连接信息标签中可以获取到数据库端口。
驱动程序名称	是	驱动程序名称： • com.mysql.jdbc.Driver • org.postgresql.Driver
驱动文件路径	是	驱动文件在OBS上的路径。需要您自行到官网下载.jar格式驱动并上传至OBS中。 • MySQL驱动：获取地址**https://downloads.mysql.com/archives/c-j/**，建议5.1.48版本。 • PostgreSQL驱动：获取地址**https://jdbc.postgresql.org/download**，建议42.1.4版本。 **说明** 如果需要更新驱动文件，则需要先在数据集成页面重启CDM集群，然后通过编辑数据连接的方式重新选择新版本驱动，更新驱动才能生效。
用户名	是	数据库的用户名，创建集群的时候，输入的用户名。
密码	是	数据库的访问密码，创建集群的时候，输入的密码。

参数	是否必选	说明
KMS密钥	是	KMS密钥名称。 通过KMS管理控制台获取密钥名称： 1. 根据的帐号登录管理控制台。 2. 单击“密钥管理服务”，左侧列表选择密钥管理。 在密钥列表可以获取到密钥名称。
绑定Agent	是	RDS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建RDS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请先通过数据集成增量包进行创建。 CDM集群作为网络代理，必须和RDS网络互通才可以成功创建MRS连接，为确保两者网络互通，CDM集群必须和RDS处于相同的区域、可用区、VPC和子网，安全组规则需允许两者网络互通。

4. 单击“测试”，测试数据连接的连通性。如果无法连通，数据连接将无法创建。
5. 测试通过后，单击“确定”，创建数据连接。

参考

1. 创建RDS类型的数据连接时，需要注意哪些事项?

   创建RDS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。

3.3 数据集成

3.3.1 数据集成概述

DataArts Studio数据集成是一种高效、易用的数据集成服务，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。

数据集成即云数据迁移（Cloud Data Migration，后简称CDM）服务，本文中的“云数据迁移”、“CDM”均指“数据集成”。

您可以通过以下方式之一进入CDM主界面：

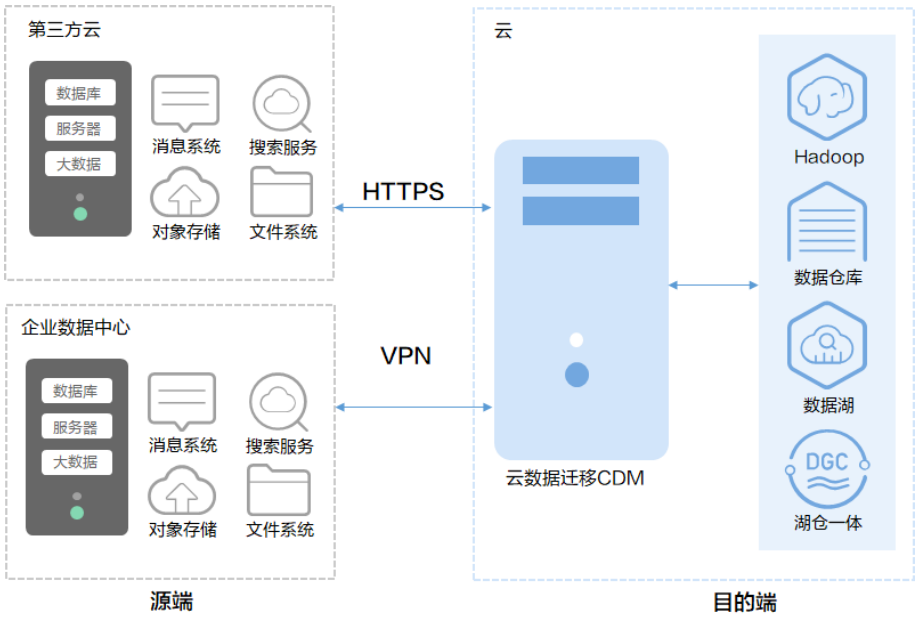
- 登录CDM控制台，单击“集群管理”，进入到CDM主界面。
- 登录DataArts Studio控制台。选择对应工作空间的“数据集成”模块，进入CDM主界面。

图 3-27 选择数据集成

云数据迁移简介

云数据迁移基于分布式计算框架，利用并行化处理技术，支持用户稳定高效地对海量数据进行移动，实现不停服数据迁移，快速构建所需的数据架构。

图 3-28 数据集成定位



产品功能

- **表/文件/整库迁移**

  支持批量迁移表或者文件，还支持同构/异构数据库之间整库迁移，一个作业即可迁移几百张表。

- **增量数据迁移**

支持文件增量迁移、关系型数据库增量迁移、HBase/CloudTable增量迁移，以及使用Where条件配合时间变量函数实现增量数据迁移。

- **事务模式迁移**

  支持当CDM作业执行失败时，将数据回滚到作业开始之前的状态，自动清理目的表中的数据。

- **字段转换**

  支持去隐私、字符串操作、日期操作等常用字段的数据转换功能。

- **文件加密**

  在迁移文件到文件系统时，CDM支持对写入云端的文件进行加密。

- **MD5校验一致性**

  支持使用MD5校验，检查端到端文件的一致性，并输出校验结果。

- **脏数据归档**

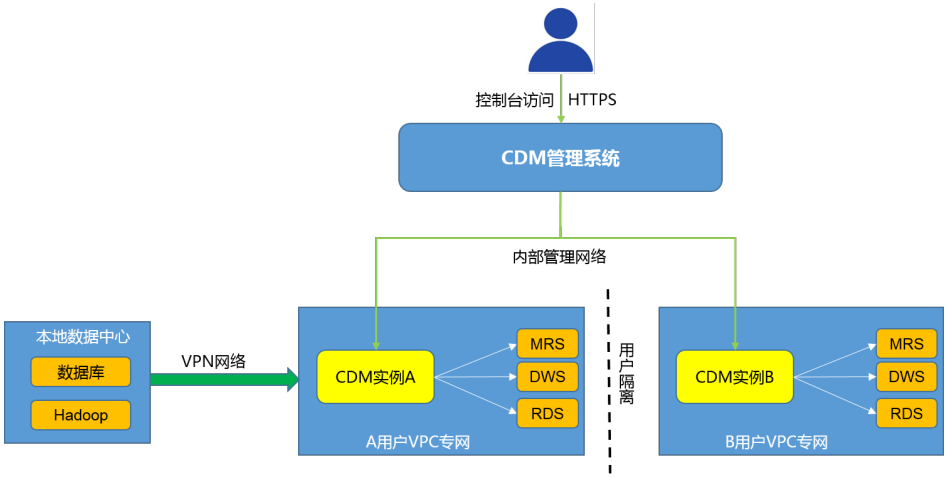
  支持将迁移过程中处理失败的、被清洗过滤掉的、不符合字段转换或者不符合清洗规则的数据单独归档到脏数据日志中，便于用户查看。并支持设置脏数据比例阈值，来决定任务是否成功。

CDM 迁移原理

用户使用CDM服务时，CDM管理系统在用户VPC中发放全托管的CDM实例。此实例仅提供控制台和Rest API访问权限，用户无法通过其他接口（如SSH）访问实例。这种方式保证了CDM用户间的隔离，避免数据泄漏，同时保证VPC内不同云服务间数据迁移时的传输安全。用户还可以使用VPN网络将本地数据中心的数据迁移到云服务，具有高度的安全性。

CDM数据迁移以抽取-写入模式进行。CDM首先从源端抽取数据然后将数据写入到目的端，数据访问操作均由CDM主动发起，对于数据源（如RDS数据源）支持SSL时，会使用SSL加密传输。迁移过程要求用户提供源端和目的端数据源的用户名和密码，这些信息将存储在CDM实例的数据库中。保护这些信息对于CDM安全至关重要。

图 3-29 CDM 迁移原理



3.3.2 约束与限制

CDM 系统级限制和约束

1. 集群创建好以后不支持修改规格，如果需要使用更高规格的，需要重新创建一个集群。
2. ARM版本的CDM集群不支持Agent功能。CDM集群为ARM或X86版本，依赖于底层资源的架构。
3. CDM暂不支持控制迁移数据的速度，请避免在业务高峰期执行迁移数据的任务。
4. 当前CDM集群cdm.large实例规格网卡的基准/最大带宽为0.8/3 Gbps，单个实例一天传输数据量的理论极限值在8TB左右。同理，cdm.xlarge实例规格网卡的基准/最大带宽为4/10 Gbps，理论极限值在40TB左右；cdm.4xlarge实例规格网卡的基准/最大带宽为36/40 Gbps，理论极限值在360TB左右。对传输速度有要求的情况下可以使用多个数据集成实例实现。

   上述数据量为理论极限值，实际传输数据量受数据源类型、源和目的数据源读写性能、带宽等多方面因素制约，实测cdm.large规格最大可达到约8TB每天（大文件迁移到OBS场景）。推荐用户在正式迁移前先用小数据量实测进行速度摸底。
5. 迁移文件或对象时支持文件级增量迁移（通过配置跳过重复文件实现），但不支持断点续传。

   例如要迁移3个文件，第2个文件迁移到一半时由于网络原因失败，再次启动迁移任务时，会跳过第1个文件，从第2个文件开始重新传，但不能从第2个文件失败的位置重新传。
6. 文件迁移时，单个任务支持千万数量的文件，如果待迁移目录下文件过多，建议拆分到不同目录并创建多个任务。
7. 用户在CDM上配置的连接和作业支持导出到本地保存，考虑到密码的安全性，CDM不会将对应数据源的连接密码导出。因此在将作业配置重新导入到CDM前，需要手工编辑导出的JSON文件补充密码或在导入窗口配置密码。
8. 不支持集群自动升级到新版本，需要用户通过作业的导出和导入功能，实现升级到新版本。
9. 在无OBS的场景下，CDM系统不会自动备份用户的作业配置，需要用户通过作业的导出功能进行备份。
10. 如果配置了VPC对等连接，可能会出现对端VPC子网与CDM管理网重叠，从而无法访问对端VPC中数据源的情况。推荐使用公网做跨VPC数据迁移，或联系管理员在CDM后台为VPC对等连接添加特定路由。
11. CDM迁移，当目的端为DWS和NewSQL的时候，不支持将源端的主键和唯一索引等约束一起迁移过去。
12. CDM迁移作业时，需确保两个集群版本的JSON文件格式保持一致，才可以从将源集群的作业导入到目标集群。

数据库迁移通用限制和约束

1. CDM以批量迁移为主，仅支持有限的数据库增量迁移，不支持数据库实时增量迁移。
2. CDM支持的数据库整库迁移，仅支持数据表迁移，不支持存储过程、触发器、函数、视图等数据库对象迁移。

   CDM仅适用于一次性将数据库迁移到云上的场景，包括同构数据库迁移和异构数据库迁移，不适合数据同步场景，比如容灾、实时同步。

3. CDM迁移数据库整库或数据表失败时，已经导入到目标表中的数据不会自动回滚，对于需要事务模式迁移的用户，可以配置“先导入到阶段表”参数，实现迁移失败时数据回滚。

   极端情况下，可能存在创建的阶段表或临时表无法自动删除，也需要用户手工清理（阶段表的表名以“_cdm_stage”结尾，例如：cdmtet_cdm_stage）。
4. CDM访问用户本地数据中心数据源时（例如本地自建的MySQL数据库），需要用户的数据源可支持Internet公网访问，并为CDM集群实例绑定弹性IP。这种方式下安全实践是：本地数据源通过防火墙或安全策略仅允许CDM弹性IP访问。
5. 仅支持常用的数据类型，字符串、数字、日期，对象类型有限支持，如果对象过大会出现无法迁移的问题。
6. 仅支持数据库字符集为GBK和UTF-8。
7. 字段名不可使用&和%。

关系数据库迁移权限配置

常见关系数据库迁移需要的最小权限级：

- MySQL：INFORMATION_SCHEMA库的读权限，以及对数据表的读权限。
- Oracle：需要该用户有resource角色，并在tablespace下有数据表的select权限。
- 达梦：具有该schema下select any table的权限。
- DWS：需要表的schema usage权限和数据表的查询权限。
- SQL Server：用户需要有sysadmin权限。
- PostgreSQL：角色拥有数据库下schema下表的select权限。

FusionInsight HD 和 Apache Hadoop 数据源约束

FusionInsight HD和Apache Hadoop数据源在用户本地数据中心部署时，由于读写Hadoop文件需要访问集群的所有节点，需要为每个节点都放通网络访问。

数据仓库服务(DWS)和 FusionInsight LibrA 数据源约束

1. 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
2. DWS字符类型字段认为空字符串（''）是空值，有非空约束的字段无法插入空字符串（''），这点与MySQL行为不一致，MySQL不认为空字符串（''）是空值。从MySQL迁移到DWS时，可能会因为上述原因导致迁移失败。
3. 使用GDS模式快速导入数据到DWS时，需要配置相关安全组或防火墙策略，允许DWS/LibrA的数据节点访问CDM IP地址的25000端口。
4. 使用GDS模式导入数据到DWS时，CDM会自动创建外表（foreign table）用于数据导入，表名以UUID结尾（例如：

cdmtest_aecf3f8n0z73dsl72d0d1dk4lcir8cd），作业失败正常会自动删除，极端情况下可能需要用户手工清理。

对象存储服务（OBS）数据源约束

1. 迁移文件时系统会自动并发，任务配置中的“抽取并发数”无效。
2. 不支持断点续传。CDM传文件失败会产生OBS碎片，需要用户到OBS控制台清理碎片文件避免空间占用。
3. 不支持对象多版本的迁移。
4. 增量迁移时，单个作业的源端目录下的文件数量或对象数量，根据CDM集群规格分别有如下限制：大规格集群30万、中规格集群20万、小规格集群10万。

   如果单目录下文件或对象数量超过限制，需要按照子目录来拆分成多个迁移作业。

DLI 数据源约束

使用CDM服务迁移数据到DLI时，当前用户需拥有OBS的读取权限。

Oracle 数据源约束

不支持Oracle实时增量数据同步。

分布式缓存服务（DCS）和 Redis 数据源约束

1. 由于分布式缓存服务（DCS）限制了获取所有Key的命令，CDM无法支持DCS作为源端，但可以作为迁移目的端，第三方云的Redis服务也无法支持作为源端。如果是用户在本地数据中心或ECS上自行搭建的Redis支持作为源端或目的端。
2. 仅支持Hash和String两种数据格式。

文档数据库服务（DDS）和 MongoDB 数据源约束

从MongoDB、DDS迁移数据时，CDM会读取集合的首行数据作为字段列表样例，如果首行数据未包含该集合的所有字段，用户需要自己手工添加字段。

云搜索服务和 Elasticsearch 数据源约束

1. CDM支持自动创建索引和类型，索引和类型名称只能全部小写，不能有大写。
2. 索引下的字段类型创建后不能修改，只能创建新字段。

   如果一定要修改字段类型，需要创建新索引或到Kibana上用Elasticsearch命令删除当前索引重新创建（数据也会删除）。
3. CDM自动创建的索引，字段类型为date时，要求数据格式为“yyyy-MM-dd HH:mm:ss.SSS Z”，即“2018-08-08 08:08:08.888 +08:00”。

   迁移数据到云搜索服务时如果date字段的原始数据不满足格式要求，可以通过CDM的表达式转换功能转换为上述格式。

Kafka 数据源约束

1. 消息体中的数据是一条类似CSV格式的记录，可以支持多种分隔符。不支持二进制格式或其他格式的消息内容解析。

表格存储服务（CloudTable）和 HBase 数据源约束

1. CloudTable或HBase作为源端时，CDM会读取表的首行数据作为字段列表样例，如果首行数据未包含该表的所有字段，用户需要自己手工添加字段。
2. 由于HBase的无Schema技术特点，CDM无法获知数据类型，如果数据内容是使用二进制格式存储的，CDM会无法解析。

Hive 数据源约束

Hive作为迁移的目的时，如果存储格式为Textfile，在Hive创建表的语句中需要显式指定分隔符。例如：

```
CREATE TABLE csv_tbl(
smallint_value smallint,
tinyint_value tinyint,
int_value int,
bigint_value bigint,
float_value float,
double_value double,
decimal_value decimal(9, 7),
timestmamp_value timestamp,
date_value date,
varchar_value varchar(100),
string_value string,
char_value char(20),
boolean_value boolean,
binary_value binary,
varchar_null varchar(100),
string_null string,
char_null char(20),
int_null int
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = "\t",
"quoteChar"     = "'",
"escapeChar"    = "\\"
)
STORED AS TEXTFILE;
```

3.3.3 支持的数据源

数据集成有两种迁移方式，支持的数据源有所不同：

- 表/文件迁移：适用于数据入湖和数据上云场景下，表或文件级别的数据迁移，请参见**表/文件迁移支持的数据源类型**。
- 整库迁移：适用于数据入湖和数据上云场景下，离线或自建数据库整体迁移场景，请参见**整库迁移支持的数据源类型**。
- 另外，本章还列举了一些常见数据库迁移时所支持的数据类型，请参见**开源MySQL数据库迁移时支持的数据类型**、**Oracle数据库迁移时支持的数据类型**和**SQL Server数据库迁移时支持的数据类型**。

表/文件迁移支持的数据源类型

表/文件迁移可以实现表或文件级别的数据迁移。

表/文件迁移时支持的数据源如**表3-15**所示。

表 3-15 表/文件迁移支持的数据源

数据源分类	源端数据源	对应的目的端数据源	说明
数据仓库	数据仓库服务（DWS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）		-
Hadoop	MRS HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 支持本地存储，仅MRS Hive支持存算分离场景。 • 仅MRS Hive支持Ranger场景。 • 不支持ZK开启SSL场景。 • MRS HDFS建议使用的版本： – 2.8.X – 3.1.X • MRS HBase建议使用的版本： – 2.1.X – 1.3.X • MRS Hive暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X
	MRS HBase		
	MRS Hive		
	FusionInsight HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS）	• FusionInsight数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。
	FusionInsight HBase		

数据源分类	源端数据源	对应的目的端数据源	说明
	FusionInsight Hive	• NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • FusionInsight HDFS建议使用的版本： – 2.8.X – 3.1.X • FusionInsight HBase建议使用的版本： – 2.1.X – 1.3.X • FusionInsight Hive建议使用的版本： – 1.2.X – 3.1.X
	Apache HBase Apache Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• Apache数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • Apache HBas建议使用的版本： – 2.1.X – 1.3.X • Apache Hive暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • Apache HDFS建议使用的版本：

数据源分类	源端数据源	对应的目的端数据源	说明
	Apache HDFS		– 2.8.X – 3.1.X
对象存储	对象存储服务（OBS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。
文件系统	FTP	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 文件系统不支持作为目的端。 • FTP/SFTP到搜索的迁移仅支持如CSV等文本文件，不支持二进制文件。 • 文件系统到OBS的迁移推荐使用obsutil工具。
	SFTP		
	HTTP	Hadoop：MRS HDFS	
关系型数据库	云数据库MySQL	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server • 搜索：Elasticsearch，云搜索服务（CSS）	• OLTP数据库之间的迁移推荐通过数据复制服务DRS进行迁移。 • 云数据库 MySQL不支持SSL模式。 • Microsoft SQL Server建议使用的版本：2005以上。
	云数据库PostgreSQL		
	云数据库SQL Server		
	MySQL	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	PostgreSQL		
	Microsoft SQL Server		
	Oracle		

数据源分类	源端数据源	对应的目的端数据源	说明
	SAP HANA	• 数据仓库：数据湖探索（DLI） • Hadoop：MRS Hive	SAP HANA数据源存在如下约束： • SAP HANA不支持作为目的端。 • 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	分库	• 数据仓库：数据湖探索（DLI） • Hadoop：MRS HBase，MRS Hive • 搜索：Elasticsearch，云搜索服务（CSS） • 对象存储：对象存储服务（OBS）	分库数据源不支持作为目的端。
NoSQL	分布式缓存服务（DCS）	Hadoop：MRS HDFS，MRS HBase，MRS Hive	除了表格存储服务（CloudTable）外，其他NoSQL数据源不支持作为目的端。
	Redis		
	文档数据库服务（DDS）		
	MongoDB		

数据源分类	源端数据源	对应的目的端数据源	说明
	表格存储服务（CloudTable）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Cassandra	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
消息系统	Apache Kafka	搜索：云搜索服务（CSS）	消息系统不支持作为目的端。
	DMS Kafka		
	MRS Kafka	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• MRS Kafka不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。

数据源分类	源端数据源	对应的目的端数据源	说明
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash。

说明

上表中非云服务的数据源，例如MySQL，既可以支持用户本地数据中心自建的MySQL，也可以是用户在ECS上自建的MySQL，还可以是第三方云的MySQL服务。

整库迁移支持的数据源类型

整库迁移适用于将本地数据中心或在ECS上自建的数据库，同步到云上的数据库服务或大数据服务中，适用于数据库离线迁移场景，不适用于在线实时迁移。

数据集成支持整库迁移的数据源如表3-16所示。

表 3-16 整库迁移支持的数据源

数据源分类	数据源	读取	写入	说明
数据仓库	数据仓库服务（DWS）	支持	支持	-
	FusionInsight LibrA	支持	不支持	-
Hadoop（仅支持本地存储，不支持存算分离场景，不支持Ranger场景，不支持ZK开启SSL场景）	MRS HBase	支持	支持	整库迁移仅支持导出到MRS HBase。 建议使用的版本： • 2.1.X • 1.3.X

数据源分类	数据源	读取	写入	说明
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	FusionInsight HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	FusionInsight Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	Apache HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	Apache Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
关系数据库	云数据库 MySQL	支持	支持	不支持OLTP到OLTP迁移，此场景推荐通过数据复制服务DRS进行迁移。
	云数据库 PostgreSQL	支持	支持	
	云数据库 SQL Server	支持	支持	
	MySQL	支持	不支持	
	PostgreSQL	支持	不支持	
	Microsoft SQL Server	支持	不支持	

数据源分类	数据源	读取	写入	说明
	Oracle	支持	不支持	
	SAP HANA	支持	不支持	• 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	MYCAT	支持	不支持	-
	达梦数据库 DM	支持	不支持	仅支持导出到DWS、Hive
NoSQL	分布式缓存服务（DCS）	不支持	支持	仅支持MRS到DCS迁移。
	文档数据库服务（DDS）	支持	支持	仅支持DDS和MRS之间迁移。
	表格存储服务（CloudTable）	支持	支持	-

开源 MySQL 数据库迁移时支持的数据类型

源端为开源MySQL数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 3-17 开源 MySQL 数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	Hive	DWS
字符串	CHAR(M)	固定长度的字符串是以长度为1到255之间个字符长度(例如: CHAR(5))，存储右空格填充到指定的长度。 限定长度不是必需的，它会默认为1。	‘a’或‘aaaaa’	CHAR	CHAR
	VARCHAR(M)	可变长度的字符串是以长度为1到255之间字符数(高版本的MySQL超过255); 例如： VARCHAR(25). 创建VARCHAR类型字段时，必须定义长度。	‘a’或‘aaaaa’	VARCHAR	VARCHAR
数值	DECIMAL(M,D)	非压缩浮点数不能是无符号的。在解包小数，每个小数对应于一个字节。 定义显示长度(M)和小数(D)的数量是必需的。 NUMERIC是DECIMAL的同义词。	52.36	DECIMAL	D为0时对应BIGINT D不为0时对应NUMBERIC
	NUMBERIC	与 DECIMAL 相同	-	DECIMAL	NUMBERIC
	INTEGER	一个正常大小的整数，可以带符号。如果是有符号的，它允许的范围是从-2147483648到2147483647。 如果是无符号，允许的范围是从0到4294967295。 可以指定多达11位的宽度。	5236	INT	INTEGER
	INTEGER UNSIGNED	INTEGER 的无符号形式	-	BIGINT	INTEGER
	INT	与INTEGER相同	5236	INT	INTEGER

类别	类型	简要释义	存储格式示例	Hive	DWS
	INT UNSIGNED	与INTEGER UNSIGNED相同	-	BIGINT	INTEGER
	BIGINT	一个大的整数，可以带符号。如果有符号，允许范围为-9223372036854775808到9223372036854775807。如果无符号，允许的范围是从0到18446744073709551615. 可以指定最多20位的宽度。	5236	BIGINT	BIGINT
	BIGINT UNSIGNED	BIGINT的无符号形式	-	BIGINT	BIGINT
	MEDIUMINT	一个中等大小的整数，可以带符号。如果有符号，允许范围为-8388608至8388607。 如果无符号，允许的范围是从0到16777215，可以指定最多9位的宽度。	-128、127	INT	INTEGER
	MEDIUMINT UNSIGNED	MEDIUMINT的无符号形式	-	BIGINT	INTEGER
	TINYINT	一个非常小的整数，可以带符号。如果是有符号，它允许的范围是从-128到127。 如果是无符号，允许的范围是从0到255，可以指定多达4位数的宽度。	100	TINYINT	SMALLINT
	TINYINT UNSIGNED	TINYINT的无符号形式	-	TINYINT	SMALLINT

类别	类型	简要释义	存储格式示例	Hive	DWS
	BOOL	MySQL的bool实际上就是tinyint(1)	-128、127	SMALLINT	BYTEA
	SMALLINT	一个小的整数，可以带符号。如果有符号，允许范围为-32768至32767。 如果无符号，允许的范围是从0到65535，可以指定最多5位的宽度。	9999	SMALLINT	SMALLINT
	SMALLINT UNSIGNED	SMALLINT的无符号形式	-	INT	SMALLINT
	REAL	同DOUBLE	-	DOUBLE	-
	FLOAT(M,D)	不能使用无符号的浮点数字。可以定义显示长度(M)和小数位数(D)。这不是必需的，并且默认为10,2。其中2是小数的位数，10是数字(包括小数)的总数。小数精度可以到24个浮点。	52.36	FLOAT	FLOAT4
	DOUBLE(M,D)	不能使用无符号的双精度浮点数。可以定义显示长度(M)和小数位数(D)。 这不是必需的， 默认为16,4，其中4是小数的位数。小数精度可以达到53位的DOUBLE。 REAL是DOUBLE同义词。	52.36	DOUBLE	FLOAT8
	DOUBLE PRECISION	与DOUBLE相似	52.3	DOUBLE	FLOAT8
位	BIT(M)	存储位值的BIT类型。BIT(M)可以存储多达M位的值，M的范围在1到64之间。	B'1111100' B'1100'	TINYINT	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
日期时间	DATE	以YYYY-MM-DD格式的日期，在1000-01-01和9999-12-31之间。 例如，1973年12月30日将被存储为1973-12-30。	1999-10-01	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息	'09:10:21'或'9:10:21'	不支持（String）	TIME
	DATETIME	日期和时间组合以YYYY-MM-DD HH:MM:SS格式，在1000-01-01 00:00:00到9999-12-31 23:59:59之间。例如，1973年12月30日下午3:30，会被存储为1973-12-30 15:30:00。	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	1970年1月1日午夜之间的时间戳，到2037的某个时候。这看起来像前面的DATETIME格式，无需只是数字之间的连字符; 1973年12月30日下午3点30分将被存储为19731230153000(YYYYMMDDHHMMSS)。	19731230153000	TIMESTAMP	TIMESTAMP
	YEAR(M)	以2位或4位数字格式来存储年份。如果长度指定为2(例如YEAR(2))，年份就可以为1970至2069(70～69)。如果长度指定为4，年份范围是1901-2155，默认长度为4。	2000	不支持（String）	不支持
多媒体（二进制）	BINARY(M)	字节数为M,允许长度为0-M的变长二进制字符串，字节数为值得长度加1	0x2A3B4058 (二进制数据)	不支持	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
	VARBINARY(M)	字节数为M,允许长度为0-M的定长二进制字符串	0x2A3B4059 (二进制数据)	不支持	BYTEA
	TEXT	字段的最大长度是65535个字符。 TEXT是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。	0x5236(二进制数据)	不支持	不支持
	TINYTEXT	0-255字节短文本二进制字符串	-	-	不支持
	MEDIUMTEXT	0-167772154字节中等长度文本二进制字符串	-	-	不支持
	LONGTEXT	0-4294967295字节极大长度文本二进制字符串	-	-	不支持
	BLOB	字段的最大长度是65535个字符。 BLOB是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。BLOB大小写敏感。	0x5236(二进制数据)	不支持	BYTEA
	TINYBLOB	0-255字节短文本二进制字符串	-	-	BYTEA
	MEDIUMBLOB	0-167772154字节中等长度文本二进制字符串	-	-	BYTEA
	LONGBLOB	0-4294967295字节极大长度文本二进制字符串	0x5236(二进制数据)	不支持	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
特殊类型	SET	SET是一个字符串对象，可以有零或多个值，其值来自表创建时规定的允许的一列值。指定包括多个SET成员的SET列值时各成员之间用逗号(‘,’)间隔开。这样SET成员值本身不能包含逗号。	-	-	不支持
	JSON	-	-	不支持	不支持（TEXT）
	ENUM	当定义一个ENUM，要创建它的值的列表，这些是必须用于选择的项(也可以是NULL)。例如，如果想要字段包含“A”或“B”或“C”，那么可以定义为ENUM为ENUM(“A”，“B”，“C”)也只有这些值(或NULL)才能用来填充这个字段。	-	不支持	不支持

Oracle 数据库迁移时支持的数据类型

源端为Oracle数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 3-18 Oracle 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS
字符串	char	定长字符串，会用空格填充来达到最大长度。	CHAR	CHAR
	nchar	包含unicode格式数据的定长字符串。	CHAR	CHAR
	varchar2	是VARCHAR的同义词。这是一个变长字符串，与CHAR类型不同，它不会用空格将字段或变量填充至最大长度。	VARCHAR	VARCHAR
	nvarchar2	包含unicode格式数据的变长字符串。	VARCHAR	VARCHAR
数值	number	能存储精度最多高达38位的数字	DECIMAL	NUMERIC

类别	类型	简要释义	Hive	DWS
	binary_float	2位单精度浮点数	FLOAT	FLOAT8
	binary_double	64位双精度浮点数	DOUBLE	FLOAT8
	long	能存储最多2GB的字符数据	不支持	不支持
日期时间	date	7字节的定宽日期/时间数据类型，其中包含7个属性：世纪、世纪中的哪一年、月份、月中的哪一天、小时、分钟、秒。	DATE	TIMESTAMP
	timestamp	7字节或11字节的定宽日期/时间数据类型，它包含小数秒	TIMESTAMP	TIMESTAMP
	timestamp with time zone	3字节的timestamp，提供了时区支持。	TIMESTAMP	TIME WITH TIME ZONE
	timestamp with local time zone	7字节或11字节的定宽日期/时间数据类型，在数据的插入和读取时会发生时区转换	TIMESTAMP	不支持（TEXT）
	interval year to month	5字节的定宽数据类型，用于存储一个时段。	不支持	不支持（TEXT）
	interval day to second	11字节的定宽数据类型，用于存储一个时段。将时段存储为天/小时/分钟/秒数，还可以有9位小数秒。	不支持	不支持（TEXT）
多媒体（二进制）	raw	一种变长二进制数据类型，采用这种数据类型存储的数据不会发生字符集转换。	不支持	不支持
	long raw	能存储多达2GB的二进制信息	不支持	不支持
	blob	能够存储最多4GB的数据	不支持	不支持
	clob	在Oracle 10g及以后的版本中允许存储最多（4GB）×（数据库块大小）字节的数据。CLOB包含要进行字符集转换的信息。这种数据类型很适合存储纯文本信息。	不支持	不支持

类别	类型	简要释义	Hive	DWS
	nclob	这种类型能够存储最多4GB的数据。当字符集发生转换时，这种类型会受到影响。	不支持	不支持
	bfile	可以在数据库列中存储一个oracle目录对象和一个文件名，我们可以通过它来读取这个文件。	不支持	不支持
其他类型	rowid	实际上是数据库表中行的地址，它有10字节长。	不支持	不支持
	urowid	是一个通用的rowid，没有固定的rowid的表。	不支持	不支持

SQL Server 数据库迁移时支持的数据类型

源端为SQL Server数据库，目的端为Hive、DWS、Oracle时，支持的数据类型如下：

表 3-19 SQL Server 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS	Oracle
字符串数据类型	char	定长字符串，会用空格填充来达到最大长度。	CHAR	CHAR	CHAR
	nchar	包含unicode格式数据的定长字符串。	CHAR	CHAR	CHAR
	varchar	可变长度的字符串是以长度为1到255之间字符数(高版本的MySQL超过255); 例如： VARCHAR(25)；创建VARCHAR类型字段时，必须定义长度。	VARCHAR	VARCHAR	VARCHAR
	nvarchar	与varchar类似，存储可变长度Unicode字符数据。	VARCHAR	VARCHAR	VARCHAR
数值数据类型	int	int存储在4个字节中,其中一个二进制位表示符号位，其它31个二进制位表示长度和大小，可以表示-2的31次方~2的31次方-1范围内的所有整数。	INT	INTEGER	INT
	bigint	bigint存储在8个字节中，其中一个二进制位表示符号位，其它63个二进制位表示长度和大小，可以表示-2的63次方~2的63次方-1范围内的所有整数。	BIGINT	BIGINT	NUMBER

类别	类型	简要释义	Hive	DWS	Oracle
	smallint	smallint类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。	SMALLINT	SMALLINT	NUMBER
	tinyint	tinyint类型的数据占用了一个字节的存储空间，可以表示0~255范围内的所有整数。	TINYINT	TINYINT	NUMBER
	real	可以存储正的或者负的十进制数值。	DOUBLE	FLOAT4	NUMBER
	float	其中为用于存储float数值尾数的位数（以科学计数法表示），因此可以确定精度和存储大小。	FLOAT	FLOAT8	binary_float
	decimal	带固定精度和小数位数的数值数据类型。	DECIMAL	NUMERIC	NUMBER
	numeric	用于存储零、正负定点数	DECIMAL	NUMERIC	NUMBER
日期时间数据类型	date	存储用字符串表示的日期数据。	DATE	TIMESTAMP	DATE
	time	以字符串形式记录一天的某个时间。	不支持（String）	TIME	不支持
	datetime	用于存储时间和日期数据。	TIMESTAMP	TIMESTAMP	不支持
	datetime2	datetime的扩展类型，其数据范围更大，默认的最小精度最高，并具有可选的用户定义的精度。	TIMESTAMP	TIMESTAMP	不支持
	smalldatetime	smalldatetime类型与datetime类型相似，只是其存储范围是从1900年1月1日到2079年6月6日，当日期时间精度较小时，可以使用smalldatetime,该类型数据占用4个字节的存储空间。	TIMESTAMP	TIMESTAMP	不支持
	timestamp	时间戳数据类型	TIMESTAMP	TIMESTAMP	TIMESTAMP
	datetimeoffset	用于定义一个采用24小时制与日期相组合并可识别时区的时间。	不支持（String）	TIMESTAMP	不支持

类别	类型	简要释义	Hive	DWS	Oracle
多媒体数据类型（二进制）	text	用于存储文本数据。	不支持（String）	不支持（String）	不支持
	netxt	与text类型作用相同，为长度可变的非Unicode数据。	不支持（String）	不支持（String）	不支持
	image	长度可变的二进制数据，用于存储照片、目录图片或者图画。	不支持（String）	不支持（String）	不支持
	binary	长度为n个字节的固定长度二进制数据，其中n是从1~8000的值。	不支持（String）	不支持（String）	不支持
	varbinary	可变长度二进制数据。	不支持（String）	不支持（String）	不支持
货币数据类型	money	用于存储货币值	不支持（String）	不支持（String）	不支持
	smallmoney	与money类型相似，输入数据时在前面加上一个货币符号，如人民币为￥或其它定义的货币符号。	不支持（String）	不支持（String）	不支持
位数据类型	bit	位数据类型，只取0或1为值，长度1字节。bit值经常当作逻辑值用于判断true(1)或false(0),输入非0值时系统将其替换为1。	不支持	不支持	不支持
其他数据类型	rowversion	每个数据都有一个计数器，当对数据库中包含rowversion列的表执行插入或者更新操作时，该计数器数值就会增加。	不支持	不支持	不支持
	uniqueidentifier	16字节的GUID(Globally Unique Identifier,全球唯一标识符)，是Sql Server根据网络适配器地址和主机CPU时钟产生的唯一号码，其中，每个为都是0~9或a~f范围内的十六进制数字。	不支持	不支持	不支持
	cursor	游标数据类型。	不支持	不支持	不支持
	sql_variant	用于存储除文本，图形数据和timestamp数据外的其它任何合法的Sql Server数据，可以方便Sql Server的开发工作。	不支持	不支持	不支持
	table	用于存储对表或视图处理后的结果集。	不支持	不支持	不支持

类别	类型	简要释义	Hive	DWS	Oracle
	xml	存储xml数据的数据类型。可以在列中或者xml类型的变量中存储xml实例。存储的xml数据类型表示实例大小不能超过2GB。	不支持	不支持	不支持

3.3.4 管理集群

3.3.4.1 创建 CDM 集群

CDM采用独立集群的方式为用户提供安全可靠的数据迁移服务，各集群之间相互隔离，不可相互访问。

CDM集群可用于如下场景：

- 用于创建并运行数据迁移作业。
- 作为管理中心组件连接数据湖时的Agent代理。

DataArts Studio实例中不包含CDM集群，如果您需要使用数据集成的功能，请参考用户指南中的“准备工作 > （可选）创建DataArts Studio增量包”章节，创建批量数据迁移集群。

3.3.4.2 解绑/绑定集群的 EIP

操作场景

CDM集群创建完成后，支持解绑或绑定EIP。

- 如果CDM需要访问本地数据源、Internet的数据源，或者跨VPC的云服务，则必须要为CDM集群绑定一个弹性IP，或者使用NAT网关让CDM集群与其他弹性云服务器共享弹性IP访问Internet。
- EIP的异常通知，需要先在IAM控制台创建对应Region的VPC策略委托才能生效。也可以在CDM集群管理界面选择“弹性IP检测授权 > 创建委托”来创建。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

前提条件

- 已创建CDM集群。
- 已拥有EIP配额，才能绑定EIP。

操作步骤

步骤1 登录CDM管理控制台。单击左侧导航上的“集群管理”，进入集群管理界面。

图 3-30 集群列表

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 对相应需要操作的集群可以进行绑定EIP或解绑EIP的操作。

- 绑定EIP：单击集群操作列中的“绑定弹性IP”，进入EIP选择界面。
- 解绑EIP：选择“更多 > 解绑弹性IP”。

步骤3 单击“确定”绑定或解绑EIP。

----结束

3.3.4.3 重启集群

操作场景

在进行某些配置修改（如关闭用户隔离等）后，需要重启集群才能生效。此时您需要进行集群重启操作。

前提条件

已创建CDM集群。

重启集群

步骤1 登录CDM管理控制台。单击左侧导航上的“集群管理”，进入集群管理界面。

图 3-31 集群列表

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 选择集群操作列中的“更多 > 重启”，进入重启集群确认界面。

图 3-32 重启集群

步骤3 您可以选择重启CDM服务进程或重启集群VM，选择完成并点击确认后即可完成集群重启操作。

- 重启CDM服务进程：只重启CDM服务的进程，不会重启集群虚拟机。
- 重启集群VM：业务进程会中断，并重启集群的虚拟机。

----结束

3.3.4.4 删除集群

操作场景

当您确认不再使用当前集群后，可以删除当前CDM集群。

⚠ 注意

删除CDM集群后集群以及数据都销毁且无法恢复，请您谨慎操作！

删除集群前，请您确认如下注意事项：

- 待删除集群确认已不再使用，且其中的连接和作业数据您已通过**批量管理作业**中的导出作业功能进行备份。

前提条件

已创建CDM集群。

删除集群

步骤1 登录CDM管理控制台。单击左侧导航上的“集群管理”，进入集群管理界面。

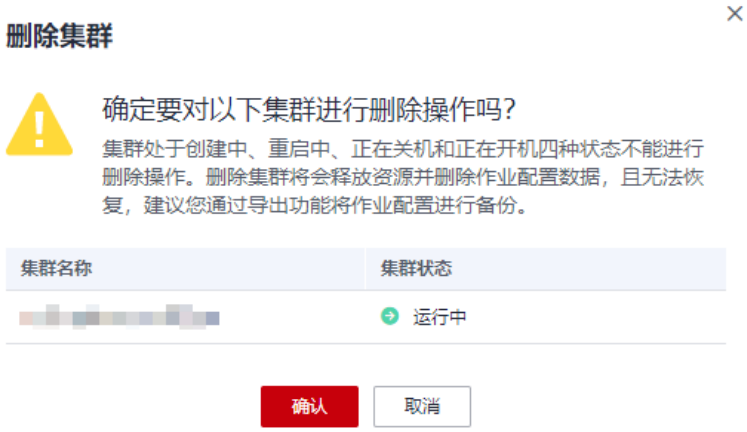
图 3-33 集群列表

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 选择集群操作列中的“更多 > 删除”，进入删除集群确认界面。

图 3-34 删除集群



步骤3 点击“确认”，即开始删除CDM集群。

----结束

3.3.4.5 下载集群日志

操作场景

本章节指导用户获取集群的日志。集群的日志可用于查看作业运行记录，定位作业失败原因等。

前提条件

已创建CDM集群。

操作步骤

步骤1 登录CDM管理控制台。单击左侧导航上的“集群管理”，进入集群管理界面。

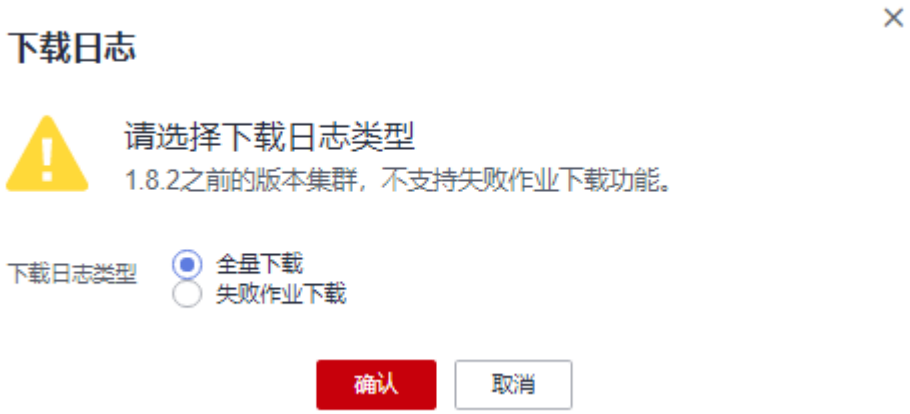
图 3-35 集群列表

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 选择集群操作列中的“更多 > 下载日志”，选择下载日志类型。

图 3-36 下载日志类型



步骤3 确认后，即可下载日志到本地。

----结束

3.3.4.6 查看集群基本信息/修改集群配置

操作场景

CDM集群已经创建成功后，您可以查看集群基本信息，并修改集群的配置。

- 查看集群基本信息：
  - 集群信息：集群版本、创建时间、项目ID、实例ID和集群ID等。
  - 节点配置：集群规格、CPU和内存配置等信息。
  - 网络信息：网络配置。
- 支持修改集群的以下配置：
  - 消息通知：CDM的迁移作业（目前仅支持表/文件迁移的作业）失败时，或者EIP异常时，会发送短信或邮件通知用户。
  - 用户隔离：控制其他用户是否能够操作该集群中的迁移作业、连接。
    - 开启该功能时，该集群中的迁移作业、连接会被隔离，帐号下的其他IAM用户无法操作该集群下的作业、连接。
    - 关闭该功能时，该集群中的迁移作业、连接信息可以用户共享，帐号下的所有拥有相应权限的IAM用户可以查看、操作。

      注意，用户隔离关闭后需要重启集群VM才能生效。
- 管理CDM集群标签：

  支持新增、修改及删除CDM集群的标签。使用标签可以标识多种云资源，后续在TMS标签系统中可筛选出同一标签的云资源。

说明

一个CDM集群最多可新增10个标签。

前提条件

已创建CDM集群。

查看集群基本信息

步骤1 登录CDM管理控制台。单击左侧导航上的“集群管理”，进入集群管理界面。

图 3-37 集群列表

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 单击集群名称，可查看集群的基本信息。

----结束

修改集群配置

步骤1 登录CDM管理控制台。单击左侧导航上的“集群管理”，进入集群管理界面。

图 3-38 集群列表

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 单击集群名称后，选择“集群配置”页签，可修改消息通知、用户是否隔离的配置。

步骤3 修改完成后单击“保存”，返回集群管理界面。

步骤4 如果是关闭用户隔离，需要重启集群VM才能生效，在集群列表处，选择操作列中的“更多 > 重启”。

图 3-39 重启集群

- 重启CDM服务进程：只重启CDM服务的进程，不会重启集群虚拟机。
- 重启集群VM：业务进程会中断，并重启集群的虚拟机。

步骤5 选择“重启集群VM”后单击“确定”。

----结束

3.3.4.7 查看监控指标

3.3.4.7.1 支持的监控指标

前提条件

使用CDM监控功能，需获取CES相关权限。

功能说明

本节定义了数据集成上报云监控的监控指标的命名空间、监控指标列表和维度定义，用户可以通过云监控提供的API接口来检索监控指标。

命名空间

SYS.CDM

监控指标

CDM集群支持的监控指标如表3-20所示。

表 3-20 CDM 支持的监控指标

指标ID	指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
bytes_in	网络流入速率	该指标用于统计每秒流入测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例	1分钟
bytes_out	网络流出速率	该指标用于统计每秒流出测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例	1分钟
cpu_usage	CPU使用率	该指标用于统计测量对象的CPU使用率。 单位：%。	0%～100%	CDM集群实例	1分钟
mem_usage	内存使用率	该指标用于统计测量对象的内存使用率。 单位：%。	0%～100%	CDM集群实例	1分钟
disk_usage	磁盘利用率	该指标为从物理机层面采集的磁盘使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~90%	CDM集群实例	1分钟
disk_io	磁盘io	该指标为从物理机层面采集的磁盘每秒读取和写入的字节数，数据准确性低于从弹性云服务器内部采集的数据。 单位：Byte/sec	0~10GB	CDM集群实例	1分钟
tomcat_heap_usage	堆内存使用率	该指标为从物理机层面采集的堆内存使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~90%	CDM集群实例	1分钟
tomcat_connect	tomcat并发连接数	该指标为从物理机层面采集的tomcat并发连接数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟

指标ID	指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
tomcat_thread_count	tomcat线程数	该指标为从物理机层面采集的tomcat所占线程数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_connect	数据库连接数	该指标为从物理机层面采集的postgres数据库连接数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_submission_row	历史记录表行数	该指标为从物理机层面采集的postgres数据库submission表行数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_failed_job_rate	失败作业率	该指标为从物理机层面sqoop进程采集的失败作业率。 单位：%。	0.001%~100%	CDM集群实例	1分钟
inodes_usage	Inodes利用率	该指标为从物理机层面采集的磁盘inodes使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~0.9%	CDM集群实例	1分钟

维度

Key	Value
instance_id	云数据迁移服务实例

3.3.4.7.2 设置告警规则

操作场景

通过设置CDM集群告警规则，用户可自定义监控目标与通知策略，及时了解CDM集群运行状况，从而起到预警作用。

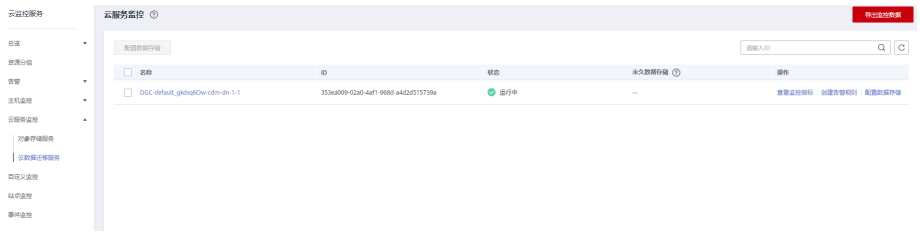
设置CDM集群的告警规则包括设置告警规则名称、监控对象、监控指标、告警阈值、监控周期和是否发送通知等参数。本节介绍了设置CDM集群告警规则的具体方法。

操作步骤

步骤1 进入CDM主界面，选择“集群管理”，选择集群操作列中的“更多 > 查看监控指标”。

步骤2 点击监控指标页面左上角的返回按钮，进入云监控服务的界面，选择“云数据迁移服务”服务监控项对应操作列的“创建告警规则”。

图 3-40 “云数据迁移服务”服务监控项



步骤3 根据界面提示设置CDM集群的告警规则。

步骤4 设置完成后，单击“确定”。当符合规则的告警产生时，系统会自动进行通知。

说明

更多关于监控告警的信息，请参见《云监控用户指南》。

----结束

3.3.4.7.3 查看监控指标

操作场景

您通过云监控服务可以对CDM集群的运行状态进行日常监控。您可以通过云监控管理控制台，直观地查看各项监控指标。

由于监控数据的获取与传输会花费一定时间，因此，监控显示的是当前时间5～10分钟前的状态。如果您的CDM集群刚刚创建完成，请等待5～10分钟后查看监控数据。

前提条件

- CDM集群正常运行。

  重启失败、不可用状态的集群，无法查看其监控指标。当集群再次启动或恢复后，即可正常查看。

- CDM集群已正常运行一段时间（约10分钟）。

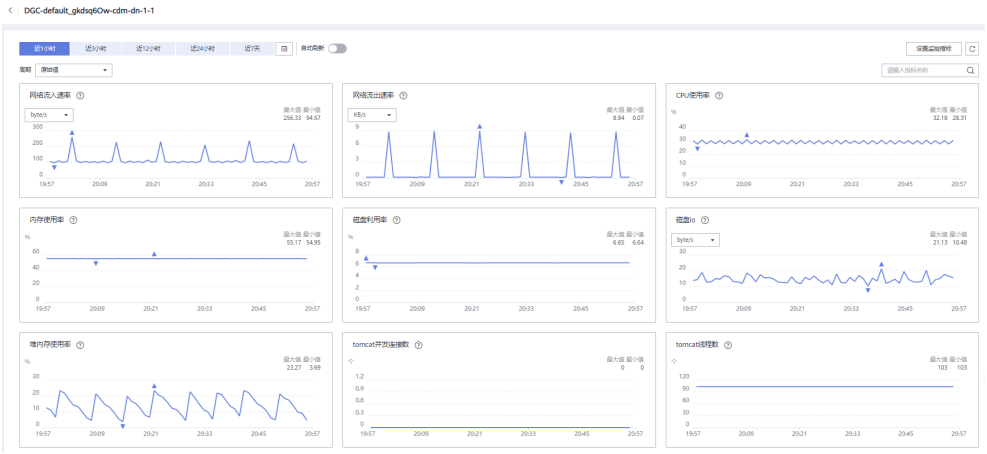
  对于新创建的集群，需要等待一段时间，才能查看上报的监控数据和监控视图。

操作步骤

步骤1 进入CDM主界面，选择“集群管理”，选择集群操作列中的“更多 > 查看监控指标”。

步骤2 在CDM监控页面，可查看所有监控指标的小图。

图 3-41 查看监控指标



步骤3 单击小图右上角的，可进入大图模式查看。

步骤4 您可以在左上角选择时长作为监控周期，查看一段时间的指标变化情况。

----结束

3.3.5 管理连接

3.3.5.1 新建连接

操作场景

用户在创建数据迁移的任务前，需要先创建连接，让CDM集群能够读写数据源。一个迁移任务，需要建立两个连接，源连接和目的连接。不同的迁移方式（表或者文件迁移），哪些数据源支持导出（即作为源连接），哪些数据源支持导入（即作为目的连接），详情请参见**支持的数据源**。

不同类型的数据源，创建连接时的配置参数也不相同，本章节指导用户根据数据源类型创建对应的连接。

约束限制

当所连接的数据源发生变化（如MRS集群扩容等情况）时，您需要重新编辑并保存该连接。

前提条件

- 已具备CDM集群。
- CDM集群与目标数据源可以正常通信。
  - 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。
  - 如果目标数据源为云上服务（如DWS、MRS及ECS等），则网络互通需满足如下条件：

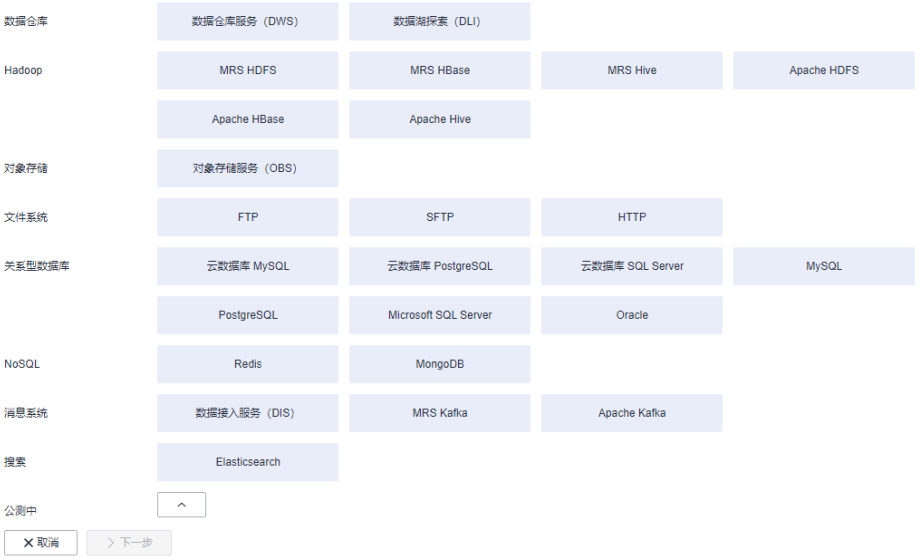
  - CDM集群与云上服务处于不同区域的情况下，需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP，数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
  - CDM集群与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC) 使用指南》中的“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC) 使用指南》中的“安全组 > 添加安全组规则”章节。
  - 此外，您还必须确保该云服务的实例与CDM集群所属的企业项目必须相同，如果不同，需要修改工作空间的企业项目。
- 已获取待连接数据源的地址、用户名和密码，且该用户拥有数据导入、导出的操作权限。
- 使用Agent时需用主账户给子账户赋予CDM操作权限。

新建连接

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择CDM集群后的“作业管理 > 连接管理 > 新建连接”。选择连接器类型。

这里的连接器类型，是根据待连接的数据源类型分类的，包含了CDM目前支持导入/导出的所有数据源类型。

图 3-42 选择连接器类型



步骤2 选择数据源类型后，单击“下一步”配置连接参数，这里以创建MySQL连接为例。

每种数据源的连接参数不同，您可以根据所选择的连接器类型在表3-21中查找对应参数。

表 3-21 连接参数分类

连接器类型	参数说明
• 数据仓库服务（DWS） • 云数据库 MySQL • 云数据库 PostgreSQL • 云数据库 SQL Server • PostgreSQL • Microsoft SQL Server • SAP HANA	由于连接这些关系型数据库，所采用的JDBC驱动相同，所以他们的连接参数也一样，具体参数请参见**配置常见关系数据库连接**。
MySQL	连接MySQL数据库时，具体参数请参见**配置MySQL数据库连接**。
Oracle	连接Oracle数据库时，具体参数请参见**配置Oracle数据库连接**。
分库	连接达梦数据库时，具体参数请参见**配置分库连接**。
对象存储服务（OBS）	连接OBS时，具体参数请参见**配置OBS连接**。
• MRS HDFS • FusionInsight HDFS • Apache HDFS	连接MRS、Apache Hadoop或FusionInsight HD上的HDFS时，具体参数请参见**配置HDFS连接**。
• MRS HBase • FusionInsight HBase • Apache HBase	连接MRS、Apache Hadoop或FusionInsight HD上的HBase时，具体参数请参见**配置HBase连接**。
• MRS Hive • FusionInsight Hive • Apache Hive	连接MRS、Apache Hadoop或FusionInsight HD上的Hive时，具体参数请参见**配置Hive连接**。
表格存储服务（CloudTable）	连接CloudTable时，具体参数请参见**配置CloudTable连接**。
• FTP • SFTP	连接FTP或SFTP服务器时，具体参数请参见**配置FTP/SFTP连接**。
HTTP	用于读取一个公网HTTP/HTTPS URL的文件，包括第三方对象存储的公共读取场景和网盘场景。 当前创建HTTP连接时，只需要配置连接名称，具体URL在创建作业时配置。
MongoDB	连接本地MongoDB数据库时，具体参数请参见**配置MongoDB连接**。
文档数据库服务（DDS）	连接DDS时，具体参数请参见**配置DDS连接**。

连接器类型	参数说明
• Redis • 分布式缓存服务（DCS）	连接Redis或DCS时，具体参数请参见**配置Redis/DCS连接**。
• MRS Kafka • Apache Kafka	连接MRS Kafka或Apache Kafka数据源时，具体参数请参见**配置Kafka连接**。
云搜索服务 Elasticsearch	连接云搜索服务或Elasticsearch时，具体参数请参见**配置Elasticsearch/云搜索服务（CSS）连接**。
数据湖探索（DLI）	连接数据湖探索服务时，具体参数请参见**配置DLI连接**。
DMS Kafka	连接DMS的Kafka队列时，具体参数请参见**配置DMS Kafka连接**。
Cassandra	连接Cassandra时，具体参数请参见**配置Cassandra连接**。

说明

目前以下数据源处于公测阶段：FunsionInsight HDFS、FunsionInsight HBase、FunsionInsight Hive、SAP HANA、文档数据库服务（DDS）、表格存储服务（CloudTable）、Cassandra、DMS Kafka、云搜索服务、分库。

步骤3 连接的参数配置完成后单击“测试”，可测试连接是否可用。或者直接单击“保存”，保存时也会先检查连接是否可用。

受网络和数据源的影响，部分连接测试的时间可能需要30～60秒。

----结束

管理连接

CDM支持对已创建的连接进行以下操作：

- 删除：支持删除未被任何作业使用的连接，也支持批量删除连接。
- 编辑：支持修改已创建好的连接参数，但不支持重新选择连接器。修改连接时，需要重新输入数据源的登录密码。
- 测试连通性：支持直接测试已保存连接的连通性。
- 查看连接JSON：以JSON文件格式查看连接参数的配置。
- 编辑连接JSON：以直接修改JSON文件的方式，修改连接参数。
- 查看后端连接：查看该连接对应的后端连接。例如已开启后端连接的MYCAT连接，就可以查询到对应的后端连接详情。

在管理连接前，您需要确保该连接未被任何作业使用，避免影响现有作业业务。管理连接的操作流程如下：

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择CDM集群后的“作业管理 > 连接管理”。

步骤2 在连接管理界面找到需要修改的连接：

- 删除连接：单击操作列的“删除”删除该连接，或者勾选连接后单击列表上方的“删除连接”来批量删除未被任何作业使用的连接。
- 编辑连接：单击该连接名称，或者单击操作列的“编辑”进入修改连接的界面，修改连接时需要重新输入数据源的登录密码。
- 测试连通性：单击操作列的“测试连通性”，直接测试已保存连接的连通性。
- 查看连接JSON：选择操作列的“更多 > 查看连接JSON”，以JSON文件格式查看连接参数的配置。
- 编辑连接JSON：选择操作列的“更多 > 编辑连接JSON”，以直接修改JSON文件的方式，修改连接参数。
- 查看后端连接：选择操作列的“更多 > 查看后端连接”，查看该连接对应的后端连接。

----结束

3.3.5.2 管理驱动

JDBC即Java DataBase Connectivity，java数据库连接；JDBC提供的API可以让JAVA通过API方式访问关系型数据库，执行SQL语句，获取数据。

CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。

前提条件

- 已创建集群。
- 已参见**表3-22**下载对应的驱动。
- 已参见**配置FTP/SFTP连接**创建SFTP连接并将对应的驱动上传至线下文件服务器（可选）。

如何获取驱动

不同类型的关系数据库，需要适配不同类型的驱动。注意，上传的驱动版本不必与待连接的数据库版本相匹配，直接参考**表3-22**获取建议版本的JDK8 .jar格式驱动即可。

表 3-22 获取驱动

关系数据库类型	驱动名称	获取地址	建议版本
• 云数据库MySQL • MySQL	MYSQL MYCAT	**https://downloads.mysql.com/archives/c-j/**	5.1.48，获取mysql-connector-java-5.1.48.jar

关系数据库类型	驱动名称	获取地址	建议版本
Oracle	ORACLE_6 ORACLE_7 ORACLE_8	驱动包下载地址：**https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html** 历史版本驱动包下载地址：**https://repo1.maven.org/maven2/com/oracle/database/jdbc/ojdbc8/12.2.0.1/**	ojdbc8的12.2.0.1版本，获取ojdbc8.jar **说明** 不支持使用新版本（如Oracle Database 21c (21.3) drivers），会导致创建作业时无法获取模式名。
• 云数据库PostgreSQL • PostgreSQL	POSTGRESQL	**https://jdbc.postgresql.org/download**	42.1.4的JDBC 4.2版本，获取postgresql-42.1.4.jar
• 云数据库SQL Server • Microsoft SQL Server	SQLServer	驱动包下载地址： **https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15** 历史版本驱动包下载地址： **https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases**	4.2，获取sqljdbc42.jar

操作步骤

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择CDM集群后的“作业管理 > 连接管理 > 驱动管理”，进入驱动管理页面。

图 3-43 上传驱动

步骤2 方式一：单击对应驱动名称右侧操作列的“上传”，选择本地已下载的驱动。

方式二：单击对应驱动名称右侧操作列的“从sftp复制”，配置sftp连接器名称和驱动文件路径。

步骤3 （可选）在驱动更新场景下，上传驱动后必须在CDM集群列表中重启集群才能更新生效。

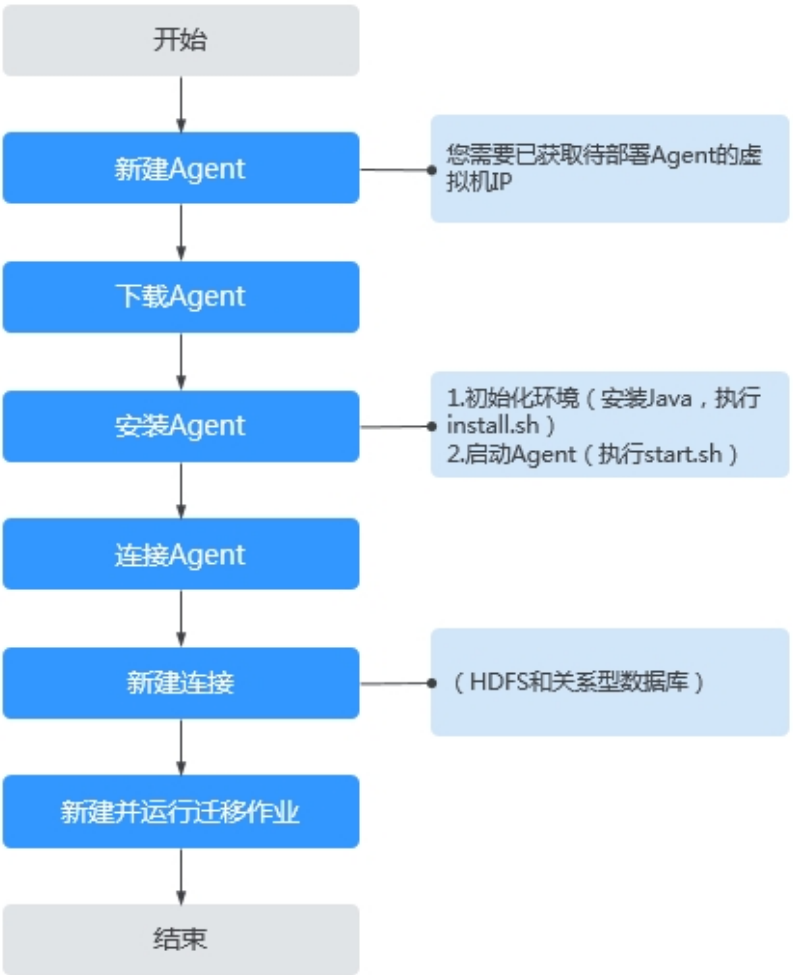
----结束

3.3.5.3 管理 Agent

对于HDFS和关系型数据库类型的数据源，不方便暴露节点的场景，可选择在源端网络中部署Agent。CDM通过Agent拉取客户内部数据源的数据，但不支持写入数据。

Agent的使用流程如图3-44所示。

图 3-44 Agent 使用流程



前提条件

已具备CDM集群。

新建 Agent

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择集群后的“作业管理 > Agent管理 > 新建Agent”，配置Agent相关信息。

图 3-45 配置 Agent

- IP地址：配置为源端网络中部署Agent的IP地址。
- 端口：Agent自定义的端口。建议范围：1024~65535。
- 启用压缩：是否对数据使用gzip算法进行压缩传输。
  - 对于文本数据（基于字符编码的数据，例如MySQL的INT等数据类型，详见相关数据库的说明文档），建议开启此选项，gzip压缩可以达到较好的压缩效果。
  - 对于二进制数据（基于值编码的数据，例如MySQL的BINARY等数据类型，详见相关数据库的说明文档），由于其本身已经压缩过，不推荐再开启gzip压缩，压缩后可能会导致压缩效果较差，同时会增大客户端解压缩的压力，带来不必要的性能损耗。
- 启用SSL：是否启用SSL双向认证，保证数据的安全性。如果对安全性要求较高，则可以开启SSL。
- 限流：设置agent的最大下行速率，默认不限流。

步骤2 单击“确定”，完成Agent的创建。在Agent管理页面可查看已成功创建的Agent。

----结束

安装并启动 Agent

步骤1 在Agent管理页面，找到已成功创建的Agent。如图3-46所示，下载Agent。

图 3-46 下载 Agent

步骤2 准备部署Agent的主机。该主机对vCPUs、内存、磁盘等规格无特殊要求，但须满足以下条件：

- 需要已安装64位版本java 8并配置java环境变量。
- 授予Ruby用户（若无Ruby用户则需手动创建）在/tmp目录下的写权限。

步骤3 将下载的Agent压缩包，上传至部署Agent的主机上。

步骤4 解压安装包后执行如下命令安装Agent。

sh sbin/install.sh

步骤5 如果需要通过Agent连接关系数据库，则需要将对应的驱动（参考**管理驱动**获取）上传至Agent安装目录下的/server/jdbc，并修改同目录下properties文件里对应数据库驱动的版本号。

步骤6 安装完成后，执行如下命令启动Agent。

su Ruby

sh sbin/start.sh

步骤7 执行如下命令检查Agent进程是否启动。

ps -ef | grep agent

如果命令执行完成后返回了正在运行的Agent进程，说明Agent进程已启动。

----结束

连接 Agent

步骤1 在Agent管理页面，找到已成功创建的Agent。如图3-47所示，连接Agent。

图 3-47 连接 Agent

步骤2 Agent连接成功后，即可在创建连接中选择Agent。

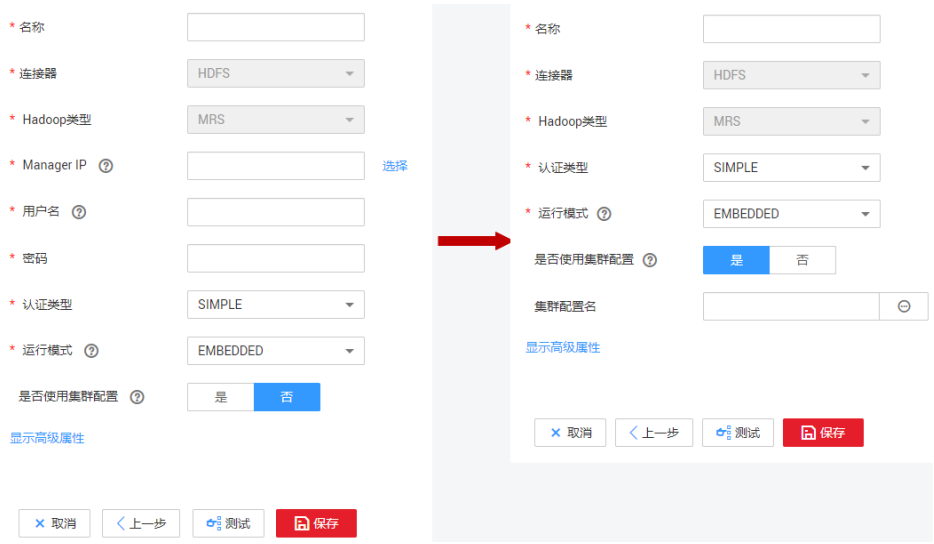
----结束

3.3.5.4 管理集群配置

集群配置管理支持新建、编辑或删除Hadoop集群配置。

Hadoop集群配置主要用于新建Hadoop类型连接时，能够简化复杂的连接参数配置，如图3-48所示。

图 3-48 使用集群配置前后对比



CDM支持的Hadoop类型连接主要包括以下几类：

- MRS集群：MRS HDFS，MRS HBase，MRS Hive。
- FusionInsight集群：FusionInsight HDFS，FusionInsight HBase，FusionInsight Hive。
- Apache集群：Apache HDFS，Apache HBase，Apache Hive。

操作场景

当需要新建Hadoop类型连接时，建议先创建集群配置，以简化复杂的连接参数配置。

前提条件

- 已创建集群。
- 已参见表1获取相应Hadoop集群配置文件和Keytab文件。

获取集群配置文件和 Keytab 文件

不同Hadoop类型的集群配置文件和Keytab文件获取方式有所不同，请参见表1获取相应Hadoop集群配置文件和Keytab文件。

表 3-23 集群配置文件和 Keytab 文件获取方式

Hadoop类型连接	集群配置文件获取方式	Keytab文件获取方式
MRS集群 ● MRS HDFS ● MRS HBase ● MRS Hive	针对MRS 3.x版本集群： 1. 登录FusionInsight Manager。 2. 选择“集群 > > 待操作的集群名称 > 概览 > > 更多 > > 下载客户端”，界面显示“下载集群客户端”对话框。 3. 对话框中选择“仅配置文件”，平台类型和服务端保持一致，单击确认后进行本地下载。 4. 获取下载的tar包，此即为FusionInsight集群配置文件。 针对MRS 2.x及之前版本集群： 1. 登录MRS管理控制台。 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，单击“组件管理”。 3. 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”或“远端主机”，自定义文件保存路径后，单击“确定”开始生成客户端配置文件。 4. 将生成的配置文件，保存到本地路径。 具体可参见MapReduce服务文档。	针对MRS 3.x版本集群： 1. 登录FusionInsight Manager。 2. 通过“系统 > > 权限 > 用户”，选择所需用户所在行，点击“更多 > >下载认证凭据”下载认证凭据文件。 3. 获取下载的tar包，此即为FusionInsight集群Keytab文件。 针对MRS 2.x及之前版本集群： 1. 登录MRS服务的Manager，单击“系统设置”。在“权限配置”区域，单击“用户管理”。 2. 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。 具体可参见MapReduce服务文档。

Hadoop类型连接	集群配置文件获取方式	Keytab文件获取方式
FusionInsight集群 ● FusionInsight HDFS ● FusionInsight HBase ● FusionInsight Hive	1. 登录FusionInsight Manager。 2. 选择“集群 > > 待操作的集群名称 > 概览 > > 更多 > > 下载客户端”，界面显示“下载集群客户端”对话框。 3. 对话框中选择“仅配置文件”，平台类型和服务端保持一致，单击确认后进行本地下载。 4. 获取下载的tar包，此即为FusionInsight集群配置文件。 具体可参见FusionInsight文档。	1. 登录FusionInsight Manager。 2. 通过“系统 > > 权限 > 用户”，选择所需用户所在行，点击“更多 > >下载认证凭据”下载认证凭据文件。 3. 获取下载的tar包，此即为FusionInsight集群Keytab文件。 具体可参见FusionInsight文档。

Hadoop类型连接	集群配置文件获取方式	Keytab文件获取方式
Apache集群 ● Apache HDFS ● Apache HBase ● Apache Hive	Apache集群场景下，此处仅说明需要哪些配置文件与打包原则，各配置文件的具体获取方式请参见对应版本说明文档。 ● HDFS需要将以下文件压缩为无目录格式的zip包： – hosts – core-site.xml – hdfs-site.xml – yarm-site.xml – mapred-site.xml – krb5.conf（可选，安全模式集群使用） ● HBase需要将以下文件压缩为无目录格式的zip包： – hosts – core-site.xml – hdfs-site.xml – yarm-site.xml – mapred-site.xml – hbase-site.xml – krb5.conf（可选，安全模式集群使用） ● Hive需要将以下文件压缩为无目录格式的zip包： – hosts – core-site.xml – hdfs-site.xml – yarm-site.xml – mapred-site.xml – hive-site.xml – hivemetastore-site.xml – krb5.conf（可选，安全模式集群使用）	Apache集群场景下，此处仅说明认证凭据文件打包原则，认证凭据文件具体获取方式请参见对应版本说明文档。 1. 将用户的认证凭据文件重命名为user.keytab。 2. 将user.keytab文件压缩为无目录格式的zip包：user.keytab.zip。

说明

- 集群配置文件包含集群的配置参数。如果修改了集群的配置参数，需重新获取获取配置文件。
- Keytab文件为认证凭据文件。获取Keytab文件前，需要在集群上至少修改过一次此用户的密码，否则下载获取的keytab文件可能无法使用。另外，修改用户密码后，之前导出的keytab将失效，需要重新导出。
- Keytab文件在仅安全模式集群下使用，普通模式集群下无需准备Keytab文件。

操作步骤

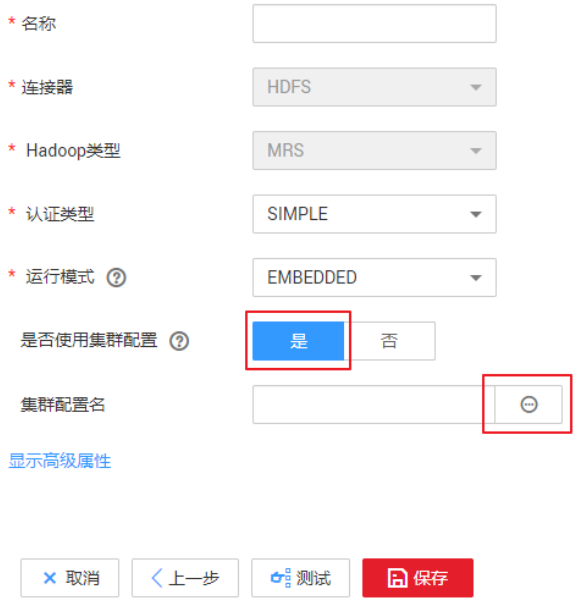
1. 进入CDM主界面，进入集群管理界面。选择CDM集群后的“作业管理 > 连接管理 > 集群配置管理”。
2. 在集群配置管理界面，选择“新建集群配置”，配置参数填写如下：

   图 3-49 新建集群配置

   - 集群配置名：根据连接的数据源类型，用户可自定义便于记忆、区分的集群配置名。
   - 上传集群配置：单击“添加文件”以选择本地的集群配置文件，然后通过操作框右侧的“上传文件”进行上传。
   - Principal：**仅安全模式集群需要填写该参数**。Principal即Kerberos安全模式下的用户名，需要与Keytab文件保持一致。
   - 上传Keytab文件：**仅安全模式集群需要上传该文件**。单击“添加文件”以选择本地的Keytab文件，然后通过操作框右侧的“上传文件”进行上传。
   - 描述：用户可添加对此集群配置的描述，用于标识和区分该集群配置。
3. 确认后集群配置新建成功。后续在新建Hadoop类型连接时，认证模式根据实际情况选择，将“是否使用集群配置”选择为“是”，然后选择对应的“集群配置名”， 即可快速完成Hadoop类型连接创建。

图 3-50 使用集群配置



3.3.5.5 配置常见关系数据库连接

常见关系数据库包括数据仓库服务（DWS）、云数据库 MySQL、云数据库 PostgreSQL、云数据库 SQLServer、PostgreSQL、Microsoft SQL Server、IBM Db2、SAP HANA。

前提条件

已参考**管理驱动**上传对应的驱动。

常见关系数据库连接参数

连接参数如**表3-24**所示。

表 3-24 常见关系数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mysql_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的DWS、RDS等实例列表。	192.168.0.1

参数名	说明	取值样例
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。例如： SQLServer默认端口：1433 PostgreSQL默认端口：5432
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
一次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
SSL加密	可选参数，支持通过SSL加密方式连接数据库，暂不支持自建的数据库。 RDS上的PostgreSQL数据库服务做了一些安全增强，在创建RDS上的PostgreSQL的连接时，该参数需要配置为“是”。	是
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 **说明** CDM作业默认打开了useCursorFetch开关，即JDBC连接器与关系型数据库的通信使用二进制协议。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

3.3.5.6 配置分库连接

分库指的是同时连接多个后端数据源，该连接可作为作业源端，将多个数据源的数据合一迁移到其他数据源上。连接参数如**表3-25**所示。

表 3-25 分库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	my_link
用户名	待连接数据库的用户。 仅当“数据源列表”中某个后端数据库A未配置用户名密码时，该配置对A生效。如果后端数据库B已配置用户名密码，此处配置不对B生效。	cdm
密码	待连接数据库的用户密码。 仅当“数据源列表”中某个后端数据库A未配置用户名密码时，该配置对A生效。如果后端数据库B已配置用户名密码，此处配置不对B生效。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
后端数据源	输入后端数据库的类型，当前仅支持MYSQL。	MYSQL
数据源列表	输入后端数据库的IP、端口、数据库名称、账户名、密码，以“:”隔开。即ip:port:dbs:username:password，其中username:password可以不填，此时以“用户名”、“密码”配置为准。 如果此处有多个后端数据库，需要确保表结构一致，并使用“|”分隔数据源。如果密码包含“|”或者“:”，可使用“\”转义。 例如“192.168.2.1:3306:cdm|192.168.2.2:3306:cdm:user:password”表示，第一个后端数据库IP为192.168.2.1，端口为3306，数据库名称为cdm，账户名密码以“用户名”、“密码”处配置为准；第二个后端数据库IP为192.168.2.2，端口为3306，数据库名称为cdm，账户名为“user”、密码为“password”。	192.168.2.1:3306:cdm|192.168.2.2:3306:cdm:user:password
一次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

3.3.5.7 配置 MYCAT 连接

MYCAT是一个开源的分布式数据库系统，其核心功能是分表分库，即将一个大表水平分割为多个小表，存储在后端MySQL或者其他数据库里。连接参数如表3-26所示。

表 3-26 MYCAT 数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mycat_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的DWS、RDS实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	3306
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户密码。	-
使用本地API	可选参数，选择是否使用数据库本地API加速。 创建连接时，CDM会自动尝试启用MySQL数据库的local_infile系统变量，开启MySQL的LOAD DATA功能加快数据导入，提高导入数据到MySQL数据库的性能。 如果CDM自动启用失败，请联系数据库管理员启用local_infile参数或选择不使用本地API加速。	是
后端连接	选择是否使用后端连接。	是
管理帐号	输入MYCAT管理帐号。	root
管理密码	输入MYCAT管理密码。	123456
管理端口	输入MYCAT管理端口。	9066
后端数据源	输入MYCAT后端数据库的类型。	MYSQL
后端用户名	输入MYCAT后端数据库的用户名。	cdm
后端密码	输入MYCAT后端数据库的密码。	-
一次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

3.3.5.8 配置达梦（DM）数据库连接

连接达梦（DM）数据库时，相关参数如表3-27所示。

表 3-27 达梦（DM）数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dm_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的DWS、RDS等实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
Agent	单击“选择”，选择管理Agent中已创建的Agent。	-
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
一次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
SSL加密	可选参数，支持通过SSL加密方式连接数据库，暂不支持自建的数据库。 RDS上的PostgreSQL数据库服务做了一些安全增强，在创建RDS上的PostgreSQL的连接时，该参数需要配置为“是”。	是
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 **说明** CDM作业默认打开了useCursorFetch开关，即JDBC连接器与关系型数据库的通信使用二进制协议。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

3.3.5.9 配置 MySQL 数据库连接

连接MySQL数据库连接时，相关参数如表3-28所示。

表 3-28 MySQL 数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mysql_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的MySQL数据库实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	3306
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用本地API	可选参数，选择是否使用数据库本地API加速。 创建MySQL连接时，CDM会自动尝试启用MySQL数据库的local_infile系统变量，开启MySQL的LOAD DATA功能加快数据导入，提高导入数据到MySQL数据库的性能。 如果CDM自动启用失败，请联系数据库管理员启用local_infile参数或选择不使用本地API加速。 如果是导入到RDS上的MySQL数据库，由于RDS上的MySQL默认没有开启LOAD DATA功能，所以同时需要修改MySQL实例的参数组，将“local_infile”设置为“ON”，开启该功能。 **说明** 如果RDS上的“local_infile”参数组不可编辑，则说明是默认参数组，需要先创建一个新的参数组，再修改该参数值，并应用到RDS的MySQL实例上，具体操作请参见《关系型数据库用户指南》。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择管理Agent中已创建的Agent。	-
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000

参数名	说明	取值样例
单次提交行数	可选参数，单击“显示高级属性”后显示。 指定每次批量提交的行数，根据数据目的端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	-
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 **说明** CDM作业默认打开了useCursorFetch开关，即JDBC连接器与关系型数据库的通信使用二进制协议。 开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

3.3.5.10 配置 Oracle 数据库连接

连接Oracle数据库时，连接参数如表3-29所示。

表 3-29 Oracle 数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	oracle_link
数据库服务器	配置为要连接的数据库的IP地址或域名。	192.168.0.1
端口	配置为要连接的数据库的端口。	默认端口：1521
数据库连接类型	选择Oracle数据库连接类型： ● Service Name：通过SERVICE_NAME连接Oracle数据库。 ● SID：通过SID连接Oracle数据库。	SID
实例名称	配置Oracle实例ID，用于实例区分各个数据库。“数据库连接类型”选择“SID”时才有该参数。	dbname
数据库名称	配置为要连接的数据库名称。“数据库连接类型”选择“Service Name”时才有该参数。	dbname

参数名	说明	取值样例
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
Oracle版本	创建Oracle连接时才有该参数，根据您Oracle数据库的版本来选择。当出现“java.sql.SQLException: Protocol violation异常”时，可以尝试更换版本号。	高于12.1
一次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。 Oracle到DWS迁移时，可能出现目的端写太久导致迁移超时的情况。此时请减少Oracle源端“一次请求行数”参数值的设置。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 **说明** CDM作业默认打开了useCursorFetch开关，即JDBC连接器与关系型数据库的通信使用二进制协议。 • 开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

3.3.5.11 配置 DLI 连接

连接数据湖探索（DLI）服务时，相关参数如**表3-30**所示。

表 3-30 DLI 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dli_link
访问标识(AK)	访问DLI数据库时鉴权所需的AK和SK。 您需要先创建当前帐号的访问密钥，并获得对应的AK和SK。	-
密钥(SK)		-

参数名	说明	取值样例
项目ID	DLI服务所在区域的项目ID。 项目ID表示租户的资源，帐号ID对应当前帐号。用户可在对应页面下查看不同Region对应的项目ID和帐号ID。 1. 注册并登录管理控制台。 2. 在用户名的下拉列表中单击“我的凭证”。 3. 在“我的凭证”页面，查看帐号名和帐号ID，在项目列表中查看项目ID。	-

3.3.5.12 配置 Hive 连接

目前CDM支持连接的Hive数据源有以下几种：

- **MRS Hive**
- **FusionInsight Hive**
- **Apache Hive**

MRS Hive

用户具有MRS Hive连接的表的访问权限时，才能在字段映射时看到表。

MRS Hive连接适用于云上的MapReduce服务。MRS Hive的连接参数如**表3-31**所示。

说明

- 新建MRS连接前，需在MRS中添加一个kerberos认证用户并登录MRS管理页面更新其初始密码，然后使用该新建用户创建MRS连接。
- 如需连接MRS 2.x版本的集群，请先创建2.x版本的CDM集群。CDM 1.8.x版本的集群无法连接MRS 2.x版本的集群。
- 由于当前CDM Hive连接是从MRS HDFS组件获取core-site.xml配置信息，所以在MRS侧使用的是Hive over OBS场景时，在创建Hive连接前，需要用户在MRS管理界面的HDFS组件中配置OBS的AK、SK信息。
- 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件：
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。
  - 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。

表 3-31 MRS Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 ● 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。	cdm
密码	访问MRS Manager的用户密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： • EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 • STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
检查Hive JDBC连通性	是否需要测试Hive JDBC连通。	否
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hive_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

FusionInsight Hive

FusionInsight Hive连接适用于用户在本地数据中心自建的FusionInsight HD，需通过专线连接。

FusionInsight Hive的连接参数如**表3-32**所示。

表 3-32 FusionInsight Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
Manager IP	FusionInsight Manager平台的地址。	127.0.0.1
Manager端口	FusionInsight Manager平台的端口。	28443
CAS Server端口	与FusionInsight对接的CAS Server的端口。	20009

参数名	说明	取值样例
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。	HIVE_3_X
用户名	登录FusionInsight Manager平台的用户名。	cdm
密码	FusionInsight Manager平台的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hive_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

Apache Hive

Apache Hive连接适用于用户在本地数据中心或ECS上自建的第三方Hadoop，其中本地数据中心的Hadoop需通过专线连接。

Apache Hive的连接参数如**表3-33**所示。

表 3-33 Apache Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
URI	NameNode URI地址。	hdfs://hacluster
Hive元数据地址	设置Hive元数据地址，参考 hive.metastore.uris配置项。例如：thrift://host-192-168-1-212:9083	-
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。	HIVE_3_X
IP与主机名映射	如果Hadoop配置文件使用主机名，需要配置IP与主机的映射。格式：IP与主机名之间使用空格分隔，多对映射使用分号或回车换行分隔。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
Principal	认证类型为“KERBEROS”时，需要填写Principal。Principal即Kerberos安全模式下的用户名，可以联系Hadoop管理员获取。此处填写的Principal需要与Keytab文件保持一致。	-
Keytab文件	认证类型为“KERBEROS”时，需要上传Keytab文件。Keytab文件为认证凭据文件，可以联系Hadoop管理员获取。获取Keytab文件前，需要在集群上至少修改过一次此用户的密码，否则下载获取的keytab文件可能无法使用。另外，修改用户密码后，之前导出的keytab将失效，需要重新导出。	-
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否

参数名	说明	取值样例
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hive_01
Hive JDBC连接串	连接Hive JDBC的url，默认使用匿名用户连接。	-

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

3.3.5.13 配置 HBase 连接

目前CDM支持连接的HBase数据源有以下几种：

- **MRS HBase**
- **FusionInsight HBase**
- **Apache HBase**

MRS HBase

连接MRS上的HBase数据源时，相关参数如**表3-34**所示。

说明

- 新建MRS连接前，需在MRS中添加一个kerberos认证用户并登录MRS管理页面更新其初始密码，然后使用该新建用户创建MRS连接。
- 如需连接MRS 2.x版本的集群，请先创建2.x版本的CDM集群。CDM 1.8.x版本的集群无法连接MRS 2.x版本的集群。
- 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件：
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。
  - 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。

表 3-34 MRS 上的 HBase 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mrs_hbase_link

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 ● 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。	cdm
密码	访问MRS Manager的用户密码。	-
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
HBase版本	HBase版本。	HBASE_2_X

参数名	说明	取值样例
运行模式	“HBASE_2_X”版本支持该参数。选择HBase连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	STANDALONE
是否使用集群配置	用户可以在“连接管理”处创建集群配置，用于简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hbase_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

FusionInsight HBase

连接FusionInsight HD上的HBase数据源时，相关参数如**表3-35**所示。

表 3-35 FusionInsight HBase 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	FI_hbase_link
Manager IP	FusionInsight Manager平台的地址。	127.0.0.1
Manager端口	FusionInsight Manager平台的端口。	28443
CAS Server端口	与FusionInsight对接的CAS Server的端口。	20009
用户名	登录FusionInsight Manager平台的用户名。	cdm
密码	FusionInsight Manager平台的密码。	-

参数名	说明	取值样例
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
HBase版本	HBase版本。	HBASE_2_X
运行模式	“HBASE_2_X”版本支持该参数。选择HBase连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	STANDALONE
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hbase_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

Apache HBase

连接Apache Hadoop上的HBase数据源时，相关参数如**表3-36**所示。

表 3-36 Apache HBase 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hadoop_hbase_link

参数名	说明	取值样例
ZK链接地址	HBase的Zookeeper链接地址。 格式： <host1>:<port>,<host2>:<port>,<host3>:<port>	zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
Principal	认证类型为“KERBEROS”时，需要填写Principal。Principal即Kerberos安全模式下的用户名，可以联系Hadoop管理员获取。此处填写的Principal需要与Keytab文件保持一致。	-
Keytab文件	认证类型为“KERBEROS”时，需要上传Keytab文件。Keytab文件为认证凭据文件，可以联系Hadoop管理员获取。获取Keytab文件前，需要在集群上至少修改过一次此用户的密码，否则下载获取的keytab文件可能无法使用。另外，修改用户密码后，之前导出的keytab将失效，需要重新导出。	-
IP与主机名映射	如果配置文件使用主机名，需要配置IP与主机的映射。格式：IP与主机名之间使用空格分隔，多对映射使用分号或回车换行分隔。	10.3.6.9 hostname01 10.4.7.9 hostname02
HBase版本	HBase版本。	HBASE_2_X
运行模式	“HBASE_2_X”版本支持该参数。选择HBase连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	STANDALONE
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否

参数名	说明	取值样例
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hbase_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

3.3.5.14 配置 HDFS 连接

目前CDM支持连接的HDFS数据源有以下几种：

- **MRS HDFS**
- **FusionInsight HDFS**
- **Apache HDFS**

MRS HDFS

连接MRS上的HDFS数据源时，相关参数如**表3-37**所示。

说明

- 新建MRS连接前，需在MRS中添加一个kerberos认证用户并登录MRS管理页面更新其初始密码，然后使用该新建用户创建MRS连接。
- 如需连接MRS 2.x版本的集群，请先创建2.x版本的CDM集群。CDM 1.8.x版本的集群无法连接MRS 2.x版本的集群。
- 需确保MRS集群和DataArts Studio实例之间网络互通，网络互通需满足如下条件：
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，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，MRS集群可以访问公网且防火墙规则已开放连接端口。
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC)使用指南》中的“自定义路由（Region Type Ⅰ）> 添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。
  - 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。

表 3-37 MRS 上的 HDFS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mrs_hdfs_link

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 ● 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。	cdm
密码	访问MRS Manager的用户密码。	-
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE

参数名	说明	取值样例
运行模式	选择HDFS连接的运行模式： • EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 • STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。 • Agent：连接实例运行在Agent上。 若不使用AGENT运行模式，且在一个CDM中同时连接两个及以上开启Kerberos认证且realm相同的集群，只能使用EMBEDDED运行模式连接其中一个集群，其余需使用STANDALONE。	STANDALONE
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。运行模式选择Agent时显示此参数。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

FusionInsight HDFS

连接FusionInsight HD上的HDFS数据源时，相关参数如**表3-38**所示。

表 3-38 FusionInsight HDFS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	FI_hdfs_link
Manager IP	FusionInsight Manager平台的地址。	127.0.0.1

参数名	说明	取值样例
Manager端口	FusionInsight Manager平台的端口。	28443
CAS Server端口	与FusionInsight对接的CAS Server的端口。	20009
用户名	登录FusionInsight Manager平台的用户名。 从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。	cdm
密码	FusionInsight Manager平台的密码。	-
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
运行模式	选择HDFS连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。 ● Agent：连接实例运行在Agent上。	STANDALONE
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。运行模式选择Agent时显示此参数。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

Apache HDFS

连接Apache Hadoop上的HDFS数据源时，相关参数如表3-39所示。

表 3-39 Apache HDFS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hadoop_hdfs_link
URI	表示NameNode URI地址。可以填写为：hdfs://***namenode实例的ip***:8020。	hdfs://***IP***:8020
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
Principal	认证类型为“KERBEROS”时，需要填写Principal。Principal即Kerberos安全模式下的用户名，可以联系Hadoop管理员获取。此处填写的Principal需要与Keytab文件保持一致。	-
Keytab文件	认证类型为“KERBEROS”时，需要上传Keytab文件。Keytab文件为认证凭据文件，可以联系Hadoop管理员获取。获取Keytab文件前，需要在集群上至少修改过一次此用户的密码，否则下载获取的keytab文件可能无法使用。另外，修改用户密码后，之前导出的keytab将失效，需要重新导出。	-
运行模式	选择HDFS连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。 ● Agent：连接实例运行在Agent上。	STANDALONE
IP与主机名映射	运行模式选择“EMBEDDED”、“STANDALONE”时，该参数有效。 如果HDFS配置文件使用主机名，需要配置IP与主机的映射。格式：IP与主机名之间使用空格分隔，多对映射使用分号或回车换行分隔。	10.1.6.9 hostname01 10.2.7.9 hostname02

参数名	说明	取值样例
Agent	运行模式选择“Agent”时，单击“选择”，选择**连接Agent**中已创建的Agent。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

3.3.5.15 配置 OBS 连接

OBS连接目的端OBS桶需添加读写权限，并在连接时不需要认证文件。

连接OBS时，相关连接参数如**表3-40**所示。

表 3-40 OBS 连接的参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	obs_link
OBS终端节点	您可以通过以下任一方式获取Endpoint信息： • OBS桶的Endpoint，可以进入OBS控制台概览页，点击桶名称后查看桶的基本信息获取。 • 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。Endpoint可从获取。 这里支持用户输入桶级别的域名，例如：test.xx.com，则在查询OBS桶的时候，只能查询到test这个桶。	-
端口	数据传输协议端口，https是443，http是80。	443
OBS桶类型	用户下拉选择即可，一般选择为“对象存储”。	对象存储
访问标识(AK)	AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前帐号的访问密钥，并获得对应的AK和SK。 您可以通过如下方式获取访问密钥。	-
密钥(SK)		-

3.3.5.16 配置 FTP/SFTP 连接

FTP/SFTP连接适用于从线下文件服务器或ECS服务器上迁移文件到OBS或数据库。

说明

当前仅支持Linux操作系统的FTP 服务器。

连接FTP或SFTP服务器时，他们的连接参数相同，如表3-41所示。

表 3-41 FTP/SFTP 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	ftp_link
主机名或IP	FTP或SFTP服务器的IP地址或者主机名。	ftp.apache.org
端口	FTP或SFTP服务器的端口，默认值为21。	21
用户名	登录FTP或SFTP服务器的用户名。	cdm
密码	登录FTP或SFTP服务器的密码。	-

3.3.5.17 配置 Redis/DCS 连接

Redis连接适用于用户在本地数据中心或ECS上自建的Redis，适用于将数据库或文件中的数据加载到Redis。

DCS适用于将数据库或文件中的数据加载到云上的DCS缓存中，从第三方云Redis服务迁移到DCS推荐使用备份恢复方式。

连接本地Redis数据库或DCS时，相关参数如表3-42所示。

表 3-42 Redis 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	redis_link
Redis部署方式	Redis部署方式： • Single：表示单机部署。 • Cluster：表示集群部署。 • Proxy：表示通过代理部署。	Single
Redis服务器列表	MongoDB服务器地址列表，输入格式为“数据库服务器域名或IP地址：端口”。多个服务器列表间以“;”分隔。	192.168.0.1:7300;192.168.0.2:7301
密码	连接Redis的密码。	-

参数名	说明	取值样例
Redis数据库索引	Redis分库的索引标识。 Redis的分库，相当于关系型数据库中的database。分库总数可以在Redis配置文件中设置，默认是16个，分库名称是一个整数（0～15），不是一个字符串。	0

3.3.5.18 配置 DDS 连接

DDS连接适用于云上的文档数据库服务，常用于从DDS同步数据到大数据平台。

连接云服务DDS时，相关参数如表3-43所示。

表 3-43 DDS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dds_link
服务器列表	服务器地址列表，输入格式为“数据库服务器域名或IP地址：端口”。多个服务器列表间以“;”分隔。	192.168.0.1:7300;192.168.0.2:7301
数据库名称	要连接的DDS数据库名称。	DB_dds
用户名	连接DDS的用户名。	cdm
密码	连接DDS的密码。	-

3.3.5.19 配置 CloudTable 连接

连接CloudTable时，相关参数如表3-44所示。

表 3-44 CloudTable 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	cloudtable_link
ZK链接地址	可通过CloudTable服务的集群管理界面获取该参数值。	cloudtable-cdm-zk1.cloudtable.com:2181,cloudtable-cdm-zk2.cloudtable.com:2181

参数名	说明	取值样例
IAM统一身份认证	如果所需连接的CloudTable集群在创建时开启了“IAM统一身份认证”，该参数需设置为“是”，否则设置为“否”。 当选择IAM统一身份认证时，需要输入用户名、AK和SK。	否
用户名	登录CloudTable集群的用户名。	admin
AK	登录CloudTable集群的访问标识。 您需要先创建当前帐号的访问密钥，并获得对应的AK和SK。	-
SK	登录CloudTable集群的密钥。 您需要先创建当前帐号的访问密钥，并获得对应的AK和SK。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hadoop_01

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

3.3.5.20 配置 CloudTable OpenTSDB 连接

连接CloudTable OpenTSDB时，相关参数如**表3-45**所示。

表 3-45 CloudTable OpenTSDB 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	TSDB_link
OpenTSDB链接地址	OpenTSDB的ZK链接地址。	opentsdb-sp8afz7bgbps5ur.cloudtable.com:4242
安全模式	选择安全或非安全模式。 选择安全模式时，需要输入项目ID、用户名、AK/SK。	Nonsecurity

参数名	说明	取值样例
项目ID	CloudTable服务所在区域的项目ID。 项目ID表示租户的资源，帐号ID对应当前帐号。用户可在对应页面下查看不同Region对应的项目ID和帐号ID。 1. 注册并登录管理控制台。 2. 在用户名的下拉列表中单击“我的凭证”。 3. 在“我的凭证”页面，查看帐号名和帐号ID，在项目列表中查看项目ID。	-
用户名	访问CloudTable服务的用户名。	admin
访问标识(AK)	访问CloudTable服务的AK和SK。 您需要先创建当前帐号的访问密钥，并获得对应的AK和SK。	-
密钥(SK)		-

3.3.5.21 配置 MongoDB 连接

MongoDB连接适用于第三方云MongoDB服务，以及用户在本地数据中心或ECS上自建的MongoDB，常用于从MongoDB同步数据到大数据平台。

连接本地MongoDB数据库时，相关参数如表3-46所示。

表 3-46 MongoDB 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mongodb_link
服务器列表	MongoDB服务器地址列表，输入格式为“数据库服务器域名或IP地址：端口”。多个服务器列表间以“;”分隔。	192.168.0.1:7300;192.168.0.2:7301
数据库名称	要连接的MongoDB数据库名称。	DB_mongodb
用户名	连接MongoDB的用户名。	cdm
密码	连接MongoDB的密码。	-

3.3.5.22 配置 Cassandra 连接

表 3-47 Cassandra 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mongodb_link

参数名	说明	取值样例
服务节点	一个或者多个节点的地址，以“;”分隔。建议同时配置多个节点。	192.168.0.1;192.168.0.2
端口	连接的Cassandra节点的端口号。	9042
用户名	连接Cassandra的用户名。	cdm
密码	连接Cassandra的密码。	-
连接超时时长	可选参数，单击“显示高级属性”后显示。 连接超时时长，单位秒。	5
读取超时时长	可选参数，单击“显示高级属性”后显示。 读取超时时长，单位秒。小于或等于0表示不超时。	12

3.3.5.23 配置 Kafka 连接

MRS Kafka

连接MRS上的Kafka数据源时，相关参数如表3-48所示。

表 3-48 MRS Kafka 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	kafka_link
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	-

参数名	说明	取值样例
用户名	需要配置MRS Manager的用户名和密码。 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 • 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。	-
密码	访问MRS Manager的用户密码。	-
认证类型	访问MRS的认证类型： • SIMPLE：非安全模式选择Simple鉴权。 • KERBEROS：安全模式选择Kerberos鉴权。	是

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

Apache Kafka

Apache Kafka连接适用于用户在本地数据中心或ECS上自建的第三方Kafka，其中本地数据中心的Kafka需通过专线连接。

连接Apache Hadoop上的Kafka数据源时，相关参数如表3-49所示。

表 3-49 Apache Kafka 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	kafka_link
Kafka broker	Kafka broker的IP地址和端口。	192.168.1.1:9092

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

3.3.5.24 配置 DMS Kafka 连接

连接DMS的Kafka队列时，相关参数如表3-50所示。

表 3-50 DMS Kafka 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dms_link
服务类型	选择DMS Kafka版本，目前只有专享版。	专享版
Kafka Broker	Kafka专享版实例的地址，格式为 host:port。	-
Kafka SASL_SSL	选择是否打开客户端连接Kafka专享版实例时SSL认证的开关。 开启Kafka SASL_SSL，则数据加密传输，安全性更高，但性能会下降。	是
用户名	开启Kafka SASL_SSL时显示该参数，表示连接DMS Kafka的用户名。	-
密码	开启Kafka SASL_SSL时显示该参数，表示连接DMS Kafka的密码。	-

3.3.5.25 配置 Elasticsearch/云搜索服务（CSS）连接

Elasticsearch

Elasticsearch连接适用于Elasticsearch服务，以及用户在本地数据中心或ECS上自建的Elasticsearch。

说明

Elasticsearch连接器只支持非安全模式。

连接Elasticsearch时，相关参数如表3-51所示。

表 3-51 Elasticsearch 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	css_link
Elasticsearch服务器列表	配置为一个或多个Elasticsearch服务器的IP地址或域名，包括端口号，格式为“ip:port”，多个地址之间使用“;”分隔。	192.168.0.1:9200; 192.168.0.2:9200

云搜索服务（CSS）

云搜索服务基于Elasticsearch引擎，该连接适用于将各类日志文件或数据库记录迁移到Elasticsearch引擎进行搜索和分析。

连接云搜索服务(CSS)时，相关参数如表3-52所示。

表 3-52 云搜索服务(CSS)连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	css_link
Elasticsearch服务器列表	配置为一个或多个Elasticsearch服务器的IP地址或域名，包括端口号，格式为“ip:port”，多个地址之间使用“;”分隔。	192.168.0.1:9200; 192.168.0.2:9200
安全模式认证	是否开启安全模式认证。 如果所需连接的CSS集群在创建时开启了“安全模式”，该参数需设置为“是”，否则设置为“否”。	是
用户名	CSS集群开启安全认证模式时显示此参数。该参数表示连接云搜索服务的用户名。	admin
密码	CSS集群开启安全认证模式时显示此参数。该参数表示连接云搜索服务的密码。	-
https访问	CSS集群开启安全认证模式时显示此参数。该参数表示开启https访问，https访问相较于http访问更安全。	是

3.3.6 管理作业

3.3.6.1 新建表/文件迁移作业

操作场景

CDM可以实现在同构、异构数据源之间进行表或文件级别的数据迁移，支持表/文件迁移的数据源请参见**表/文件迁移支持的数据源类型**。

约束限制

- 记录脏数据功能依赖于OBS服务。
- 作业导入时，JSON文件大小不超过1MB。

前提条件

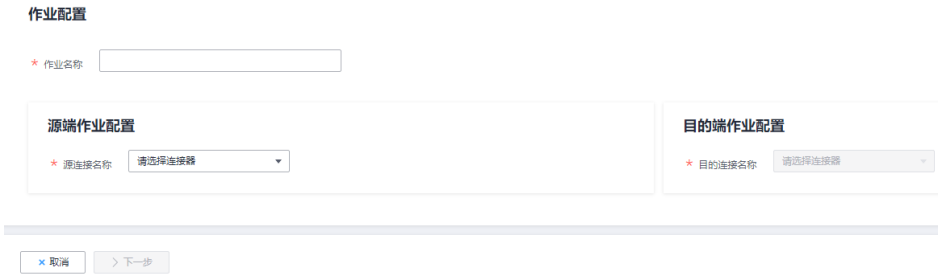
- 已新建连接。
- CDM集群与待迁移数据源可以正常通信。

操作步骤

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择集群后的“作业管理”。

步骤2 选择“表/文件迁移 > 新建作业”，进入作业配置界面。

图 3-51 新建表/文件迁移的作业



步骤3 选择源连接、目的连接：

- 作业名称：用户自定义任务名称，名称由中文、数字、字母、中划线、下划线、点号，且首字符不能是中划线或点号组成，长度必须在1到240个字符之间，例如“oracle2obs_t”。
- 源连接名称：选择待迁移数据的数据源，作业运行时将从此端复制导出数据。
- 目的连接名称：选择将数据迁移到哪个数据源，作业运行时会将数据导入此端。

步骤4 选择源连接后，配置作业参数。

每种数据源对应的作业参数不一样，其它类型数据源的作业参数请根据表3-53和表3-54选择。

表 3-53 源端作业参数说明

源端类型	说明	参数配置
OBS	支持以CSV、JSON或二进制格式抽取数据，其中二进制方式不解析文件内容，性能快，适合文件迁移。	参见配置OBS源端参数。
• MRS HDFS • FusionInsight HDFS • Apache HDFS	支持以CSV、Parquet或二进制格式抽取HDFS数据，支持多种压缩格式。	参见配置HDFS源端参数。

源端类型	说明	参数配置
• MRS HBase • FusionInsight HBase • Apache HBase • CloudTable	支持从MRS、FusionInsight HD、开源Apache Hadoop的HBase，或CloudTable服务导出数据，用户需要知道HBase表的所有列族和字段名。	参见**配置HBase/CloudTable源端参数**。
• MRS Hive • FusionInsight Hive • Apache Hive	支持从Hive导出数据，使用JDBC接口抽取数据。 Hive作为数据源，CDM自动使用Hive数据分片文件进行数据分区。	参见**配置Hive源端参数**。
DLI	支持从DLI导出数据。	参见**配置DLI源端参数**。
• FTP • SFTP	支持以CSV、JSON或二进制格式抽取FTP/SFTP的数据。	参见**配置FTP/SFTP源端参数**。
• HTTP	用于读取一个公网HTTP/HTTPS URL的文件，包括第三方对象存储的公共读取场景和网盘场景。 当前只支持从HTTP URL导出数据，不支持导入。	参见**配置HTTP源端参数**。
• 数据仓库 DWS • 云数据库 MySQL • 云数据库 SQL Server • 云数据库 PostgreSQL	支持从云端的数据库服务导出数据。	从这些数据源导出数据时，CDM使用JDBC接口抽取数据，源端作业参数相同，详细请参见**配置常见关系数据库源端参数**。
• FusionInsight LibrA	支持从FusionInsight LibrA导出数据。	
• MySQL • PostgreSQL • Oracle • Microsoft SQL Server • SAP HANA • MYCAT • 分库	这些非云服务的数据库，既可以是用户在本地数据中心自建的数据库，也可以是用户在ECS上部署的，还可以是第三方云上的数据库服务。	
• MongoDB • 文档数据库服务（DDS）	支持从MongoDB或DDS导出数据。	参见**配置MongoDB/DDS源端参数**。

源端类型	说明	参数配置
Redis	支持从开源Redis导出数据。	参见**配置Redis源端参数**。
• Apache Kafka • DMS Kafka • MRS Kafka	仅支持导出数据到云搜索服务。	参见**配置Kafka/DMS Kafka源端参数**。
• 云搜索服务 • Elasticsearch	支持从云搜索服务或Elasticsearch导出数据。	参见**配置Elasticsearch或云搜索服务源端参数**。

步骤5 配置目的端作业参数，根据目的端数据类型配置对应的参数，具体如**表3-54**所示。

表 3-54 目的端作业参数说明

目的端类型	说明	参数配置
OBS	支持使用CSV或二进制格式批量传输大量文件到OBS。	参见**配置OBS目的端参数**。
MRS HDFS	导入数据到HDFS时，支持设置压缩格式。	参见**配置HDFS目的端参数**。
MRS HBase CloudTable	支持导入数据到HBase，创建新HBase表时支持设置压缩算法。	参见**配置HBase/CloudTable目的端参数**。
MRS Hive	支持快速导入数据到MRS的Hive。	参见**配置Hive目的端参数**。
数据湖探索（DLI）	支持导入数据到DLI服务。	参见**配置DLI目的端参数**。
• 数据仓库 DWS • 云数据库 MySQL • 云数据库 SQL Server • 云数据库 PostgreSQL	支持导入数据到云端的数据库服务。	使用JDBC接口导入数据，参见**配置常见关系数据库目的端参数**。
文档数据库服务（DDS）	支持导入数据到DDS，不支持导入到本地MongoDB。	参见**配置DDS目的端参数**。
分布式缓存服务（DCS）	支持导入数据到DCS，支持“String”或“Hashmap”两种值存储方式。不支持导入数据到本地Redis。	参见**配置DCS目的端参数**。
云搜索服务（CSS）	支持导入数据到云搜索服务。	参见**配置云搜索服务目的端参数**。

步骤6 作业参数配置完成后，单击“下一步”进入字段映射的操作页面。

如果是文件类数据源（FTP/SFTP/HDFS/OBS）之间相互迁移数据，且源端“文件格式”配置为“二进制格式”（即不解析文件内容直接传输），则没有字段映射这一步骤。

其他场景下，CDM会自动匹配源端和目的端数据表字段，需用户检查字段映射关系和时间格式是否正确，例如：源字段类型是否可以转换为目的字段类型。

图 3-52 字段映射

说明

- 如果字段映射关系不正确，用户可以通过拖拽字段来调整映射关系。
- 如果在字段映射界面，CDM通过获取样值的方式无法获得所有列（例如从HBase/CloudTable/MongoDB导出数据时，CDM有较大概率无法获得所有列），则可以单击后选择“添加新字段”来手动增加，确保导入到目的端的数据完整。
- 如果是导入到数据仓库服务（DWS），则还需在目的字段中选择分布列，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。

步骤7 CDM支持字段内容转换，如果需要可单击操作列下，进入转换器列表界面，再单击“新建转换器”。

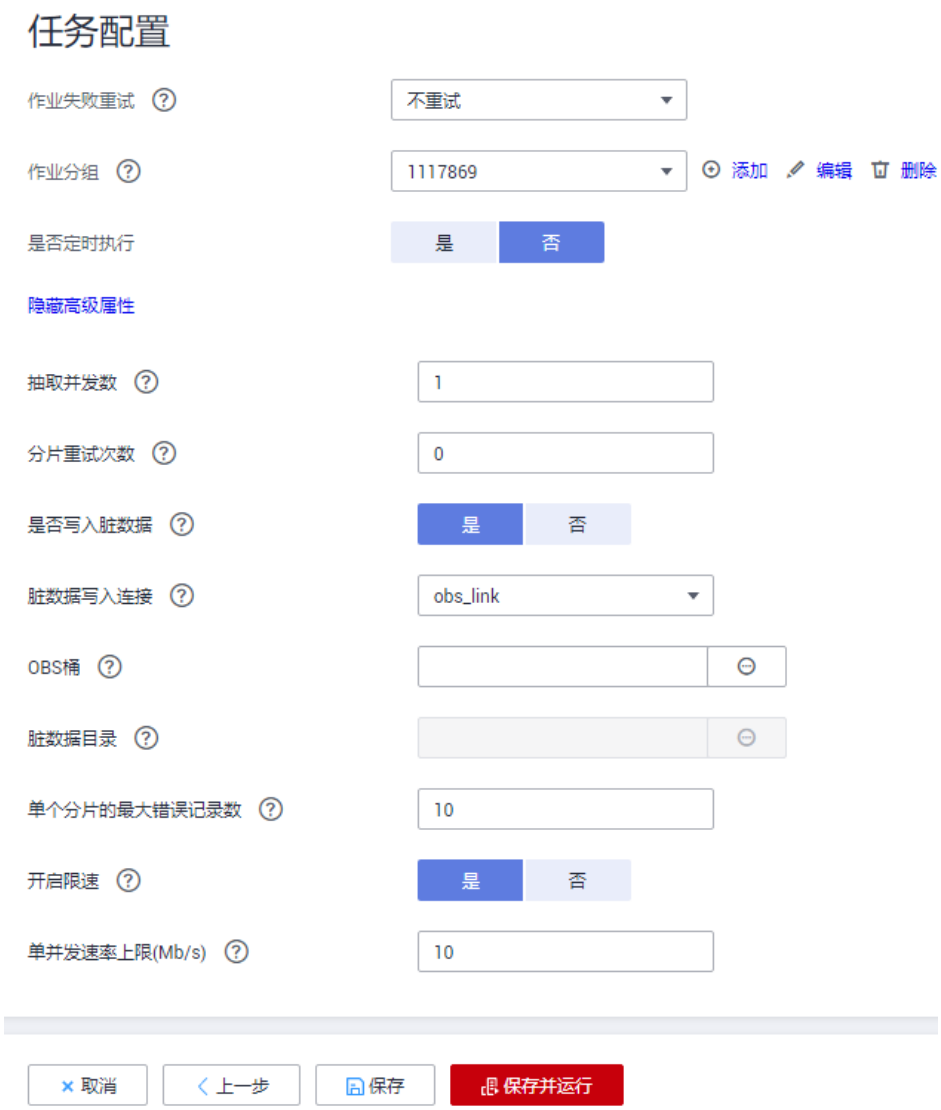
图 3-53 新建转换器

CDM支持以下转换器：

- 脱敏：隐藏字符串中的关键数据。
  例如要将“12345678910”转换为“123****8910”，则参数配置如下：
  - “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：自动删除字符串前后的空值。
- 字符串反转：自动反转字符串，例如将“ABC”转换为“CBA”。
- 字符串替换：将选定的字符串替换。
- 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。
- 去换行：将字段中的换行符（\n、\r、\r\n）删除。

步骤8 单击“下一步”配置任务参数，单击“显示高级属性”展开可选参数。

图 3-54 任务参数



各参数说明如表3-55所示。

表 3-55 任务配置参数

参数	说明	取值样例
作业失败重试	如果作业执行失败，可选择自动重试三次或者不重试。 建议仅对文件类作业或启用了导入阶段表的数据库作业配置自动重试，避免自动重试重复写入数据导致数据不一致。 **说明** 如果通过DataArts Studio数据开发使用参数传递并调度CDM迁移作业时，不能在CDM迁移作业中配置“作业失败重试”参数，如有需要请在数据开发中的CDM节点配置“失败重试”参数。	不重试
作业分组	选择作业的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。	DEFAULT
是否定时执行	如果选择“是”，可以配置作业自动启动的时间、重复周期和有效期，具体请参见配置定时任务。 **说明** 如果通过DataArts Studio数据开发调度CDM迁移作业，此处也配置了定时任务，则两种调度均会生效。为了业务运行逻辑统一和避免调度冲突，推荐您启用数据开发调度即可，无需配置CDM定时任务。	否
抽取并发数	设置同时执行的抽取任务数。并发抽取数取值范围为1-300，若配置过大，则以队列的形式进行排队。 CDM迁移作业的抽取并发量，与集群规格和表大小有关。 • 按集群规格建议每1CUs（1CUs=1核4G）配置为4。 • 表每行数据大小为1MB以下的可以多并发抽取，超过1MB的建议单线程抽取数据。 **说明** • 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。 • 单作业的抽取并发数，受到作业“配置管理”中所配置的“最大抽取并发数”影响。“最大抽取并发数”配置的是抽取并发总数。	1
加载（写入）并发数	加载（写入）时并发执行的Loader数量。 仅当HBase或Hive作为目的数据源时该参数才显示。	3
分片重试次数	每个分片执行失败时的重试次数，为0表示不重试。	0

参数	说明	取值样例
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据。 CDM中脏数据指的是数据格式非法的数据。当源数据中存在脏数据时，建议您打开此配置。否则可能导致迁移作业失败。	是
脏数据写入连接	当“是否写入脏数据”为“是”才显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	obs_link
OBS桶	当“脏数据写入连接”为OBS类型的连接时，才显示该参数。 写入脏数据的OBS桶的名称。	dirtydata
脏数据目录	“是否写入脏数据”选择为“是”时，该参数才显示。 OBS上存储脏数据的目录，只有在配置了脏数据目录的情况下才会记录脏数据。 用户可以进入脏数据目录，查看作业执行过程中处理失败的数据或者被清洗过滤掉的数据，针对该数据可以查看源数据中哪些数据不符合转换、清洗规则。	/user/dirtydir
单个分片的最大错误记录数	当“是否写入脏数据”为“是”才显示该参数。 单个map的错误记录超过设置的最大错误记录数则任务自动结束，已经导入的数据不支持回退。推荐使用临时表作为导入的目标表，待导入成功后再改名或合并到最终数据表。	0
开启限速	设置限速可以保护源端读取压力，速率代表CDM传输速率，而非网卡流量。 **说明** • 支持对Hive\DLI\关系数据库\OBS\HDFS作为目的端的作业进行单并发限速。 • 如果作业配置多并发则实际限制速率需要乘以并发数。	是
单并发速率上限(Mb/s)	开启限速情况下设置的单并发速率上限值，如果配置多并发则实际速率限制需要乘以并发数。 **说明** 限制速率为大于1的整数。	20

步骤9 单击“保存”，或者“保存并运行”回到作业管理界面，可查看作业状态。

说明

作业状态有New，Pending，Booting，Running，Failed，Succeeded。

其中“Pending”表示正在等待系统调度该作业，“Booting”表示正在分析待迁移的数据。

----结束

3.3.6.2 新建整库迁移作业

操作场景

CDM支持在同构、异构数据源之间进行整库迁移，迁移原理与**新建表/文件迁移作业**相同，关系型数据库的每张表、Redis的每个键前缀、Elasticsearch的每个类型、MongoDB的每个集合都会作为一个子任务并发执行。

支持整库迁移的数据源请参见**整库迁移支持的数据源类型**。

自动建表时的字段类型映射

CDM迁移数据库时支持在目的端自动建表。CDM在数据仓库服务（Data Warehouse Service，简称DWS）中自动建表时，DWS的表与源表的字段类型映射关系如**图3-55**所示。例如使用CDM将Oracle整库迁移到DWS，CDM在DWS上自动建表，会将Oracle的**NUMBER(3,0)**字段映射到DWS的**SMALLINT**。

图 3-55 DWS 端自动建表时的字段映射

源端数据库类型							目的端数据库类型
Oracle	MySQL	SQL Server	PostgreSQL	Db2	GaussDB	SAP HANA	DWS
NUMBER(p,0) (p=3 or p=5)	SMALLINT,TINYINT	SMALLINT,TINYINT	SMALLINT	DECIMAL	SMALLINT	SMALLINT,TINYINT	SMALLINT
NUMBER(10,0)	INT	INT	INTEGER	INT	INTEGER	INTEGER	INTEGER
NUMBER(19,0)	BIGINT	BIGINT	BIGINT	DECIMAL	BIGINT	BIGINT	BIGINT
无	无	无	OID	无	OID	CHAR(128)	OID
NUMBER(p,s) (0 < p <= 38)	DECIMAL(p,s) (0 < p <= 65)	DECIMAL(p,s) (0 < p <= 30)	NUMERIC(p,s) (p <= 1000)	DECIMAL	NUMERIC(p,s) (p <= 1000)	DECIMAL(p,s) (0 < p <= 38)	NUMERIC(p,s) (p <= 1000)
RAW	BINARY	BINARY	BYTEA	BINARY	BYTEA	BINARY	BYTEA
CHAR	CHAR	CHAR	CHAR	CHAR	CHAR	CHAR(p) (p <= 2000)	CHAR
NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR(p) (p <= 5000)	NCHAR
DATE	DATE	DATE	DATE	DATE	DATE	DATE	DATE
DATE	DATETIME	DATETIME2	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP
VARCHAR2(p) (p <= 4000)	VARCHAR	VARCHAR(p) (if p >= 8000 p=max)	VARCHAR(p) (p <= 10485760)	VARCHAR	VARCHAR(p) (p <= 10485760)	VARCHAR(p) (p <= 5000)	VARCHAR(p) (p <= 10485760)
FLOAT	DOUBLE	FLOAT	DOUBLE PRECISION	FLOAT	DOUBLE PRECISION	DOUBLE	DOUBLE PRECISION
FLOAT	REAL	FLOAT	REAL	FLOAT	REAL	REAL	REAL
CLOB	TEXT	TEXT	TEXT	TEXT	TEXT	CLOB	TEXT
DATE	无	TIME	TIME	TIME	TIME	TIME	TIME
BOOLEAN	无	无	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN

前提条件

- 已**新建连接**。
- CDM集群与待迁移数据源可以正常通信。

操作步骤

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择集群后的“作业管理”。

步骤2 选择“整库迁移 > 新建作业”，进入作业参数配置界面。

步骤3 配置源端作业参数，根据待迁移的数据库类型配置对应参数，如**表3-56**所示。

表 3-56 源端作业参数

源端数据库类型	源端参数	参数说明	取值样例
• DWS • FusionInsight LibrA • MySQL • PostgreSQL • SQL Server • Oracle • SAP HANA • MYCAT	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的帐号是否有元数据查询的权限。	schema
	Where子句	该参数适用于整库迁移中的所有子表，配置子表抽取范围的Where子句，不配置时抽取整表。如果待迁移的表中没有Where子句的字段，则迁移失败。 该参数支持配置为时间宏变量，实现抽取指定日期的数据。	age > 18 and age <= 60
	分区字段是否允许空值	选择分区字段是否允许空值。	是
HIVE	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	hivedb
HBASE CloudTable	起始时间	起始时间（包含该值）。格式为'yyyy-MM-dd hh:mm:ss', 支持dateformat时间宏变量函数。例如："2017-12-31 20:00:00" 或 "${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00" 或 ${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	-
	终止时间	终止时间（不包含该值）。格式为'yyyy-MM-dd hh:mm:ss', 支持dateformat时间宏变量函数。例如："2018-01-01 20:00:00" 或 "${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00" 或 "${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}"	-
Redis	键过滤字符	填写键过滤字符后，将迁移符合条件的键。例如：a*, 迁移所有:*	-
DDS MongoDB	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	mongodb

源端数据库类型	源端参数	参数说明	取值样例
	查询筛选	创建用于匹配文档的筛选器。例如：{HTTPStatusCode:{$gt:"400",$lt:"500"},HTTPMethod:"GET"}。	-
Elasticsearch CSS	索引	待抽取数据的索引，支持配置为通配符，一次迁移多个符合通配符条件的索引。例如这里配置为cdm*时，CDM将迁移所有名称为cdm开头的索引：cdm01、cdmB3、cdm_45…… 如果源端配置为迁移多个索引时，目的端的作业参数“索引”将不允许配置。	cdm*

步骤4 配置目的端作业参数，根据待导入数据的云服务配置对应参数，如**表3-57**所示。

表 3-57 目的端作业参数

源端数据库类型	源端参数	参数说明	取值样例
• DWS • FusionInsight LibrA • MySQL • PostgreSQL • SQL Server	-	整库迁移到关系数据库时，目的端作业参数请参见**配置常见关系数据库目的端参数**。	schema
MRS HIVE	-	整库迁移到MRS HIVE时，目的端作业参数请参见**配置Hive目的端参数**。	hivedb
MRS HBASE CloudTable	-	整库迁移到MRSHBASE或CloudTable时，目的端作业参数请参见**配置HBase/CloudTable目的端参数**。	是
MRS HDFS	-	整库迁移到MRS HDFS时，目的端作业参数请参见**配置HDFS目的端参数**。	-
OBS	-	整库迁移到OBS时，目的端作业参数请参见**配置OBS目的端参数**。	-
DCS	-	整库迁移到DCS时，目的端作业参数请参见**配置DCS目的端参数**。	-
DDS	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	mongodb

源端数据库类型	源端参数	参数说明	取值样例
	迁移行为	新增 有则替换，无则新增 替换	-
CSS	索引	待抽取数据的索引，支持配置为通配符，一次迁移多个符合通配符条件的索引。例如这里配置为cdm*时，CDM将迁移所有名称为cdm开头的索引：cdm01、cdmB3、cdm_45…… 如果源端配置为迁移多个索引时，目的端的作业参数“索引”将不允许配置。	cdm*

步骤5 如果是关系型数据库整库迁移，则作业参数配置完成后，单击“下一步”会进入表的选择界面，您可以根据自己的需求选择迁移哪些表到目的端。

步骤6 单击“下一步”配置任务参数。

图 3-56 任务参数

各参数说明如表3-58所示。

表 3-58 任务配置参数

参数	说明	取值样例
同时执行的表个数	抽取时并发执行的表的数量。	3
抽取并发数	设置同时执行的抽取任务数，一般保持默认即可。	1
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据。	是
脏数据写入连接	当“是否写入脏数据”为“是”才显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	obs_link
OBS桶	当“脏数据写入连接”为OBS类型的连接时，才显示该参数。 写入脏数据的OBS桶的名称。	dirtydata
脏数据目录	“是否写入脏数据”选择为“是”时，该参数才显示。 OBS上存储脏数据的目录，只有在配置了脏数据目录的情况下才会记录脏数据。 用户可以进入脏数据目录，查看作业执行过程中处理失败的数据或者被清洗过滤掉的数据，针对该数据可以查看源数据中哪些数据不符合转换、清洗规则。	/user/dirtydir
单个分片的最大错误记录数	当“是否写入脏数据”为“是”才显示该参数。 单个map的错误记录超过设置的最大错误记录数则任务自动结束，已经导入的数据不支持回退。推荐使用临时表作为导入的目标表，待导入成功后再改名或合并到最终数据表。	0

步骤7 单击“保存”，或者“保存并运行”。

作业任务启动后，每个待迁移的表都会生成一个子任务，单击整库迁移的作业名称，可查看子任务列表。

----结束

3.3.6.3 配置作业源端参数

3.3.6.3.1 配置 OBS 源端参数

作业中源连接为**配置OBS连接**时，源端作业参数如**表3-59**所示。

高级属性里的参数为可选参数，默认隐藏，单击界面上的“显示高级属性”后显示。

表 3-59 源端为 OBS 时的作业参数

参数类型	参数名	说明	取值样例
基本参数	桶名	待迁移数据所在的桶名。	BUCKET_2
	源目录或文件	“列表文件”选择为“否”时，才有该参数。 待迁移数据的目录或单个文件路径。文件路径支持输入多个文件（最多50个），默认以“|”分隔，也可以自定义文件分隔符。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	FROM/example.csv
	文件格式	指CDM以哪种格式解析数据，可选择以下格式： ● CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 ● JSON格式：以JSON格式解析源文件，一般都是用于迁移文件到数据表的场景。	CSV格式
	列表文件	当“文件格式”选择为“二进制格式”时，才有该参数。 打开列表文件功能时，支持读取OBS桶中文件（如txt文件）的内容作为待迁移文件的列表。该文件中的内容应为待迁移文件的绝对路径（不支持目录），例如直接写为如下内容： /052101/DAY20211110.data /052101/DAY20211111.data	是
	列表文件源连接	当“列表文件”选择为“是”时，才有该参数。可选择列表文件所在的OBS连接。	OBS_test_link
	列表文件OBS桶	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶名。	01

参数类型	参数名	说明	取值样例
	列表文件或目录	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶中的绝对路径或目录。 此处建议选择为文件的绝对路径。当选择为目录时，也支持迁移子目录中的文件，但如果目录下文件量过大，可能会导致集群内存不足。	/0521/Lists.txt
	JSON类型	当“文件格式”选择为“JSON格式”时，才有该参数。JSON文件中存储的JSON对象的类型，可以选择“JSON对象”或“JSON数组”。	JSON对象
	记录节点	当“文件格式”选择为“JSON格式”并且“JSON类型”为“JSON对象”时，才有该参数。对该JSON节点下的数据进行解析，如果该节点对应的数据为JSON数组，那么系统会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。	data.list
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	使用包围符	选择“是”时，包围符内的字段分隔符会被视为字符串值的一部分，目前CDM默认的包围符为："。	否
	使用正则表达式分隔字段	选择是否使用正则表达式分隔字段，当选择“是”时，“字段分隔符”参数无效。当“文件格式”选择为“CSV格式”时，才有该参数。	是
	正则表达式	分隔字段的正则表达式。	^(\d.*\d)(\w*) \[(.*)\] ([\w\.]*)(\w.*).*
	首行为标题行	“文件格式”选择“CSV格式”时才有该参数。在迁移CSV文件到表时，CDM默认是全部写入，如果该参数选择“是”，CDM会将CSV文件的第一行数据作为标题行，不写入目的端的表。	否

参数类型	参数名	说明	取值样例
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	GBK
	压缩格式	当“文件格式”为“CSV格式”或“JSON格式”时该参数才显示。选择对应压缩格式的源文件： ● 无：表示传输所有格式的文件。 ● GZIP：表示只传输GZIP格式的文件。 ● ZIP：表示只传输ZIP格式的文件。 ● TAR.GZ：表示只传输TAR.GZ格式的文件。	无
	压缩文件后缀	压缩格式非无时，显示该参数。 该参数需要解压缩的文件后缀名。当一批文件中以该值为后缀时，才会执行解压缩操作，否则则保持原样传输。当输入*或为空时，所有文件都会被解压。	*
	源文件处理方式	作业执行成功后对源端文件的处理方式： ● 不处理。 ● 重命名：作业执行成功后将源文件重命名，添加用户名和时间戳的后缀。 ● 删除：作业执行成功后将源文件删除。	不处理
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	否
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10

参数类型	参数名	说明	取值样例
	文件分隔符	“源目录或文件”参数中如果输入的是多个文件路径，CDM使用这里配置的文件分隔符来区分各个文件，默认为|。	|
	过滤类型	满足过滤条件的路径或文件会被传输，该参数有“无”、“通配符”和“正则表达式”三种选择。	通配符
	目录过滤器	“过滤类型”选择“通配符”时，用通配符过滤目录，符合过滤器规则的目录，允许进行迁移。支持配置多个路径，中间使用“,”分隔。	*input
	文件过滤器	“过滤类型”选择“通配符”时，用通配符过滤目录下的文件，符合过滤器规则的文件，允许进行迁移。支持配置多个文件，中间使用“,”分隔。	*.csv,*.txt
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。	2019-06-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}**表示：只迁移修改时间为当前时间以前的文件。	2019-07-01 00:00:00
	加密方式	如果源端数据是被加密过的，则CDM支持解密后再导出。这里选择是否对源端数据解密，以及选择解密算法： ● 无：不解密，直接导出。 ● AES-256-GCM：使用长度为256byte的AES对称加密算法，目前加密算法只支持AES-256-GCM（NoPadding）。该参数在目的端为加密，在源端为解密。	AES-256-GCM

参数类型	参数名	说明	取值样例
	忽略不存在原路径/文件	如果将其设为是，那么作业在源路径不存在的情况下也能成功执行。	否
	数据加密密钥	“加密方式”选择“AES-256-GCM”时显示该参数，密钥由长度64的十六进制数组成，且必须与加密时配置的“数据加密密钥”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	初始化向量	“加密方式”选择“AES-256-GCM”时显示该参数，初始化向量由长度32的十六进制数组成，且必须与加密时配置的“初始化向量”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	5C91687BA886EDCD12ACBC3FF19A3C3F
	MD5文件名后缀	“文件格式”选择“二进制格式”时，该参数才显示。 校验CDM抽取的文件，是否与源文件一致。	.md5

说明

1. 迁移文件或对象时支持文件级增量迁移（通过配置跳过重复文件实现），但不支持断点续传。

   例如要迁移3个文件，第2个文件迁移到一半时由于网络原因失败，再次启动迁移任务时，会跳过第1个文件，从第2个文件开始重新传，但不能从第2个文件失败的位置重新传。
2. 文件迁移时，单个任务支持千万数量的文件，如果待迁移目录下文件过多，建议拆分到不同目录并创建多个任务。

3.3.6.3.2 配置 HDFS 源端参数

作业中源连接为**配置HDFS连接**时，即从MRS HDFS、FusionInsight HDFS、Apache HDFS导出数据时，源端作业参数如**表3-60**所示。

表 3-60 HDFS 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	源连接名称	由用户下拉选择即可。	hdfs_to_cdm

参数类型	参数名	说明	取值样例
	源目录或文件	“列表文件”选择为“否”时，才有该参数。 待迁移数据的目录或单个文件路径。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	/user/cdm/
	文件格式	传输数据时所用的文件格式，可选择以下文件格式： ● CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 ● Parquet格式：以Parquet格式解析源文件，用于HDFS数据导到表的场景。	CSV格式
	列表文件	当“文件格式”选择为“二进制格式”时，才有该参数。 打开列表文件功能时，支持读取OBS桶中文件（如txt文件）的内容作为待迁移文件的列表。该文件中的内容应为待迁移文件的绝对路径（不支持目录），文件内容示例如下： /mrs/job-properties/application_1634891604621_0014/job.properties /mrs/job-properties/application_1634891604621_0029/job.properties	是
	列表文件源连接	当“列表文件”选择为“是”时，才有该参数。可选择列表文件所在的OBS连接。	OBS_test_link
	列表文件OBS桶	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶名。	01
	列表文件或目录	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶中的绝对路径或目录。	/0521/Lists.txt
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n

参数类型	参数名	说明	取值样例
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	首行为标题行	“文件格式”选择“CSV格式”时才有该参数。在迁移CSV文件到表时，CDM默认是全部写入，如果该参数选择“是”，CDM会将CSV文件的第一行数据作为标题行，不写入目的端的表。	否
	源文件处理方式	作业执行成功后对源端文件的处理方式： ● 不处理。 ● 重命名：作业执行成功后将源文件重命名，添加用户名和时间戳的后缀。 ● 删除：作业执行成功后将源文件删除。	不处理
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	ok.txt
	过滤类型	满足过滤条件的路径或文件会被传输，该参数有“无”、“通配符”和“正则表达式”三种选择。	-
	路径过滤器	“过滤类型”选择“通配符”时，用通配符过滤目录，符合过滤器规则的目录，允许进行迁移。支持配置多个路径，中间使用“,”分隔。	*input
	文件过滤器	“过滤类型”选择“通配符”时，用通配符过滤目录下的文件，符合过滤器规则的文件，允许进行迁移。支持配置多个文件，中间使用“,”分隔。	*.csv
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是

参数类型	参数名	说明	取值样例
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。	2019-07-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}**表示：只迁移修改时间为当前时间以前的文件。	2019-07-30 00:00:00
	创建快照	如果选择“是”，CDM读取HDFS系统上的文件时，会先对待迁移的源目录创建快照（不允许对单个文件创建快照），然后CDM迁移快照中的数据。 需要HDFS系统的管理员权限才可以创建快照，CDM作业完成后，快照会被删除。	否
	加密方式	“文件格式”选择“二进制格式”时，该参数才显示。 如果源端数据是被加密过的，则CDM支持解密后再导出。这里选择是否对源端数据解密，以及选择解密算法： ● 无：不解密，直接导出。 ● AES-256-GCM：使用长度为256byte的AES对称加密算法，目前加密算法只支持AES-256-GCM（NoPadding）。该参数在目的端为加密，在源端为解密。	AES-256-GCM
	数据加密密钥	“加密方式”选择“AES-256-GCM”时显示该参数，密钥由长度64的十六进制数组成，且必须与加密时配置的“数据加密密钥”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B

参数类型	参数名	说明	取值样例
	初始化向量	“加密方式”选择“AES-256-GCM”时显示该参数，初始化向量由长度32的十六进制数组成，且必须与加密时配置的“初始化向量”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	5C91687BA886EDCD12ACBC3FF19A3C3F
	MD5文件名后缀	“文件格式”选择“二进制格式”时，该参数才显示。 校验CDM抽取的文件，是否与源文件一致。	.md5

说明

HDFS文件编码只能为“UTF-8”，故HDFS不支持设置文件编码类型。

3.3.6.3.3 配置 HBase/CloudTable 源端参数

作业中源连接为**配置HBase连接**或**配置CloudTable连接**时，即从MRS HBase、FusionInsight HBase、Apache HBase或者CloudTable导出数据时，源端作业参数如**表3-61**所示。

说明

1. CloudTable或HBase作为源端时，CDM会读取表的首行数据作为字段列表样例，如果首行数据未包含该表的所有字段，用户需要自己手工添加字段。
2. 由于HBase的无Schema技术特点，CDM无法获知数据类型，如果数据内容是使用二进制格式存储的，CDM会无法解析。
3. 从HBase/CloudTable导出数据时，由于HBase/CloudTable是无Schema的存储系统，CDM要求源端数值型字段是以字符串格式存储，而不能是二进制格式，例如数值100需存储格式是字符串“100”，不能是二进制“01100100”。

表 3-61 HBase/CloudTable 作为源端时的作业参数

参数名	说明	取值样例
表名	导出数据的HBase表名。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	TBL_2
列族	可选参数，导出数据所属的列族。	CF1&CF2
切分Rowkey	可选参数，选择是否拆分Rowkey，默认为“否”。	是
Rowkey分隔符	可选参数，用于拆分Rowkey的分隔符，若不设置则不切分。	|

参数名	说明	取值样例
起始时间	可选参数，起始时间（包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间及以后的数据。 该参数支持配置为时间宏变量，使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	2019-01-01 20:00:00
终止时间	可选参数，终止时间（不包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间以前的数据。 该参数支持配置为时间宏变量。	2019-02-01 20:00:00

3.3.6.3.4 配置 Hive 源端参数

作业中源连接为**配置Hive连接**时，源端作业参数如**表3-62**所示。

表 3-62 Hive 作为源端时的作业参数

参数名	说明	取值样例
数据库名称	输入或选择数据库名称。单击输入框后面的按钮可进入数据库选择界面。	default
表名	输入或选择Hive表名。单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	TBL_E
读取方式	包括HDFS和JDBC两种读取方式。默认为HDFS方式，如果没有使用WHERE条件做数据过滤及在字段映射页面添加新字段的需求，选择HDFS方式即可。 • HDFS文件方式读取数据时，性能较好，但不支持使用WHERE条件做数据过滤及在字段映射页面添加新字段。 • JDBC方式读取数据时，支持使用WHERE条件做数据过滤及在字段映射页面添加新字段。	HDFS

参数名	说明	取值样例
分区过滤条件	读取方式为HDFS时，单击“显示高级属性”后显示此参数。 该参数表示抽取指定值的partition，可以配置多个值（空格分隔），也可以配置为字段取值范围，接受时间宏函数。	• 单/多值过滤："${dateformat(yyyyMMdd, -1, DAY)} ${dateformat(yyyyMMdd)}" • 范围过滤："${value} >= ${dateformat(yyyyMMdd, -7, DAY)} && ${value} < ${dateformat(yyyyMMdd)}"
Where子句	读取方式为JDBC时，单击“显示高级属性”后显示此参数。 填写该参数表示指定抽取的WHERE子句，不指定则抽取整表。如果要迁移的表中没有WHERE子句的字段，则会迁移失败。 该参数支持配置为时间宏变量，实现抽取指定日期的数据。	age > 18 and age <= 60

说明

Hive作为数据源，CDM自动使用Hive数据分片文件进行数据分区。

3.3.6.3.5 配置 DLI 源端参数

作业中源连接为**配置DLI连接**时，源端作业参数如**表3-63**所示。

表 3-63 DLI 作为源端时的作业参数

参数名	说明	取值样例
资源队列	选择目的表所属的资源队列。 DLI的default队列无法在迁移作业中使用，您需要在DLI中新建SQL队列。	cdm
数据库名称	写入数据的数据库名称。	dli
表名	写入数据的表名。	car_detail
分区	导入前清空数据，如果设置为true时，呈现此参数。 表示分区信息。	year=2020,location=sun

3.3.6.3.6 配置 FTP/SFTP 源端参数

作业中源连接为**配置FTP/SFTP连接**时，源端作业参数如**表3-64**所示。

高级属性里的参数为可选参数，默认隐藏，单击界面上的“显示高级属性”后显示。

表 3-64 FTP/SFTP 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	源目录或文件	待迁移数据的目录或单个文件路径。文件路径支持输入多个文件（最多50个），默认以“|”分隔，也可以自定义文件分隔符。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	/ftp/a.csv|/ftp/b.txt
	文件格式	指CDM以哪种格式解析数据，可选择以下格式： • CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 • 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 • JSON格式：以JSON格式解析源文件，一般都是用于迁移文件到数据表的场景。	CSV格式
	JSON类型	当“文件格式”选择为“JSON格式”时，才有该参数。JSON文件中存储的JSON对象的类型，可以选择“JSON对象”或“JSON数组”。	JSON对象
	记录节点	当“文件格式”选择为“JSON格式”并且“JSON类型”为“JSON对象”时，才有该参数。对该JSON节点下的数据进行解析，如果该节点对应的数据为JSON数组，那么系统会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。	data.list
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	使用包围符	选择“是”时，包围符内的字段分隔符会被视为字符串值的一部分，目前CDM默认的包围符为："。	否

参数类型	参数名	说明	取值样例
	使用正则表达式分隔字段	选择是否使用正则表达式分隔字段，当选择“是”时，“字段分隔符”参数无效。当“文件格式”选择为“CSV格式”时，才有该参数。	是
	正则表达式	分隔字段的正则表达式。	^(\d.*\d) (\w*) \[(.*) \] ([\w\.]*) (\w.*).*
	首行为标题行	“文件格式”选择“CSV格式”时才有该参数。在迁移CSV文件到表时，CDM默认是全部写入，如果该参数选择“是”，CDM会将CSV文件的第一行数据作为标题行，不写入目的端的表。	是
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	UTF-8
	压缩格式	当“文件格式”为“CSV格式”或“JSON格式”时该参数才显示。选择对应压缩格式的源文件： ● 无：表示传输所有格式的文件。 ● GZIP：表示只传输GZIP格式的文件。 ● ZIP：表示只传输ZIP格式的文件。 ● TAR.GZ：表示只传输TAR.GZ格式的文件。	无
	压缩文件后缀	压缩格式非无时，显示该参数。 该参数需要解压缩的文件后缀名。当一批文件中以该值为后缀时，才会执行解压缩操作，否则则保持原样传输。当输入*或为空时，所有文件都会被解压。	*
	源文件处理方式	作业执行成功后对源端文件的处理方式： ● 不处理。 ● 重命名：作业执行成功后将源文件重命名，添加用户名和时间戳的后缀。 ● 删除：作业执行成功后将源文件删除。	不处理
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	是

参数类型	参数名	说明	取值样例
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10
	文件分隔符	“源目录或文件”参数中如果输入的是多个文件路径，CDM使用这里配置的文件分隔符来区分各个文件，默认为|。	|
	过滤类型	满足过滤条件的路径或文件会被传输，该参数有“无”、“通配符”和“正则表达式”三种选择。	无
	目录过滤器	“过滤类型”选择“通配符”时，用通配符过滤目录，符合过滤器规则的目录，允许进行迁移。支持配置多个路径，中间使用“,”分隔。	*input,*out
	文件过滤器	“过滤类型”选择“通配符”时，用通配符过滤目录下的文件，符合过滤器规则的文件，允许进行迁移。支持配置多个文件，中间使用“,”分隔。	*.csv
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。	2019-07-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}**表示：只迁移修改时间为当前时间以前的文件。	2019-07-30 00:00:00

参数类型	参数名	说明	取值样例
	加密方式	如果源端数据是被加密过的，则CDM支持解密后再导出。这里选择是否对源端数据解密，以及选择解密算法： ● 无：不解密，直接导出。 ● AES-256-GCM：使用长度为256byte的AES对称加密算法，目前加密算法只支持AES-256-GCM（NoPadding）。该参数在目的端为加密，在源端为解密。	AES-256-GCM
	忽略不存在原路径/文件	如果将其设为是，那么作业在源路径不存在的情况下也能成功执行。	否
	数据加密密钥	“加密方式”选择“AES-256-GCM”时显示该参数，密钥由长度64的十六进制数组成，且必须与加密时配置的“数据加密密钥”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	初始化向量	“加密方式”选择“AES-256-GCM”时显示该参数，初始化向量由长度32的十六进制数组成，且必须与加密时配置的“初始化向量”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	5C91687BA886EDCD12ACBC3FF19A3C3F
	MD5文件名后缀	“文件格式”选择“二进制格式”时，该参数才显示。 校验CDM抽取的文件，是否与源文件一致。	.md5

3.3.6.3.7 配置 HTTP 源端参数

作业中源连接为HTTP连接时，源端作业参数如表3-65所示。当前只支持从HTTP URL导出数据，不支持导入。

表 3-65 HTTP/HTTPS 作为源端时的作业参数

参数名	说明	取值样例
文件URL	通过使用GET方法，从HTTP/HTTPS协议的URL中获取数据。 用于读取一个公网HTTP/HTTPS URL的文件，包括第三方对象存储的公共读取场景和网盘场景。	-

参数名	说明	取值样例
列表文件	选择“是”，将待上传的文本文件中所有URL对应的文件拉取到OBS，文本文件记录的是HDFS上的文件路径。	是
列表文件源连接	文本文件存储在OBS桶中，这里需要选择已建立的OBS连接。	obs_link
列表文件OBS桶	存储文本文件的OBS桶名称。	obs-cdm
列表文件或目录	在OBS中存储文本文件文件的自定义目录，多级目录可用“/”进行分隔。	test1
文件格式	当前CDM只支持选择“二进制格式”，不解析文件内容直接传输，不要求原文件格式必须为二进制。	二进制格式
压缩格式	选择对应压缩格式的源文件进行迁移： ● 无：表示传输所有格式的文件。 ● GZIP：表示只传输GZIP格式的文件。 ● ZIP：表示只传输ZIP格式的文件。 ● TAR.GZ：表示只传输TAR.GZ格式的文件。	无
压缩文件后缀	压缩格式非无时，显示该参数。 该参数需要解压缩的文件后缀名。当一批文件中以该值为后缀时，才会执行解压缩操作，否则则保持原样传输。当输入*或为空时，所有文件都会被解压。	*
文件分隔符	传输多个文件时，CDM使用这里配置的文件分隔符来区分各个文件，默认为|。列表文件选择“是”时，不显示该参数。	|
QUERY参数	● 该参数设置为“是”时，上传到OBS的对象使用的对象名，为去掉query参数后的字符。 ● 该参数设置为“否”时，上传到OBS的对象使用的对象名，包含query参数。	否
加密方式	如果源端数据是被加密过的，则CDM支持解密后再导出。这里选择是否对源端数据解密，以及选择解密算法： ● 无：不解密，直接导出。 ● AES-256-GCM：使用长度为256byte的AES对称加密算法，目前加密算法只支持AES-256-GCM（NoPadding）。该参数在目的端为加密，在源端为解密。	AES-256-GCM
忽略不存在原路径/文件	如果将其设为是，那么作业在源路径不存在的情况下也能成功执行。	否

参数名	说明	取值样例
数据加密密钥	“加密方式”选择“AES-256-GCM”时显示该参数，密钥由长度64的十六进制数组成，且必须与加密时配置的“数据加密密钥”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
初始化向量	“加密方式”选择“AES-256-GCM”时显示该参数，初始化向量由长度32的十六进制数组成，且必须与加密时配置的“初始化向量”一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	5C91687BA886EDCD12ACBC3FF19A3C3F
MD5文件名后缀	校验CDM抽取的文件，是否与源文件一致。	.md5

3.3.6.3.8 配置常见关系数据库源端参数

常见关系数据库作为源端包括数据仓库服务（DWS）、云数据库 MySQL、云数据库 PostgreSQL、云数据库 SQLServer、达梦数据库 DM、FusionInsight LibrA、PostgreSQL、Microsoft SQL Server、SAP HANA、MYCAT。

从以上数据库导出数据时，源端作业参数如**表3-66**所示。

表 3-66 常见关系数据库作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	使用SQL语句	导出关系型数据库的数据时，您可以选择使用自定义SQL语句导出。	否

	SQL语句	“使用SQL语句”选择“是”时，您可以在这里输入自定义的SQL语句，CDM将根据该语句导出数据。
		说明
		• SQL语句只能查询数据，支持join和嵌套写法，但不能有多条查询语句，比如 select * from table a; select * from table b。
		• 不支持with语句。
		• 不支持注释，比如 "--"，“/*”。
		• 不支持增删改操作，包括但不限于以下操作：
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的帐号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的帐号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： ● **table***表示导出所有以“table”开头的表。 ● ***table**表示导出所有以“table”结尾的表。 ● ***table***表示表名中只要有“table”字符串，就全部导出。	table

参数类型	参数名	说明	取值样例
高级属性	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** ● 抽取分区字段支持CHAR、VARCHAR、LONGVARCHAR、TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。 ● 当选择CHAR、VARCHAR、LONGVARCHAR抽取分区字段类型时，字段值不支持ASCII字符代码表之外的字符。	id
	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	分区字段是否允许空值	是否允许分区字段包含空值。	是
	作业拆分字段	使用该字段将作业拆分为多个子作业并发执行。	-
	拆分字段最小值	表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	根据“作业拆分字段”的最小值和最大值限定的数据范围，将作业拆分为多少个子作业执行。	-

参数类型	参数名	说明	取值样例
	按表分区抽取	从MySQL导出数据时，支持从分区表的各个分区并行抽取数据。启用该功能时，可以通过下面的“表分区”参数指定具体的MySQL表分区。 • 该功能不支持非分区表。 • 仅支持源端数据源为云数据库PostgreSQL/云数据库 MySQL时配置该参数。 • 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否

3.3.6.3.9 配置 MySQL 源端参数

作业中源连接为**配置MySQL数据库连接**，源端作业参数如**表3-67**所示。

表 3-67 MySQL 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	使用SQL语句	导出关系型数据库的数据时，您可以选择使用自定义SQL语句导出。	否

	SQL语句	“使用SQL语句”选择“是”时，您可以在这里输入自定义的SQL语句，CDM将根据该语句导出数据。
		说明
		• SQL语句只能查询数据，支持join和嵌套写法，但不能有多条查询语句，比如 select * from table a; select * from table b。
		• 不支持with语句。
		• 不支持注释，比如 "--"，“/*”。
		• 不支持增删改操作，包括但不限于以下操作：
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的帐号是否有元数据查询的权限。 **说明** 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	SCHEMA_E
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的帐号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。 **说明** 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	table
高级属性	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** • 抽取分区字段支持CHAR、VARCHAR、LONGVARCHAR、TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。 • 当选择CHAR、VARCHAR、LONGVARCHAR抽取分区字段类型时，字段值不支持ASCII字符代码表之外的字符。	id
	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

参数类型	参数名	说明	取值样例
	分区字段是否允许空值	是否允许分区字段包含空值。	是
	作业拆分字段	使用该字段将作业拆分为多个子作业并发执行。	-
	拆分字段最小值	表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	根据“作业拆分字段”的最小值和最大值限定的数据范围，将作业拆分为多少个子作业执行。	-
	按表分区抽取	从MySQL导出数据时，支持从分区表的各个分区并行抽取数据。启用该功能时，可以通过下面的“表分区”参数指定具体的MySQL表分区。 ● 该功能不支持非分区表。 ● 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否

说明

- MySQL到DWS的场景下，MySQL Binlog方式增量迁移数据功能的使用限制如下：
  1. 单个集群在当前版本中只支持一个MySQL Binlog方式的增量迁移任务。
  2. 当前版本不支持一次性删除、更新万条记录。
  3. 不支持整库迁移。
  4. 不支持DDL操作。
  5. 不支持事件（event）迁移。
  6. 当选择增量迁移时，源MySQL数据库的“binlog_format”需要设置为“ROW”。
  7. 当选择增量迁移时，增量迁移过程中如果源MySQL实例，出现因实例跨机迁移或跨机重建等导致的binlog文件ID乱序，可能导致增量迁移数据丢失。
  8. 当目的表存在主键时，如果重启CDM集群或全量迁移过程中产生增量数据，主键可能会出现重复数据，导致迁移失败。
  9. 如果目标数据库DWS存在重启行为，会导致迁移失败，需要重启CDM集群重新拉起迁移作业。
- MySQL推荐配置如下：

```
#打开bin-log功能
log-bin=mysql-bin
#行模式
binlog-format=ROW
#gtid模式，建议版本为5.6.10以上版本可用
gtid-mode=ON
enforce_gtid_consistency = ON
```

3.3.6.3.10 配置 Oracle 源端参数

作业中源连接为**配置Oracle数据库连接**，源端作业参数如**表3-68**所示。

表 3-68 Oracle 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	使用SQL语句	导出关系型数据库的数据时，您可以选择使用自定义SQL语句导出。	否

	SQL语句	“使用SQL语句”选择“是”时，您可以在这里输入自定义的SQL语句，CDM将根据该语句导出数据。
		说明
		• SQL语句只能查询数据，支持join和嵌套写法，但不能有多条查询语句，比如 select * from table a; select * from table b。
		• 不支持with语句。
		• 不支持注释，比如 "--"，“/*”。
		• 不支持增删改操作，包括但不限于以下操作：
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的帐号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： • **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 • ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 • ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的帐号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： ● **table***表示导出所有以“table”开头的表。 ● ***table**表示导出所有以“table”结尾的表。 ● ***table***表示表名中只要有“table”字符串，就全部导出。	table
高级属性	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** ● 抽取分区字段支持CHAR、VARCHAR、LONGVARCHAR、TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。 ● 当选择CHAR、VARCHAR、LONGVARCHAR抽取分区字段类型时，字段值不支持ASCII字符代码表之外的字符。	id
	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	分区字段是否允许空值	是否允许分区字段包含空值。	是

参数类型	参数名	说明	取值样例
	按表分区抽取	从Oracle导出数据时，支持从分区表的各个分区并行抽取数据。启用该功能时，可以通过下面的“表分区”参数指定具体的Oracle表分区。 • 该功能不支持非分区表。 • 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否
	表分区	输入需要迁移数据的Oracle表分区，多个分区以&分隔，不填则迁移所有分区。 如果有子分区，以“分区.子分区”的格式填写，例如“P2.SUBP1”。	P0&P1&P2.SUBP1&P2.SUBP3
	作业拆分字段	使用该字段将作业拆分为多个子作业并发执行。	-
	拆分字段最小值	表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	根据“作业拆分字段”的最小值和最大值限定的数据范围，将作业拆分为多少个子作业执行。	-

说明

Oracle作为源端时，如果未配置“抽取分区字段”或者“按表分区抽取”这2个参数，CDM自动使用ROWID进行数据分区。

3.3.6.3.11 配置分库源端参数

作业中源连接为**配置分库连接**，源端作业参数如**表3-69**所示。

表 3-69 分库作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，分库连接时此处默认展示对应第一个后端连接的表空间。用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的帐号是否有元数据查询的权限。 **说明** 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	SCHEMA_E
	表名	表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的帐号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。 **说明** 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	table
高级属性	Where子句	表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

说明

- 选择源连接名称为分库连接对应的后端连接时，此作业即为普通的MySQL作业。
- 新建源端为分库连接的作业时，在字段映射阶段，可以在源字段新增样值为“${custom(host)}”样式的自定义字段，用于在多个数据库中的多张表迁移到同一张表后，查看表的数据来源。支持的样值包括：
  - ${custom(host)}
  - ${custom(database)}
  - ${custom(fromLinkName)}
  - ${custom(schemaName)}
  - ${custom(tableName)}

3.3.6.3.12 配置 MongoDB/DDS 源端参数

从MongoDB、DDS迁移数据时，CDM会读取集合的首行数据作为字段列表样例，如果首行数据未包含该集合的所有字段，用户需要自己手工添加字段。

作业中源连接为**配置MongoDB连接**，即从本地MongoDB或DDS导出数据时，源端作业参数如**表3-70**所示。

表 3-70 MongoDB/DDS 作为源端时的作业参数

参数名	说明	取值样例
数据库名称	选择待迁移的数据库。	mongodb
集合名称	相当于关系数据库的表名。单击输入框后面的按钮可进入选择集合名的界面，用户也可以直接输入集合名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的帐号是否有元数据查询的权限。	COLLECTION
查询筛选	创建用于匹配文档的筛选条件，CDM只迁移符合条件的数据。例如： 1. 按表达式对象筛选：例如{'last_name': 'Smith'}，表示查找所有“last_name”属性值为“Smith”的文档。 2. 按参数选项筛选：例如{ x : "john" }, { z : 1 }，表示查找x=john的所有z字段。 3. 按条件筛选：例如{ "field" : { $gt: 5 } }，表示查找field字段中大于5的值。 4. 按时间宏筛选：例如{"ts":{$gte:ISODate("${dateformat(yyyy-MM-dd'T'HH:mm:ss.SSS'Z',-1,HOUR)}")}}，表示查找ts字段中大于 时间宏转换后的值。	{'last_name': 'Smith'}

3.3.6.3.13 配置 Redis 源端参数

由于分布式缓存服务（DCS）限制了获取所有Key的命令，CDM无法支持DCS作为源端，但可以作为迁移目的端，第三方云的Redis服务也无法支持作为源端。如果是用户在本地数据中心或ECS上自行搭建的Redis支持作为源端或目的端。

从本地Redis导出数据时，源端作业参数如**表3-71**所示。

表 3-71 Redis 作为源端时的作业参数

参数名	说明	取值样例
Redis键前缀	键的前缀，类似关系型数据库的表名。	TABLE
值存储类型	仅支持以下数据格式： • STRING：不带列名，如“值1，值2”形式。 • HASH：带列名，如“列名1=值1，列名2=值2”的形式。	STRING

参数名	说明	取值样例
键分隔符	用来分隔关系型数据库的表和列名。	_
值分隔符	以STRING方式存储时，列之间的分隔符。	;
字段相同	“值存储类型”参数值为“HASH”显示该参数。 哈希键内有相同的字段。	是

3.3.6.3.14 配置 Kafka/DMS Kafka 源端参数

作业中源连接为**配置Kafka连接**或**配置DMS Kafka连接**时，源端作业参数如**表3-72**所示。

表 3-72 Kafka 作为源端时的作业参数

参数	说明	取值样例
Topics	支持单个或多个topic。	est1,est2
偏移量参数	从Kafka拉取数据时的初始偏移量： • 最新：最大偏移量，即拉取最新的数据。 • 最早：最小偏移量，即拉取最早的数据。 • 已提交：拉取已提交的数据。 • 时间范围：拉取时间范围内的数据。	最新
是否持久运行	用户自定义是否永久运行。	是
消费组ID	用户指定消费组ID。 如果是从DMS Kafka导出数据，专享版请任意输入，标准版请输入有效的消费组ID。	sumer-group
数据格式	解析数据时使用的格式： • 二进制格式：适用于文件迁移场景，不解析数据内容原样传输。 • CSV格式：以CSV格式解析源数据。 • JSON：以JSON格式解析源数据。 • CDC（DRS_JSON）：以DRS_JSON格式解析源数据。	二进制格式
字段分隔符	默认为空格，使用Tab键作为分隔符请输入“\t”。	,
最大消息数/poll	可选参数，每次向Kafka请求数据限制最大请求记录数。	100
最大时间间隔/poll	可选参数，向Kafka请求数据的最大时间间隔。	100

3.3.6.3.15 配置 Elasticsearch 或云搜索服务源端参数

作业中源连接为**配置Elasticsearch/云搜索服务（CSS）连接**时，源端作业参数如**表3-73**所示。

表 3-73 Elasticsearch 或云搜索服务作为源端时的作业参数

参数名	说明	取值样例
索引	Elasticsearch的索引，类似关系数据库中的数据库名称。索引名称只能全部小写，不能有大写。	index
类型	Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。	type
拆分nested类型字段	可选参数，选择是否将nested字段的json内容拆分，例如：将“a:{ b:{ c:1, d:{ e:2, f:3 } } }”拆成三个字段“a.b.c”、“a.b.d.e”、“a.b.d.f”。	否

参数名	说明	取值样例
过滤条件	可选参数，CDM只迁移满足过滤条件的数据。 ● 当前仅支持通过Elasticsearch的query string（即q语法）方式对源数据进行过滤。q语法使用方式介绍如下： – 精确匹配时，直接使用***column*:*data***格式进行匹配过滤。其中column表示字段名，data表示查询条件，例如“last_name:Smith”。 另外，如果查询条件data为带空格的字符串，则需要用双引号包围。如果不指定column，则会对所有字段以data进行匹配。 – 多条查询条件时，可通过连接词组合多个查询条件，格式为***column1*:*data1* AND *column2*:*data2***。其中，中间的连接词必须用全大写，可以为“AND”、“OR”或“NOT”，且连接词前后要有空格。 例如：“last_name:Smith AND last_name:John”。 – 范围匹配时，可以直接使用条件表达式的方式进行过滤，格式为***column*:>*data***。其中，操作符支持“>”、“>=”、“<”或“<=”。 例如：“time:>=1636905600000 AND time:<1637078400000”。也可以配合时间宏变量使用，如“createTime:>=${timestamp(dateformat(yyyyMMdd,-1,DAY))} AND createTime:< ${timestamp(dateformat(yyyyMMdd))}”。 – 范围匹配时，也支持使用范围区间语法的方式进行过滤，格式为***column*:{*data1* TO *data2*}**。其中，{”、“}”代表不包含该值，“[”、“]”代表包含该值，TO必须大写且前后要有空格，*代表所有。 例如：“time:{1636992000000 TO *]”,表示过滤time字段中大于1636992000000的所有数据。也可以配合时间宏变量使用，如“createTime:[${timestamp(dateformat(yyyyMMdd,-1,DAY))} TO ${timestamp(dateformat(yyyyMMdd))}}”。 ● 暂不支持通过Elasticsearch的query DSL（即DSL语法，Domain Sepcified Language）查询方式对源数据进行过滤。	last_name:Smith
抽取元字段	表示是否抽取索引的元字段，目前只支持（_index、_type、_id、_score）例如：_index、_type、_id、_score	是

在下一步的字段映射中，源端和目的端均支持配置自定义字段。

图 3-57 配置自定义字段

3.3.6.3.16 配置 OpenTSDB 源端参数

作业中源连接为**配置CloudTable OpenTSDB连接**时，源端作业参数如**表3-74**所示。

表 3-74 OpenTSDB 作为源端时的作业参数

参数名	说明	取值样例
开始时间	查询的起始时间，格式为yyyyMMddHHmmdd的字符串或时间戳。	20180920145505
结束时间	可选参数，查询的终止时间，格式为yyyyMMddHHmmdd的字串或时间戳。	20180921145505
指标	输入迁移哪个指标的数据，或选择OpenTSDB中已存在的指标。	city.temp
聚合函数	输入聚合函数。	sum
标记	可选参数，如果这里有输入标记，则只迁移标记的数据。	tagk1:tagv1,tagk2:tagv2

3.3.6.4 配置作业目的端参数

3.3.6.4.1 配置 OBS 目的端参数

作业中目的连接为**配置OBS连接**时，即导入数据到云服务OBS时，目的端作业参数如**表3-75**所示。

高级属性里的参数为可选参数，默认隐藏，单击界面上的“显示高级属性”后显示。

表 3-75 OBS 作为目的端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	桶名	写入数据的OBS桶名。	bucket_2
	写入目录	写入数据到OBS服务器的目录，目录前面不加“/”。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	directory/
	文件格式	写入后的文件格式，可选择以下文件格式： ● CSV格式：按CSV格式写入，适用于数据表到文件的迁移。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，CDM会原样写入文件，不改变原始文件格式，适用于文件到文件的迁移。 如果是文件类数据源（FTP/SFTP/HDFS/OBS）之间相互迁移数据，此处的“文件格式”只能选择与源端的文件格式一致。	CSV格式
	重复文件处理方式	只有文件名和文件大小都相同才会判定为重复文件。写入时如果出现文件重复，可选择如下处理方式： ● 替换重复文件 ● 跳过重复文件 ● 停止任务	跳过重复文件
高级属性	加密方式	选择是否对上传的数据进行加密，以及加密方式： ● 无：不加密，直接写入数据。 ● KMS：使用数据加密服务中的KMS进行加密。如果启用KMS加密则无法进行数据的MD5校验。 ● AES-256-GCM：使用长度为256byte的AES对称加密算法，目前加密算法只支持AES-256-GCM（NoPadding）。该参数在目的端为加密，在源端为解密。	KMS

参数类型	参数名	说明	取值样例
	KMS ID	写入文件时加密使用的密钥，“加密方式”选择“KMS”时显示该参数。单击输入框后面的，可以直接选择在数据加密服务中已创建好的KMS密钥。 ● 当使用与CDM集群相同项目下的KMS密钥时，不需要修改下面的“项目ID”参数。 ● 当用户使用其它项目下的KMS密钥时，需要修改下面的“项目ID”参数。	53440ccb-3e73-4700-98b5-71ff5476e621
	项目ID	KMS ID所属的项目ID，该参数默认值为当前CDM集群所属的项目ID。 ● 当“KMS ID”与CDM集群在同一个项目下时，这里的“项目ID”保持默认即可。 ● 当“KMS ID”使用的是其它项目下的KMS ID时，这里需要修改为KMS所属的项目ID。	9bd7c4bd54e5417198f9591bef07ae67
	数据加密密钥	“加密方式”选择“AES-256-GCM”时显示该参数，密钥由长度64的十六进制数组成。 请您牢记这里配置的“数据加密密钥”，解密时的密钥与这里配置的必须一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	初始化向量	“加密方式”选择“AES-256-GCM”时显示该参数，初始化向量由长度32的十六进制数组成。 请您牢记这里配置的“初始化向量”，解密时的初始化向量与这里配置的必须一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	5C91687BA886EDCD12ACBC3FF19A3C3F
	复制Content-Type属性	“文件格式”为“二进制”，且源端、目的端都为对象存储时，才有该参数。 选择“是”后，迁移对象文件时会复制源文件的Content-Type属性，主要用于静态网站的迁移场景。 归档存储的桶不支持设置Content-Type属性，所以如果开启了该参数，目的端选择写入的桶时，必须选择非归档存储的桶。	否
	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。“文件格式”为“二进制格式”时该参数值无效。	\n
	字段分隔符	文件中的字段分隔符。“文件格式”为“二进制格式”时该参数值无效。	,

参数类型	参数名	说明	取值样例
	写入文件大小	源端为数据库时该参数才显示，支持按大小分成多个文件存储，避免导出的文件过大，单位为MB。	1024
	校验MD5值	使用“二进制格式”传输文件时，才能校验MD5值。选择校验MD5值时，无法使用KMS加密。 计算源文件的MD5值，并与OBS返回的MD5值进行校验。如果源端已经存在MD5文件，则直接读取源端的MD5文件与OBS返回的MD5值进行校验。	是
	记录校验结果	当选择校验MD5值时，可以选择是否记录校验结果。	是
	校验结果写入连接	可以指定任意一个OBS连接，将MD5校验结果写入该连接的桶下。	obslink
	OBS桶	写入MD5校验结果的OBS桶。	cdm05
	写入目录	写入MD5校验结果的目录。	/md5/
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。“文件格式”为“二进制格式”时该参数值无效。	GBK
	使用包围符	“文件格式”为“CSV格式”，才有该参数，用于将数据库的表迁移到文件系统的场景。 选择“是”时，如果源端数据表中的某一个字段内容包含字段分隔符或换行符，写入目的端时CDM会使用双引号（"）作为包围符将该字段内容括起来，作为一个整体存储，避免其中的字段分隔符误将一个字段分隔成两个，或者换行符误将字段换行。例如：数据库中某字段为hello,world，使用包围符后，导出到CSV文件的时候数据为"hello,world"。	否
	首行为标题行	从关系型数据库导出数据到OBS，“文件格式”为“CSV格式”时，才有该参数。 在迁移表到CSV文件时，CDM默认是不迁移表的标题行，如果该参数选择“是”，CDM在才会将表的标题行数据写入文件。	否
	作业成功标识文件	当作业执行成功时，会在写入目录下生成一个标识文件，文件名由用户指定。不指定时默认关闭该功能。	finish.txt
	自定义目录层次	选择“是”时，支持迁移后的文件按照自定义的目录存储。即只迁移文件，不迁移文件所归属的目录。	是

参数类型	参数名	说明	取值样例
	目录层次	自定义迁移后文件的存储路径，支持时间宏变量。	${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}
	自定义文件名	从关系型数据库导出数据到OBS，且“文件格式”为“CSV格式”时，才有该参数。 用户可以通过该参数自定义OBS端生成的文件名，支持以下自定义方式： • 字符串，支持特殊字符。例如“cdm#”，则生成的文件名为“cdm#.csv”。 • 时间宏，例如“${timestamp()}”，则生成的文件名为“1554108737.csv”。 • 表名宏，例如“${tableName}”，则生成的文件名为“sqltabname.csv”。 • 版本宏，例如“${version}”，则生成的文件名为“v1.csv”。 • 字符串和宏（时间宏/表名宏/版本宏）任意组合，例如“cdm#${timestamp()}_${version}”，则生成的文件名为“cdm#1554108737_v1.csv”。	cdm

3.3.6.4.2 配置 HDFS 目的端参数

作业中目的连接为**配置HDFS连接**时，即导入数据到以下数据源时，目的端作业参数如**表3-76**所示。

表 3-76 HDFS 作为目的端时的作业参数

参数名	说明	取值样例
写入目录	写入数据到HDFS服务器的目录。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	/user/output

参数名	说明	取值样例
文件格式	写入后的文件格式，可选择以下文件格式： ● CSV格式：按CSV格式写入，适用于数据表到文件的迁移。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，CDM会原样写入文件，不改变原始文件格式，适用于文件到文件的迁移。 如果是文件类数据源（FTP/SFTP/HDFS/OBS）之间相互迁移数据，此处的“文件格式”只能选择与源端的文件格式一致。	CSV格式
重复文件处理方式	只有文件名和文件大小都相同才会判定为重复文件。写入时如果出现文件重复，可选择如下处理方式： ● 替换重复文件 ● 跳过重复文件 ● 停止任务	停止任务
压缩格式	写入文件后，选择对文件的压缩格式。支持以下压缩格式： ● NONE：不压缩。 ● DEFLATE：压缩为DEFLATE格式。 ● GZIP：压缩为GZIP格式。 ● BZIP2：压缩为BZIP2格式。 ● LZ4：压缩为LZ4格式。 ● SNAPPY：压缩为SNAPPY格式。	SNAPPY
换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。“文件格式”为“二进制格式”时该参数值无效。	\n
字段分隔符	文件中的字段分隔符。“文件格式”为“二进制格式”时该参数值无效。	,
使用包围符	“文件格式”为“CSV格式”，才有该参数，用于将数据库的表迁移到文件系统的场景。 选择“是”时，如果源端数据表中的某一个字段内容包含字段分隔符或换行符，写入目的端时CDM会使用双引号（"）作为包围符将该字段内容括起来，作为一个整体存储，避免其中的字段分隔符误将一个字段分隔成两个，或者换行符误将字段换行。例如：数据库中某字段为hello,world，使用包围符后，导出到CSV文件的时候数据为"hello,world"。	否
首行为标题行	在迁移表到CSV文件时，CDM默认是不迁移表的标题行，如果该参数选择“是”，CDM在才会将表的标题行数据写入文件。	否

参数名	说明	取值样例
写入到临时文件	将二进制文件先写入到临时文件（临时文件以“.tmp”作为后缀），迁移成功后，再进行rename或move操作，在目的端恢复文件。	否
作业成功标识文件	当作业执行成功时，会在写入目录下生成一个标识文件，文件名由用户指定。不指定时默认关闭该功能。	finish.txt
自定义目录层次	支持用户自定义文件的目录层次。例如：【表名】/【年】/【月】/【日】/【数据文件名】. csv	-
目录层次	指定文件的目录层次，支持时间宏（时间格式为yyyy/MM/dd）。不填默认为不带层次目录。例如：${dateformat(yyyy/MM/dd, -1, DAY)}	-
加密方式	“文件格式”选择“二进制格式”时，该参数才显示。 选择是否对写入的数据进行加密： • 无：不加密，直接写入数据。 • AES-256-GCM：使用长度为256byte的AES对称加密算法，目前加密算法只支持AES-256-GCM（NoPadding）。该参数在目的端为加密，在源端为解密。	AES-256-GCM
数据加密密钥	“加密方式”选择“AES-256-GCM”时显示该参数，密钥由长度64的十六进制数组成。 请您牢记这里配置的“数据加密密钥”，解密时的密钥与这里配置的必须一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
初始化向量	“加密方式”选择“AES-256-GCM”时显示该参数，初始化向量由长度32的十六进制数组成。 请您牢记这里配置的“初始化向量”，解密时的初始化向量与这里配置的必须一致。如果不一致系统不会报异常，只是解密出来的数据会错误。	5C91687BA886EDCD12ACBC3FF19A3C3F

说明

HDFS文件编码只能为“UTF-8”，故HDFS不支持设置文件编码类型。

3.3.6.4.3 配置 HBase/CloudTable 目的端参数

作业中目的连接为**配置HBase连接**或**配置CloudTable连接**时，即导入数据到以下数据源时，目的端作业参数如**表3-77**所示。

表 3-77 HBase/CloudTable 作为目的端时的作业参数

参数名	说明	取值样例
表名	写入数据的HBase表名。如果是创建新HBase表，支持从源端拷贝字段名。单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	TBL_2
导入前清空数据	选择目的端表中数据的处理方式： ● 是：任务启动前会清除目标表中数据。 ● 否：导入前不清空目标表中的数据，如果选“否”且表中有数据，则数据会追加到已有的表中。	是
Row key拼接分隔符	可选参数，用于多列合并作为rowkey，默认为空格。	,
Rowkey冗余	可选参数，是否将选做Rowkey的数据同时写入HBase的列，默认值“否”。	否
压缩算法	可选参数，创建新HBase表时采用的压缩算法，默认为值“NONE”。 ● NONE：不压缩。 ● SNAPPY：压缩为Snappy格式。 ● GZ：压缩为GZ格式。	NONE
WAL开关	选择是否开启HBase的预写日志机制（WAL，Write Ahead Log）。 ● 是：开启后如果出现HBase服务器宕机，则可以从WAL中回放执行之前没有完成的操作。 ● 否：关闭时能提升写入性能，但如果HBase服务器宕机可能会造成数据丢失。	否
匹配数据类型	● 是：源端数据库中的Short、Int、Long、Float、Double、Decimal类型列的数据，会转换为Byte[]数组（二进制）写入HBase，其他类型的按字符串写入。如果这几种类型中，有合并做rowkey的，则依然当字符串写入。 该功能作用是：降低存储占用空间，存储更高效；特定场景下rowkey分布更均匀。 ● 否：源端数据库中所有类型的数据，都会按照字符串写入HBase。	否

3.3.6.4.4 配置 Hive 目的端参数

作业中目的连接为**配置Hive连接**时，目的端作业参数如**表3-78**所示。

表 3-78 Hive 作为目的端时的作业参数

参数名	说明	取值样例
数据库名称	输入或选择写入数据的数据库名称。单击输入框后面的按钮可进入数据库选择界面。	default
自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 ● 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。	不自动创建
表名	输入或选择写入数据的目标表名。 单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	TBL_X
导入前清空数据	选择目的端表中数据的处理方式： ● 是：任务启动前会清除目标表中数据。 ● 否：导入前不清空目标表中的数据，如果选“否”且表中有数据，则数据会追加到已有的表中。	是
待清空分区	“导入前清空数据”设置为“是”时，呈现此参数。 填写待清空分区信息后，表示清空该分区的数据。	单分区：year=2020,location=sun; 多分区：['year=2020,location=sun','year=2021,location=earth'].

说明

1. Hive作为目的端时，会自动创建存储格式为ORC的表。
2. Hive作为迁移的目的时，如果存储格式为Textfile，在Hive创建表的语句中需要显式指定分隔符。例如：

```
CREATE TABLE csv_tbl(
smallint_value smallint,
tinyint_value tinyint,
int_value int,
bigint_value bigint,
float_value float,
double_value double,
decimal_value decimal(9, 7),
timestmamp_value timestamp,
date_value date,
varchar_value varchar(100),
string_value string,
char_value char(20),
boolean_value boolean,
binary_value binary,
varchar_null varchar(100),
string_null string,
char_null char(20),
int_null int
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = "\t",
"quoteChar"     = "'",
"escapeChar"    = "\\"
)
STORED AS TEXTFILE;
```

3.3.6.4.5 配置常见关系数据库目的端参数

常见关系数据库作为目的端包括云数据库 MySQL、云数据库 SQL Server、云数据库 PostgreSQL。

将数据导入到以上数据源时，目的端作业参数如表3-79所示。

表 3-79 常见关系型数据库作为目的端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	待写入数据的数据库名称，支持自动创建Schema。单击输入框后面的按钮可选择模式或表空间。	schema
	自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 ● 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。	不自动创建

参数类型	参数名	说明	取值样例
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	table
	导入开始前	导入数据前，选择是否清除目的表的数据： ● 不清除：写入数据前不清除目标表中数据，数据追加写入。 ● 清除全部数据：写入数据前会清除目标表中数据。 ● 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
	where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
	约束冲突处理	当迁移数据出现冲突时的处理方式。 ● insert into：当存在主键、唯一性索引冲突时，数据无法写入并将以脏数据的形式存在。 ● replace into：当存在主键、唯一性索引冲突时，会先删除原有行、再插入新行，替换原有行的所有字段。 ● on duplicate key update，当存在主键、唯一性索引冲突时，目的表中约束冲突的行除开唯一约束列的其他数据列将被更新。	insert into
	loader线程数	每个loader内部启动的线程数，可以提升写入并发数。 **说明** 不支持“约束冲突处理”策略为“replace into”或“on duplicate key update”的并发场景。	1
高级参数	先导入阶段表	如果选择“是”，则启用事务模式迁移，CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中，导入失败则将目的表回滚到作业开始之前的状态。 默认为“否”，CDM直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否

参数类型	参数名	说明	取值样例
	扩大字符字段长度	选择自动创表时，迁移过程中可将字符类型的字段长度扩大为原来的3倍，再写入到目的表中。如果源端数据库与目的端数据库字符编码不一样，但目的表字符类型字段与源表一样，在迁移数据时，可能会有出现长度不足的错误。 **说明** 当启动该功能时，也会导致部分字段消耗用户相应的3倍存储空间。	否
	使用非空约束	当选择自动创建目的表时，如果选择使用非空约束，则目的表字段的是否非空约束，与原表具有相应非空约束的字段保持一致。	是
	导入前准备语句	执行任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句。	create temp table
	导入后完成语句	执行任务之后执行的SQL语句，目前仅允许执行一条SQL语句。	merge into

3.3.6.4.6 配置 DWS 目的端参数

作业中目的连接为**配置DWS连接**，目的端作业参数如**表3-80**所示。

表 3-80 目的端为 DWS 时的作业参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	待写入数据的数据库名称，支持自动创建Schema。单击输入框后面的按钮可选择模式或表空间。	schema
	自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 ● 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。 当选择在DWS端自动创表时，DWS的表与源表的字段类型映射关系见**在DWS端自动建表时的字段类型映射**。	不自动创建

参数类型	参数名	说明	取值样例
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据。	table
	是否压缩	导入数据到DWS且选择自动创表时，用户可以指定是否压缩存储。	否
	存储模式	导入数据到DWS且选择自动创表时，用户可以指定存储模式： ● 行模式：表的数据将以行式存储，适用于点查询（返回记录少，基于索引的简单查询），或者增删改比较多的场景。 ● 列模式：表的数据将以列式存储，适用于统计分析类查询（group、join多的场景），或者即席查询（查询条件不确定，行模式表扫描难以使用索引）的场景。	行模式
	导入模式	导入数据到DWS时，用户可以指定导入模式： ● COPY模式，源数据经过管理节点后，复制到DWS的DataNode节点。 ● UPSERT模式，数据发生主键或唯一约束冲突时，更新除了主键和唯一约束列的其他列数据。	COPY
	导入开始前	导入数据前，选择是否清除目的表的数据： ● 不清除：写入数据前不清除目标表中数据，数据追加写入。 ● 清除全部数据：写入数据前会清除目标表中数据。 ● 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
	where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
	约束冲突处理	当迁移数据出现冲突时的处理方式。 ● insert into：当存在主键、唯一性索引冲突时，数据无法写入并将以脏数据的形式存在。 ● replace into：当存在主键、唯一性索引冲突时，会先删除原有行、再插入新行，替换原有行的所有字段。 ● on duplicate key update，当存在主键、唯一性索引冲突时，目的表中约束冲突的行除开唯一约束列的其他数据列将被更新。	insert into

参数类型	参数名	说明	取值样例
	loader线程数	每个loader内部启动的线程数，可以提升写入并发数。 **说明** 不支持“约束冲突处理”策略为“replace into”或“on duplicate key update”的并发场景。	1
高级参数	先导入阶段表	如果选择“是”，则启用事务模式迁移，CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中，导入失败则将目的表回滚到作业开始之前的状态。 默认为“否”，CDM直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否
	扩大字符字段长度	选择自动创表时，迁移过程中可将字符类型的字段长度扩大为原来的3倍，再写入到目的表中。如果源端数据库与目的端数据库字符编码不一样，但目的表字符类型字段与源表一样，在迁移数据时，可能会有出现长度不足的错误。 应用场景主要是将字符字段导入到DWS时，需要自动将字符长度放大3倍。 在导入字符到DWS时，如果作业执行失败，且日志中出现类似“value too long for type character varying”的错误，则可以通过启用该功能解决。 **说明** 当启动该功能时，也会导致部分字段消耗用户相应的3倍存储空间。	否
	使用非空约束	当选择自动创建目的表时，如果选择使用非空约束，则目的表字段的是否非空约束，与原表具有相应非空约束的字段保持一致。	是
	导入前准备语句	执行任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句。	create temp table
	导入后完成语句	执行任务之后执行的SQL语句，目前仅允许执行一条SQL语句。	merge into

在 DWS 端自动建表时的字段类型映射

CDM在数据仓库服务（Data Warehouse Service，简称DWS）中自动建表时，DWS的表与源表的字段类型映射关系如图3-58所示。例如使用CDM将Oracle整库迁移到

DWS，CDM在DWS上自动建表，会将Oracle的**NUMBER(3,0)**字段映射到DWS的**SMALLINT**。

图 3-58 自动建表的字段映射

源端数据库类型							目的端数据库类型
Oracle	MySQL	SQL Server	PostgreSQL	Db2	GaussDB	SAP HANA	DWS
NUMBER(p,0) (p=3 or p=5)	SMALLINT,TINYINT	SMALLINT,TINYINT	SMALLINT	DECIMAL	SMALLINT	SMALLINT,TINYINT	SMALLINT
NUMBER(10,0)	INT	INT	INTEGER	INT	INTEGER	INTEGER	INTEGER
NUMBER(19,0)	BIGINT	BIGINT	BIGINT	DECIMAL	BIGINT	BIGINT	BIGINT
无	无	无	OID	无	OID	CHAR(128)	OID
NUMBER(p,s) (0 < p <= 38)	DECIMAL(p,s) (0 < p <= 65)	DECIMAL(p,s) (0 < p <= 30)	NUMERIC(p,s) (p <= 1000)	DECIMAL	NUMERIC(p,s) (p <= 1000)	DECIMAL(p,s) (0 < p <= 38)	NUMERIC(p,s) (p <= 1000)
RAW	BINARY	BINARY	BYTEA	BINARY	BYTEA	BINARY	BYTEA
CHAR	CHAR	CHAR	CHAR	CHAR	CHAR	CHAR(p) (p <= 2000)	CHAR
NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR(p) (p <= 5000)	NCHAR
DATE	DATE	DATE	DATE	DATE	DATE	DATE	DATE
DATE	DATETIME	DATETIME2	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP
VARCHAR2(p) (p <= 4000)	VARCHAR	VARCHAR(p) (if p >= 8000 p=max)	VARCHAR(p) (p <= 10485760)	VARCHAR	VARCHAR(p) (p <= 10485760)	VARCHAR(p) (p <= 5000)	VARCHAR(p) (p <= 10485760)
FLOAT	DOUBLE	FLOAT	DOUBLE PRECISION	FLOAT	DOUBLE PRECISION	DOUBLE	DOUBLE PRECISION
FLOAT	REAL	FLOAT	REAL	FLOAT	REAL	REAL	REAL
CLOB	TEXT	TEXT	TEXT	TEXT	TEXT	CLOB	TEXT
DATE	无	TIME	TIME	TIME	TIME	TIME	TIME
BOOLEAN	无	无	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN

3.3.6.4.7 配置 DDS 目的端参数

作业中目的连接为**配置DDS连接**，即导入数据到文档数据库服务（DDS）时，目的端作业参数如**表3-81**所示。

表 3-81 DDS 作为目的端时的作业参数

参数名	说明	取值样例
数据库名称	选择待导入数据的数据库。	mongodb
集合名称	选择待导入数据的集合，相当于关系数据库的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的帐号是否有元数据查询的权限。	COLLECTION

3.3.6.4.8 配置 DCS 目的端参数

当作业将数据导入到分布式缓存服务（DCS）时，目的端作业参数如**表3-82**所示。

表 3-82 DCS 作为目的端时的作业参数

参数名	说明	取值样例
Redis键前缀	键的前缀，类似关系型数据库的表名。	TABLE
值存储类型	仅支持以下数据格式： • STRING：不带列名，如“值1，值2”形式。 • HASH：带列名，如“列名1=值1，列名2=值2”的形式。	STRING
键分隔符	用来分隔关系型数据库的表和列名。	_
值分隔符	以STRING方式存储时，列之间的分隔符。	;

3.3.6.4.9 配置云搜索服务目的端参数

作业中目的连接为**配置Elasticsearch/云搜索服务（CSS）连接**，即将数据导入到云搜索服务时，目的端作业参数如**表3-83**所示。

表 3-83 Elasticsearch 作为目的端时的作业参数

参数名	说明	取值样例
索引	待写入数据的Elasticsearch的索引，类似关系数据库中的数据库名称。CDM支持自动创建索引和类型，索引和类型名称只能全部小写，不能有大写。	index
类型	待写入数据的Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。	type
管道ID	需要先在kibana中创建管道ID，这里才可以选择，该参数用于数据传到Elasticsearch后，通过Elasticsearch的数据转换pipeline进行数据格式变换。	pipeline_id

参数名	说明	取值样例
定时创索引	对于持续写入数据到Elasticsearch的流式作业，CDM支持在Elasticsearch中定时创建新索引并写入数据，方便用户后期删除过期的数据。支持按以下周期创建新索引： • 每小时：每小时整点创建新索引，新索引的命名格式为“索引名+年+月+日+小时”，例如“index2018121709”。 • 每天：每天零点零分创建新索引，新索引的命名格式为“索引名+年+月+日”，例如“index20181217”。 • 每周：每周周一的零点零分创建新索引，新索引的命名格式为“索引名+年+周”，例如“index201842”。 • 每月：每月一号零点零分创建新索引，新索引的命名格式为“索引名+年+月”，例如“index201812”。 • 不创建：选择此项表示不创建定时索引。 从文件类抽取数据时，必须配置单个抽取（“抽取并发数”参数配置为1），否则该参数无效。	每小时

3.3.6.4.10 配置 DLI 目的端参数

作业中目的连接为**配置DLI连接**，即将数据导入到数据湖探索服务（DLI）时，目的端作业参数如**表3-84**所示。

说明

使用CDM服务迁移数据到DLI时，当前用户需要先开通OBS读取权限。

表 3-84 DLI 作为目的端时的作业参数

参数名	说明	取值样例
资源队列	选择目的表所属的资源队列。 DLI的default队列无法在迁移作业中使用，您需要在DLI中新建SQL队列。	cdm
数据库名称	写入数据的数据库名称。	dli
表名	写入数据的表名。	car_detail
导入前清空数据	选择导入前是否清空目的表的数据。 如果设置为是，任务启动前会清除目标表中数据。	否

参数名	说明	取值样例
清空数据方式	导入前清空数据，如果设置为true时，呈现此参数。 TRUNCATE：删除标准数据。 INSERT_OVERWRITE：新增数据插入，同主键数据覆盖。	TRUNCATE
分区	“导入前清空数据”设置为“是”时，呈现此参数。 填写分区信息后，表示清空该分区的数据。	year=2020,location=sun

3.3.6.4.11 配置 OpenTSDB 目的端参数

作业中目的连接为**配置CloudTable OpenTSDB连接**时，目的端作业参数如**表3-85**所示。

表 3-85 OpenTSDB 作为目的端时的作业参数

参数名	说明	取值样例
指标	可选参数，输入指标名称，或选择OpenTSDB中已存在的指标。	city.temp
时间	可选参数，记录数据的时间点，格式为yyyyMMddHHmmdd的字符串或时间戳。	1598870800
标记	可选参数，可在这里自定义数据的标签。	tagk:tagv, tagk2:tagv2

3.3.6.5 配置定时任务

在表/文件迁移的任务中，CDM支持定时执行作业，按重复周期分为：分钟、小时、天、周、月。

说明

- CDM在配置定时作业时，不要为大量任务设定相同的定时时间，应该错峰调度，避免出现异常。
- 如果通过DataArts Studio数据开发调度CDM迁移作业，此处也配置了定时任务，则两种调度均会生效。为了业务运行逻辑统一和避免调度冲突，推荐您启用数据开发调度即可，无需配置CDM定时任务。

分钟

CDM支持配置每几分钟执行一次作业，定时任务周期不建议小于5分钟。

- 开始时间：表示定时配置生效的时间，也是第一次自动执行作业的时间。
- 重复周期（分）：从开始时间起，每多少分钟执行一次作业。
- 结束时间：该参数为可选参数，如果不配置则表示一直自动执行。如果配置了结束时间，则会在该时间停止自动执行作业。

小时

CDM支持配置每几小时执行一次作业。

- 重复周期（时）：表示每多少个小时自动执行一次定时任务。
- 触发时间（分）：表示每小时的第几分钟触发定时任务。该参数值取值范围是“0～59”，可配置多个值但不可重复，最多60个，中间使用“,”分隔。

  如果触发时间不在有效期内，则第一次自动执行的时间取有效期内最近的触发时间，例如：
  - 有效期的“开始时间”为“1:20”。
  - “重复周期（时）”为“3”。
  - “触发时间（分）”为“10”。
- 有效期：分为开始时间和结束时间。
  - 开始时间：表示定时配置生效的时间。
  - 结束时间：该参数是可选参数，表示停止自动执行的时间。如果不配置，则表示一直自动执行。

天

CDM支持配置每几天执行一次作业。

- 重复周期（天）：从开始时间起，每多少天执行一次作业。
- 有效期：分为开始时间和结束时间。
  - 开始时间：表示定时配置生效的时间，也是第一次自动执行作业的时间。
  - 结束时间：该参数是可选参数，表示停止自动执行的时间。如果不配置，则表示一直自动执行。

周

CDM支持配置每几周执行一次作业。

- 重复周期（周）：表示从开始时间起，每多少周执行一次定时任务。
- 触发时间（天）：选择每周几自动执行作业，可单选或多选。
- 有效期：分为开始时间和结束时间。
  - 开始时间：表示定时配置生效的时间。
  - 结束时间：该参数是可选参数，表示停止自动执行的时间。如果不配置，则表示一直自动执行。

月

CDM支持配置每几月执行一次作业。

- 重复周期（月）：从开始时间起，每多少个月自动执行定时任务。
- 触发时间（天）：选择每月的几号执行作业，该参数值取值范围是“1～31”，可配置多个值但不可重复，中间使用“,”分隔。
- 有效期：分为开始时间和结束时间。
  - 开始时间：表示定时配置生效的时间。其中的时、分、秒也是每次自动执行的时间。
  - 结束时间：该参数为可选参数，表示停止自动执行定时任务的时间。如果没有配置，则表示一直自动执行。

3.3.6.6 作业配置管理

CDM作业管理界面的“配置管理”页签，主要操作如下：

- **CDM作业最大抽取并发数**
- **CDM作业定时备份/恢复**
- **CDM作业参数的环境变量**

CDM 作业最大抽取并发数

最大抽取并发数取值范围为1-300，用于限制作业运行的总抽取并发数。如果当前所有作业总并发数超过限制，超过部分将排队等待。请您参考各单作业抽取并发数估算最大总抽取并发数。

单作业的抽取并发量配置原则如下：

CDM迁移作业的抽取并发数，与集群规格和表大小有关。并发抽取数取值范围为1-300，若配置过大，则以队列的形式进行排队。

建议每1CUs（1CUs=1核4G）配置为4，如**表3-86**所示，您也可以根据实际情况进行调整。另外，每行数据大小为1MB以下的可以多并发抽取，超过1MB的建议单线程抽取数据。

说明

- 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。
- 单作业的抽取并发数，受到作业“配置管理”中所配置的“最大抽取并发数”影响。“最大抽取并发数”配置的是抽取并发总数。

表 3-86 抽取并发数参考配置

CDM集群规格	vCPUs/内存	抽取并发数参考配置
cdm.large	8核 16GB	16
cdm.xlarge	16核 32GB	32
cdm.4xlarge	64核 128GB	128

CDM 作业定时备份/恢复

该功能依赖于OBS服务。

- 前提条件
  已创建**配置OBS连接**。
- 定时备份
  在CDM作业管理界面，单击“配置管理”页签，配置定时备份的参数。

表 3-87 定时备份参数

参数	说明	配置样例
定时备份	自动备份功能的开关，该功能只备份作业，不会备份连接。	开
备份策略	• 所有作业：不管作业处于什么状态，CDM会备份所有表/文件迁移作业、整库迁移的作业。不备份历史作业。 • 分组作业：选择备份某一个或多个分组下的作业。	所有作业
备份周期	选择备份周期： • 日：每天零点执行一次。 • 周：每周一零点执行一次。 • 月：每月1号零点执行一次。	日
备份写入OBS连接	CDM通过该连接，将作业备份到OBS，需要用户提前在“连接管理”界面创建好OBS连接。	obslink
OBS桶	存储备份文件的OBS桶。	cdm
备份数据目录	存储备份文件的目录。	/cdm-bk/

- 恢复作业

  如果之前执行过自动备份，“配置管理”页签下会显示备份列表：显示备份文件所在的OBS桶、路径、备份时间。

  您可以单击备份列表操作列的“恢复备份”来恢复CDM作业。

CDM 作业参数的环境变量

CDM在创建迁移作业时，可以手动输入的参数（例如OBS桶名、文件路径等）、参数中的某个字段、或者字段中的某个字符，都支持配置为一个全局变量，方便您批量更改作业中的参数值，以及作业导出/导入后进行批量替换。

这里以批量替换作业中OBS桶名为例进行介绍。

1. 在CDM作业管理界面，单击“配置管理”页签，配置环境变量。

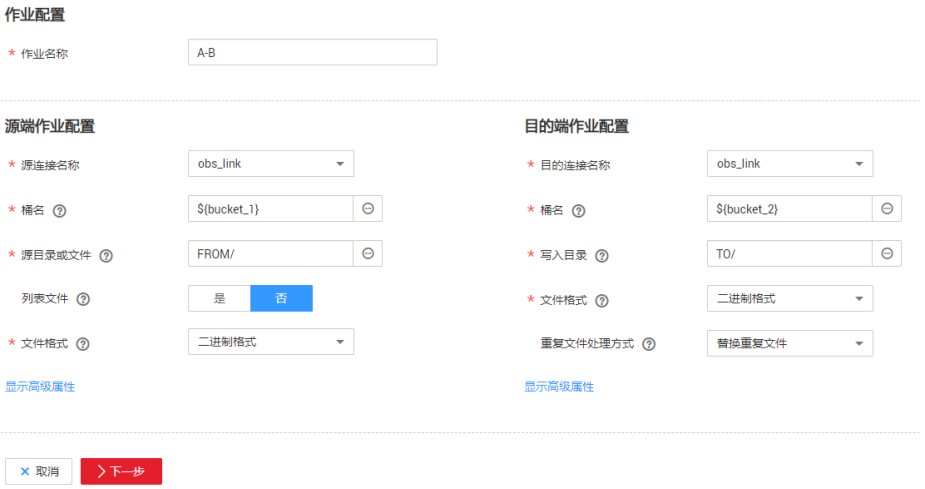
   ```
   bucket_1=A
   bucket_2=B
   ```

   这里以变量“bucket_1”表示桶A，变量“bucket_2”表示桶B。

2. 在创建CDM迁移作业的界面，迁移桶A的数据到桶B。

   源端桶名配置为**${bucket_1}**，目的端桶名配置为**${bucket_2}**。

图 3-59 桶名配置为环境变量



3. 如果下次要迁移桶C数据到桶D，则无需更改作业参数，只需要在“配置管理”界面将环境变量改为如下即可：

```
bucket_1=C
bucket_2=D
```

3.3.6.7 管理单个作业

已存在的CDM作业支持查看、修改、删除、启动、停止等操作，这里主要介绍作业的查看和修改。

查看

- **查看作业状态**

  作业状态有New，Pending，Booting，Running，Failed，Succeeded。

  其中“Pending”表示正在等待系统调度该作业，“Booting”表示正在分析待迁移的数据。

- **查看历史记录**

  查看作业的历史执行记录、读取和写入的统计数据，在历史记录界面还可查看作业执行的日志信息。

- **查看作业日志**

  在历史记录界面可查看作业所有的日志。

  也可以在作业列表界面，选择“更多 > 日志”来查看该作业最近的一次日志。

- **查看作业JSON**

  直接编辑作业的JSON文件，作用等同于修改作业的参数配置

- **源目的统计查询**

  可对已经配置好的数据库类作业打开预览窗口，预览最多1000条数据内容。可对比源和目的端的数据，也可以通过对比记录数来看迁移结果是否成功、数据是否丢失。

- **查看历史作业**

  CDM可以保留最近1个月已执行的作业，包括一次性作业（运行完自动删除的作业）和周期重复执行的作业，都支持在“历史作业”页签下查看、重新执行。

对于周期重复执行的作业，每次执行时（无论成功失败）都会在“历史作业”的页签下生成一个历史作业，执行了多少次便生成多少个历史作业。由于原作业名相同，所以历史作业的作业名会随机增加一个字符串以做区分。

修改

- **修改作业参数**

  可重新配置作业参数，但是不能重新选择源连接和目的连接。

- **编辑作业JSON**

  直接编辑作业的JSON文件，作用等同于修改作业的参数配置。

操作步骤

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择集群后的“作业管理”。

步骤2 单击“历史作业”可以查看最近1个月所有执行过的历史作业。

CDM可以保留最近1个月已执行的作业，包括一次性作业（运行完自动删除的作业）和周期重复执行的作业，都支持在“历史作业”页签下查看、重新执行。

对于周期重复执行的作业，每次执行时（无论成功失败）都会在“历史作业”的页签下生成一个历史作业，执行了多少次便生成多少个历史作业。由于原作业名相同，所以历史作业的作业名会随机增加一个字符串以做区分。

步骤3 单击“表/文件迁移”显示作业列表，可对单个作业执行如下操作：

- 修改作业参数：单击作业操作列的“编辑”可修改作业参数。
- 运行作业：单击作业操作列的“运行”可手动启动作业。
- 查看历史记录：单击作业操作列的“历史记录”进入历史记录界面，可查看该作业的历史执行记录、读取和写入的统计数据。在历史记录界面单击“日志”，可查看作业执行的日志信息。
- 删除作业：选择作业操作列的“更多 > 删除”可删除作业。
- 停止作业：选择作业操作列的“更多 > 停止”可停止作业。
- 查看作业JSON：选择作业操作列的“更多 > 查看作业JSON”，可查看该作业的JSON定义。
- 编辑作业JSON：选择作业操作列的“更多 > 编辑作业JSON”，可直接编辑该作业的JSON文件，作用等同于修改作业的参数配置。
- 配置定时任务：选择作业操作列的“更多 > 配置定时任务”，可选择在有效期内周期性启动作业，具体请参考**配置定时任务**。

步骤4 修改完成后单击“保存”或“保存并运行”。

----结束

3.3.6.8 批量管理作业

操作场景

这里以表/文件迁移的作业为例进行介绍，指导用户批量管理CDM作业，提供以下操作：

- 作业分组管理

- 批量运行作业
- 批量删除作业
- 批量导出作业
- 批量导入作业

批量导出、导入作业的功能，适用以下场景：

- CDM集群间作业迁移：例如需要将作业从老版本集群迁移到新版本的集群。
- 备份作业：例如需要将CDM集群停掉或删除来降低成本时，可以先通过批量导出把作业脚本保存下来，仅在需要的时候再重新创建集群和重新导入作业。
- 批量创建作业任务：可以先手工创建一个作业，导出作业配置（导出的文件为JSON格式），然后参考该作业配置，在JSON文件中批量复制出更多作业，最后导入CDM以实现批量创建作业。

操作步骤

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择集群后的“作业管理”。

步骤2 单击“表/文件迁移”显示作业列表，提供以下批量操作：

- **作业分组**

  CDM支持对分组进行新增、修改、查找、删除。删除分组时，会将组内的所有作业都删除。

  创建作业的第三步任务配置中，如果已经将作业分配到了不同的分组中，则这里可以按分组显示作业、按组批量启动作业、按分组导出作业等操作。

- **批量运行作业**

  勾选一个或多个作业后，单击“运行”可批量启动作业。

- **批量删除作业**

  勾选一个或多个作业后，单击“删除”可批量删除作业。

- **批量导出作业**

  单击“导出”，弹出批量导出页面，如图3-60。

  图 3-60 批量导出页面

- 全部作业和连接：勾选此项表示一次性导出所有作业和连接。
- 全部作业：勾选此项表示一次性导出所有作业。
- 全部连接：勾选此项表示一次性导出所有连接。
- 按作业名导出：勾选此项并选择需要导出的作业，单击确认即可导出所选作业。
- 按分组导出：勾选此项并下拉选择需要导出的分组，单击确认即可导出所选分组。

批量导出可将需要导出的作业导出保存为JSON文件，用于备份或导入到别的集群中。

说明

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。

- **批量导入作业**

  单击“导入”，选择JSON格式的文件导入或文本导入。

  - 文件导入：待导入的作业文件必须为JSON格式（大小不超过1M）。如果待导入的作业文件是之前从CDM中导出的，则导入前必须先编辑JSON文件，将“Add password here”替换为对应连接的正确密码，再执行导入操作。
  - 文本导入：无法正确上传本地JSON文件时可选择该方式。将作业的JSON文本直接粘贴到输入框即可。

----结束

3.3.7 审计

3.3.7.1 支持云审计的关键操作

云审计服务（Cloud Trace Service，简称CTS）为用户提供了云账户下资源的操作记录，可以帮您记录云数据迁移相关的操作事件，便于日后的查询、审计和回溯。

表 3-88 云审计服务支持的 CDM 操作列表

操作名称	资源类型	事件名称
创建集群	cluster	createCluster
删除集群	cluster	deleteCluster
修改集群配置	cluster	modifyCluster
开机	cluster	startCluster
重启	cluster	startStopCluster
导入作业	cluster	clusterImportJob
绑定弹性IP	cluster	bindEip
解绑弹性IP	cluster	unbindEip
创建连接	link	createLink

操作名称	资源类型	事件名称
修改连接	link	modifyLink
删除连接	link	deleteLink
创建任务	job	createJob
修改任务	job	modifyJob
删除任务	job	deleteJob
启动任务	job	startJob
停止任务	job	stopJob

3.3.7.2 如何查看审计日志

操作场景

在您开启了云审计服务后，系统开始记录CDM的相关操作，云审计服务的管理控制台保存最近7天的操作记录。

本节介绍如何在云审计服务管理控制台查看最近7天的操作记录。

操作步骤

1. 登录管理控制台。
2. 单击“服务列表”，选择“管理与部署 > 云审计服务”，进入云审计服务信息页面。
3. 单击左侧导航树的“事件列表”，进入事件列表信息页面。

   事件列表支持通过筛选来查询对应的操作事件。
4. 在需要查看的事件左侧，单击事件名称左边的箭头，展开该记录的详细信息。
5. 在需要查看的记录右侧，单击“查看事件”，弹窗中显示了该操作事件结构的详细信息。

   更多关于云审计的信息，请参见《云审计服务用户指南》。

3.3.8 使用教程

3.3.8.1 创建 MRS Hive 连接器

MRS Hive连接适用于MapReduce服务，本教程为您介绍如何创建MRS Hive连接器。

前提条件

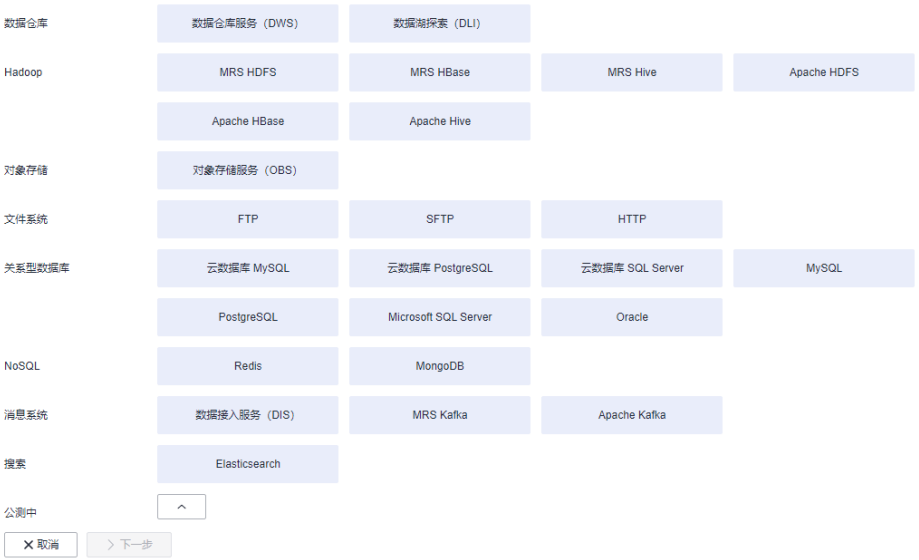
- 已创建CDM集群。
- 已获取MRS集群的Manager IP、管理员帐号和密码，且该帐号拥有数据导入、导出的操作权限。
- MRS集群和CDM集群之间网络互通，网络互通需满足如下条件：

- CDM集群与云上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。通过公网互通时，需确保CDM集群已绑定EIP，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。
- CDM集群与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC) 使用指南》中的“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC) 使用指南》中的“安全组 > 添加安全组规则”章节。
- 此外，您还必须确保该云服务的实例与CDM集群所属的企业项目必须相同，如果不同，需要修改工作空间的企业项目。

新建 MRS hive 连接

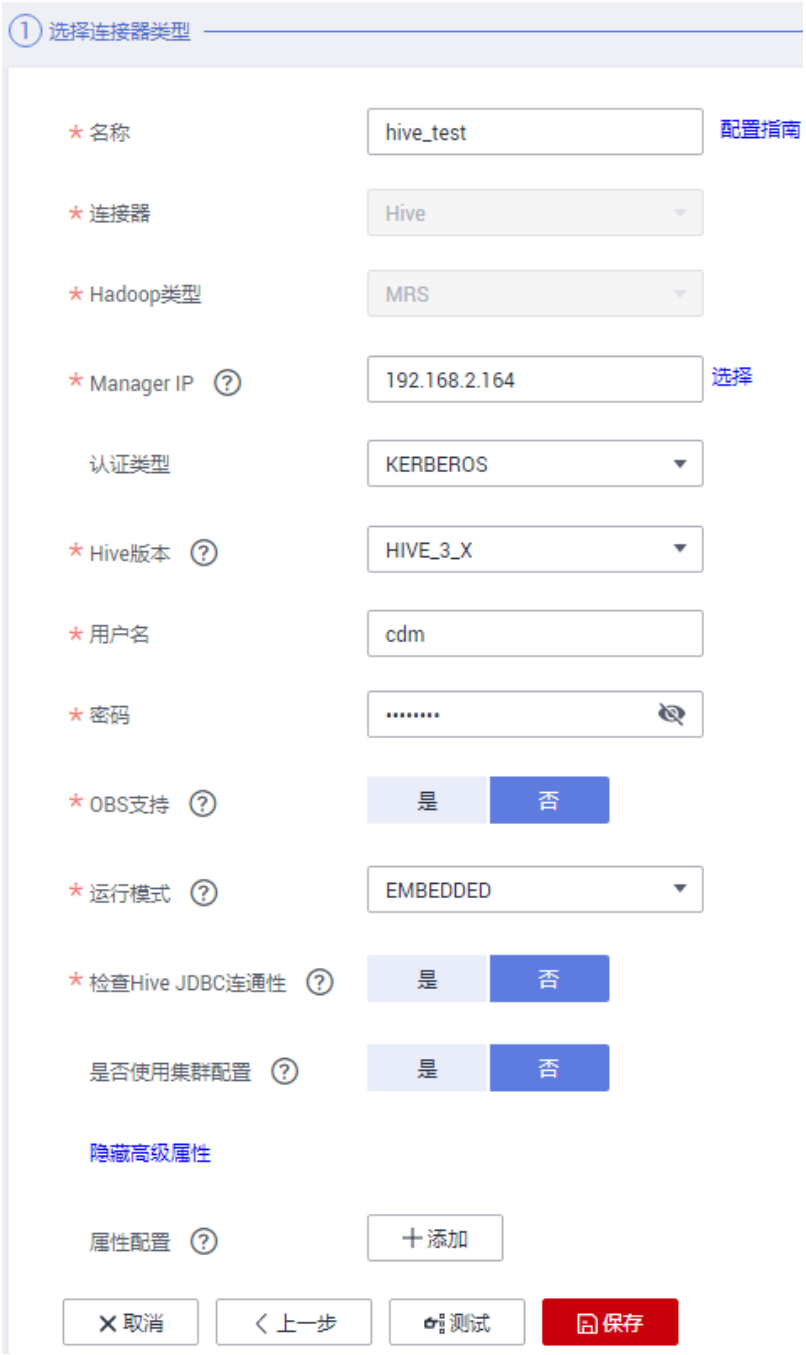
步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如图3-61所示。

图 3-61 选择连接器类型



步骤2 连接器类型选择“MRS Hive”后单击“下一步”，配置MRS Hive连接的参数，如图3-62所示。

图 3-62 创建 MRS Hive 连接



步骤3 单击“显示高级属性”可查看更多可选参数，具体请参见5.6-配置关系数据库连接。这里保持默认，必填参数如**表3-89**所示。

表 3-89 MRS Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mrs-link

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 ● 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。	cdm
密码	访问MRS Manager的用户密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
检查Hive JDBC连通性	是否需要测试Hive JDBC连通性。	否
是否使用集群配置	用户可以在“连接管理”处创建集群配置，用于简化Hadoop连接参数配置。	否
属性配置	其他Hive客户端配置属性。	-

说明

单击“显示高级属性”，然后单击“添加”，您可以添加客户端的配置属性。所添加的每个属性需配置属性名称和值。对于不再需要的属性，可单击属性后的“删除”按钮进行删除。

步骤4 单击“保存”回到连接管理界面，完成MRS Hive连接器的配置。

----结束

3.3.8.2 创建 MySQL 连接器

MySQL连接适用于第三方云MySQL服务，以及用户在本地数据中心或ECS上自建的MySQL。本教程为您介绍如何创建MySQL连接器。

前提条件

- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 本地MySQL数据库可通过公网访问。如果MySQL服务器是在本地数据中心或第三方云上，需要确保MySQL可以通过公网IP访问，或者是已经建立好了企业内部数据中心到云服务平台的VPN通道或专线。
- 已创建CDM集群。

新建 MySQL 连接器

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，选择CDM集群后的“作业管理 > 连接管理 > 驱动管理”，进入驱动管理页面。

图 3-63 上传驱动

步骤2 单击“驱动管理”页面左上角“驱动下载地址”链接下载MySQL的驱动，详情请参见**如何获取驱动**。

步骤3 在“驱动管理”页面中，选择以下方式上传MySQL驱动。

方式一：单击对应驱动名称右侧操作列的“上传”，选择本地已下载的驱动。

方式二：单击对应驱动名称右侧操作列的“从sftp复制”，配置sftp连接器名称和驱动文件路径。

步骤4 在“集群管理”界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图3-64**所示。

图 3-64 选择连接器类型

步骤5 连接器类型选择“MySQL”后单击“下一步”，配置MySQL连接的参数，参数如**表3-90**所示。

图 3-65 创建 MySQL 连接

表 3-90 MySQL 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	mysqllink
数据库服务器	MySQL数据库的IP地址或域名。	192.168.1.110
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop

参数名	说明	取值样例
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
单次请求行数	指定每次请求获取的行数。	1000
单次提交行数	支持通过agent从源端提取数据	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤6 单击“保存”回到连接管理界面，完成MySQL连接器的配置。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

3.3.8.3 MySQL 数据迁移到 MRS Hive 分区表

MapReduce服务（MapReduce Service，简称MRS）提供企业级大数据集群云服务，里面包含HDFS、Hive、Spark等组件，适用于企业海量数据分析。

其中Hive提供类SQL查询语言，帮助用户对大规模的数据进行提取、转换和加载，即通常所称的ETL（Extraction，Transformation，and Loading）操作。对庞大的数据集查询需要耗费大量的时间去处理，在许多场景下，可以通过建立Hive分区方法减少每一次扫描的总数据量，这种做法可以显著地改善性能。

Hive的分区使用HDFS的子目录功能实现，每一个子目录包含了分区对应的列名和每一列的值。当分区很多时，会有很多HDFS子目录，如果不依赖工具，将外部数据加载到Hive表各分区不是一件容易的事情。云数据迁移服务（CDM）可以请轻松将外部数据源（关系数据库、对象存储服务、文件系统服务等）加载到Hive分区表。

下面使用CDM将MySQL数据导入到MRS Hive分区表为例进行介绍。

操作场景

假设MySQL上有一张表trip_data，保存了自行车骑行记录，里面有起始时间、结束时间，起始站点、结束站点、骑手ID等信息，trip_data表字段定义如图3-66所示。

图 3-66 MySQL 表字段

Column Name	#	Data Type
TripID	1	int(11)
Duration	2	int(11)
StartDate	3	timestamp
StartStation	4	varchar(64)
StartTerminal	5	int(11)
EndDate	6	timestamp
EndStation	7	varchar(64)
EndTerminal	8	int(11)
Bike	9	int(11)
SubscriberType	10	varchar(32)
ZipCodev	11	varchar(10)

使用CDM将MySQL中的表trip_data导入到MRS Hive分区表，流程如下：

1. **在MRS Hive上创建Hive分区表**
2. **创建CDM集群并绑定EIP**
3. **创建MySQL连接**
4. **创建Hive连接**
5. **创建迁移作业**

前提条件

- 已经创建MRS。
- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 已参考**管理驱动**，上传了MySQL数据库驱动。

在 MRS Hive 上创建 Hive 分区表

在MRS的Hive上使用下面SQL语句创建一张Hive分区表，表名与MySQL上的表trip_data一致，且Hive表比MySQL表多建三个字段y、ym、ymd，作为Hive的分区字段。SQL语句如下：

```
create table trip_data(TripID int,Duration int,StartDate timestamp,StartStation varchar(64),StartTerminal int,EndDate timestamp,EndStation varchar(64),EndTerminal int,Bike int,SubscriberType varchar(32),ZipCodev varchar(10))partitioned by (y int,ym int,ymd int);
```

说明

Hive表trip_data有三个分区字段：骑行起始时间的年、骑行起始时间的年月、骑行起始时间的年月日，例如一条骑行记录的起始时间为2018/5/11 9:40，那么这条记录会保存在分区trip_data/2018/201805/20180511下面。对trip_data进行按时间维度统计汇总时，只需要对局部数据扫描，大大提升性能。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群所在VPC、子网、安全组，选择与MRS集群所在的网络一致。

步骤2 CDM集群创建完成后，选择集群操作列的“绑定弹性IP”，CDM通过EIP访问MySQL。

图 3-67 集群列表

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置常见关系数据库连接**。这里保持默认，必填参数如**表3-91**所示。

表 3-91 MySQL 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	mysqllink
数据库服务器	MySQL数据库的IP地址或域名。	192.168.1.110
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin

参数名	说明	取值样例
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
单次请求行数	指定每次请求获取的行数。	1000
单次提交行数	支持通过agent从源端提取数据	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

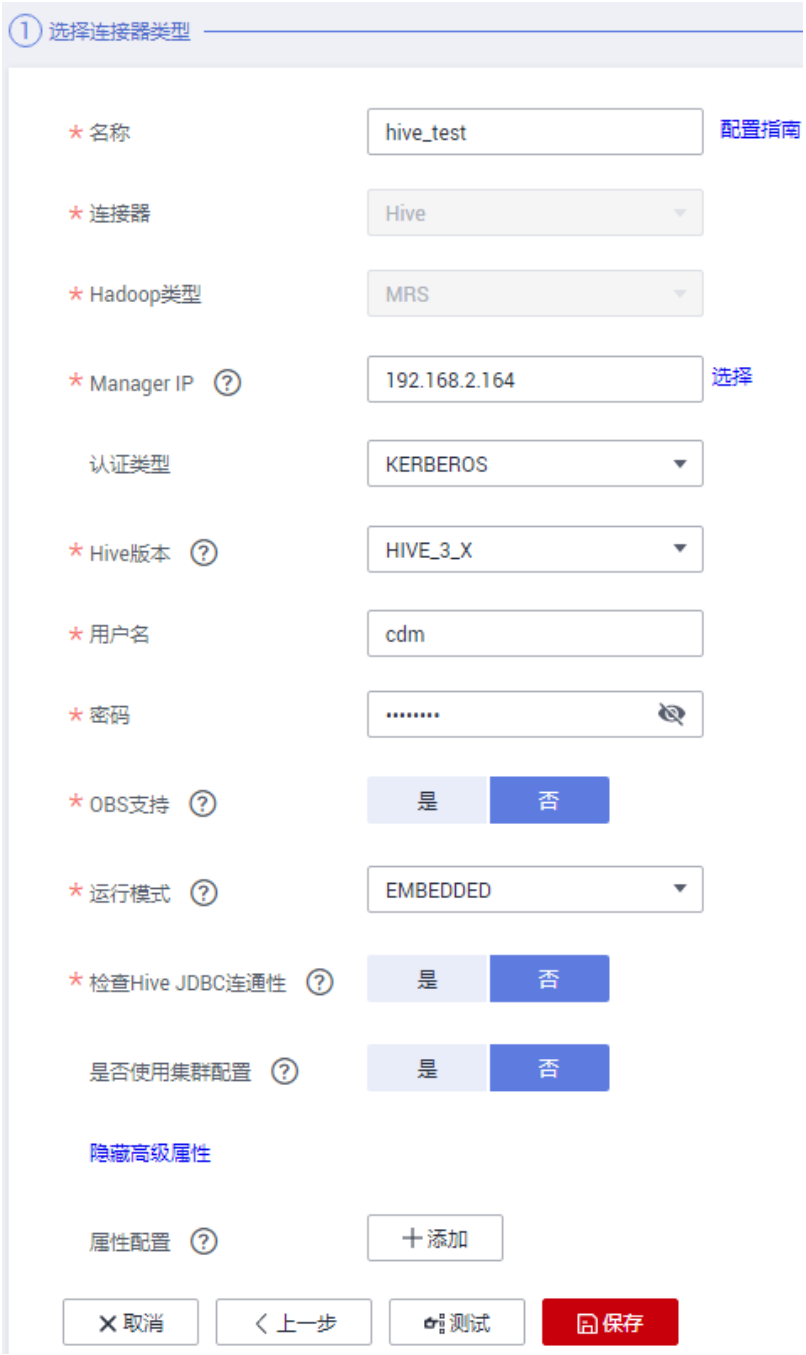
----结束

创建 Hive 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“MRS Hive”后单击“下一步”配置Hive连接参数，如**图3-68**所示。

图 3-68 创建 MRS Hive 连接



各参数说明如表3-92所示，需要您根据实际情况配置。

表 3-92 MRS Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 ● 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。	cdm
密码	访问MRS Manager的用户密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： • EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 • STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
检查Hive JDBC连通性	是否需要测试Hive JDBC连通。	否
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hive_01

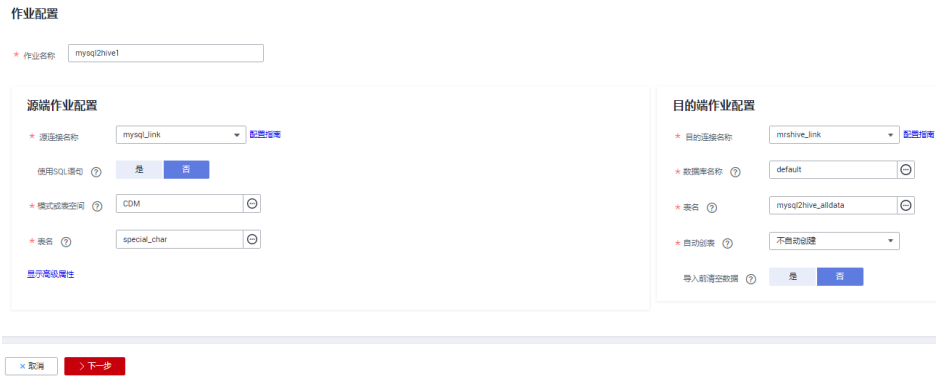
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建数据迁移任务，如**图3-69**所示。

图 3-69 创建 MySQL 到 Hive 的迁移任务



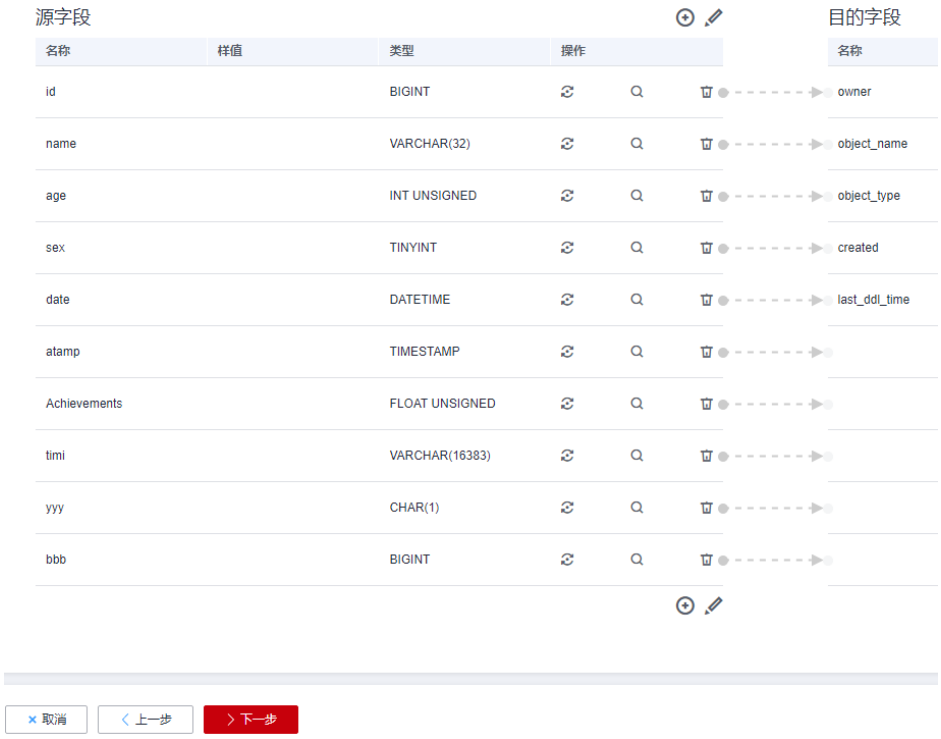
说明

“导入前清空数据”选“是”，这样每次导入前，会将之前已经导入到Hive表的数据清空。

步骤2 作业参数配置完成后，单击“下一步”，进入字段映射界面，如图3-70所示。

映射MySQL表和Hive表字段，Hive表比MySQL表多三个字段y、ym、ymd，即是Hive的分区字段。由于没有源表字段直接对应，需要配置表达式从源表的StartDate字段抽取。

图 3-70 Hive 字段映射



步骤3 单击 进入转换器列表界面，再选择“新建转换器 > 表达式转换”，如图3-71所示。

y、ym、ymd字段的表达式分别配置如下：

DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyy")

DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyyMM")

DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyyMMdd")

图 3-71 配置表达式

说明

CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换。

步骤4 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。这里保持默认值“1”。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤5 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤6 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.4 MySQL 数据迁移到 OBS

操作场景

CDM支持表到OBS的迁移，本章节以MySQL-->OBS为例，介绍如何通过CDM将表数据迁移到OBS中。流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建MySQL连接**
3. **创建OBS连接**
4. **创建迁移作业**

前提条件

- 已获取OBS的访问域名、端口，以及AK、SK。
- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 用户已参考**管理驱动**，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。

步骤2 CDM集群创建完成后，选择集群操作列的“绑定弹性IP”，CDM通过EIP访问MySQL。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置常见关系数据库连接**。这里保持默认，必填参数如**表3-93**所示。

表 3-93 MySQL 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	mysqllink
数据库服务器	MySQL数据库的IP地址或域名。	192.168.1.110

参数名	说明	取值样例
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
单次请求行数	指定每次请求获取的行数。	1000
单次提交行数	支持通过agent从源端提取数据	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

图 3-72 选择连接器类型

步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数。

- 名称：用户自定义连接名称，例如“obslink”。
- OBS服务器、端口：配置为OBS实际的地址信息。
- 访问标识（AK）、密钥（SK）：登录OBS的AK、SK。

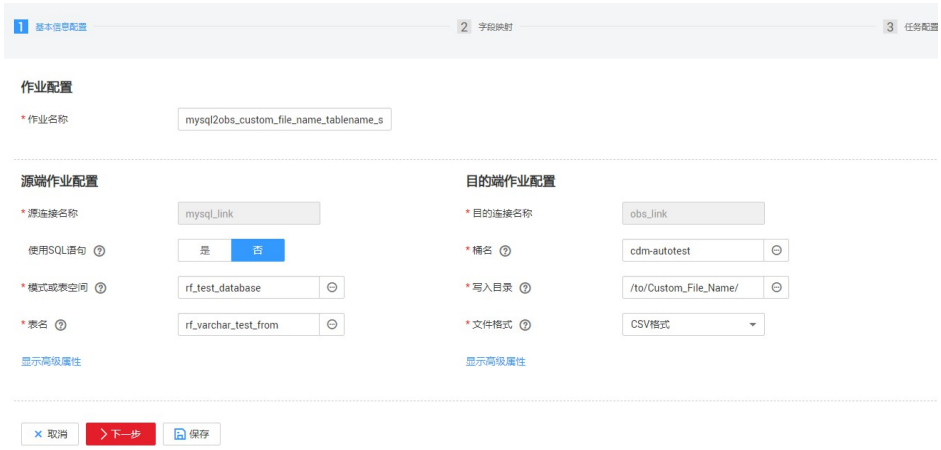
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从MySQL导出数据到OBS的任务。

图 3-73 创建 MySQL 到 OBS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建MySQL连接**中的“mysqllink”。

  - 使用SQL语句：否。
  - 模式或表空间：待抽取数据的模式或表空间名称。
  - 表名：要抽取的表名。
  - 其他可选参数一般情况下保持默认即可，详细说明请参见**配置常见关系数据库源端参数**。
- 目的端作业配置
  - 目的连接名称：选择**创建OBS连接**中的“obslink”。
  - 桶名：待迁移数据的桶。
  - 写入目录：写入数据到OBS服务器的目录。
  - 文件格式：迁移数据表到文件时，文件格式选择“CSV格式”。
  - 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置OBS目的端参数**。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图3-74**所示。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换。

图 3-74 表到文件的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。CDM支持并发抽取MySQL数据，如果源表配置了索引，可调大抽取并发数提升迁移速率。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要先配置好OBS连接。针对文件到表类迁移的数据，建议配置写入脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。根据使用场景，也可配置为“删除”，防止迁移作业堆积。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.5 MySQL 数据迁移到 DWS

操作场景

CDM支持表到表的迁移，本章节以MySQL-->DWS为例，介绍如何通过CDM将表数据迁移到表中。流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建MySQL连接**
3. **创建DWS连接**
4. **创建迁移作业**

前提条件

- 已获取DWS数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有DWS数据库的读、写和删除权限。
- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 用户已参考**管理驱动**，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群所在VPC、子网、安全组，选择与DWS集群所在的网络一致。

步骤2 CDM集群创建完成后，选择集群操作列的“绑定弹性IP”，CDM通过EIP访问MySQL。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置常见关系数据库连接**。这里保持默认，必填参数如**表3-94**所示。

表 3-94 MySQL 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	mysqllink
数据库服务器	MySQL数据库的IP地址或域名。	192.168.1.110
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
单次请求行数	指定每次请求获取的行数。	1000
单次提交行数	支持通过agent从源端提取数据	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 DWS 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 连接器类型选择“数据仓库服务（DWS）”后单击“下一步”配置DWS连接参数，必填参数如**表3-95**所示，可选参数保持默认即可。

表 3-95 DWS 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	dwslink
数据库服务器	DWS数据库的IP地址或域名。	192.168.0.3
端口	DWS数据库的端口。	8000
数据库名称	DWS数据库的名称。	db_demo
用户名	拥有DWS数据库的读、写和删除权限的用户。	dbadmin
密码	用户的密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
导入模式	COPY模式：将源数据经过DWS管理节点后拷贝到数据节点。如果需要通过Internet访问DWS，只能使用COPY模式。	COPY

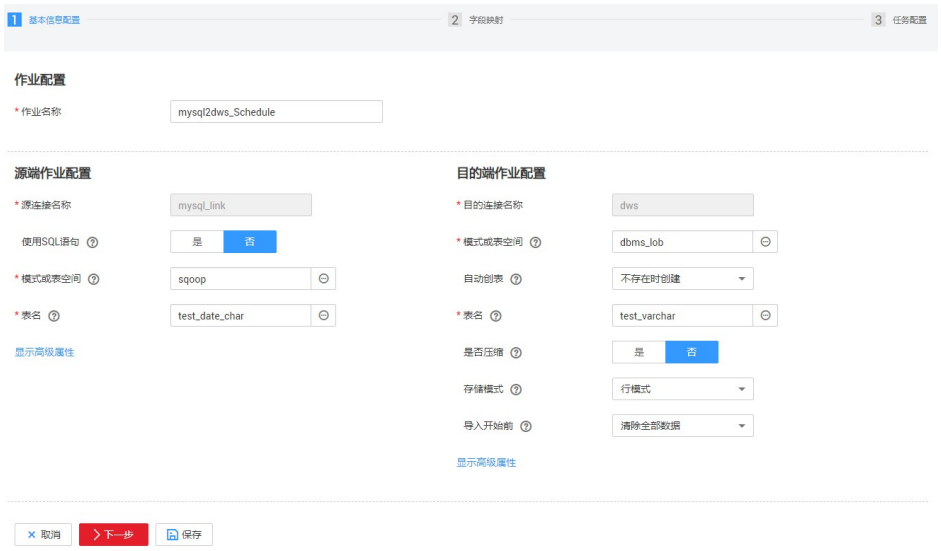
步骤3 单击“保存”完成创建连接。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从MySQL导出数据到DWS的任务。

图 3-75 创建 MySQL 到 DWS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建MySQL连接**中的“mysqllink”。
  - 使用SQL语句：否。
  - 模式或表空间：待抽取数据的模式或表空间名称。
  - 表名：要抽取的表名。
  - 其他可选参数一般情况下保持默认即可，详细说明请参见**配置常见关系数据库源端参数**。
- 目的端作业配置
  - 目的连接名称：选择**创建DWS连接**中的连接“dwslink”。
  - 模式或表空间：选择待写入数据的DWS数据库。
  - 自动创表：只有当源端和目的端都为关系数据库时，才有该参数。
  - 表名：待写入数据的表名，可以手动输入一个不存在表名，CDM会在DWS中自动创建该表。
  - 是否压缩：DWS提供的压缩数据能力，如果选择“是”，将进行高级别压缩，CDM提供了适用I/O读写量大，CPU富足（计算相对小）的压缩场景
  - 存储模式：可以根据具体应用场景，建表的时候选择行存储还是列存储表。一般情况下，如果表的字段比较多（大宽表），查询中涉及到的列不多的情况下，适合列存储。如果表的字段个数比较少，查询大部分字段，那么选择行存储比较好。
  - 扩大字符字段长度：当目的端和源端数据编码格式不一样时，自动建表的字符字段长度可能不够用，配置此选项后CDM自动建表时会将字符字段扩大3倍。
  - 导入前清空数据：任务启动前，是否清除目的表中数据，用户可根据实际需要选择。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图3-76**所示。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。

- 单击，可批量映射字段。
- CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换。

图 3-76 表到表的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。可适当调大参数，提升迁移效率。
- 是否写入脏数据：表到表的迁移容易出现脏数据，建议配置脏数据归档。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.6 MySQL 整库迁移到 RDS 服务

操作场景

本章节介绍使用CDM整库迁移功能，将本地MySQL数据库迁移到云服务RDS中。

当前CDM支持将本地MySQL数据库，整库迁移到RDS上的MySQL、PostgreSQL或者Microsoft SQL Server任意一种数据库中。这里以整库迁移到RDS上的MySQL数据库为例进行介绍，使用流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建MySQL连接**
3. **创建RDS连接**

4. 创建整库迁移作业

前提条件

- 用户拥有EIP配额。
- 用户已创建RDS数据库实例，该实例的数据库引擎为MySQL。
- 本地MySQL数据库可通过公网访问。如果MySQL服务器是在本地数据中心或第三方云上，需要确保MySQL可以通过公网IP访问，或者是已经建立好了企业内部数据中心到云服务平台的VPN通道或专线。
- 已获取本地MySQL数据库和RDS上MySQL数据库的IP地址、数据库名称、用户名和密码。
- 用户已参考**管理驱动**，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC，选择和RDS的MySQL数据库实例所在的VPC一致，且推荐子网、安全组也与RDS上的MySQL一致。
- 如果安全控制原因不能使用相同子网和安全组，则可以修改安全组规则，允许CDM访问RDS。

步骤2 CDM集群创建完成后，选择集群操作列的“绑定弹性IP”，CDM通过EIP访问本地MySQL数据库。

图 3-77 集群列表

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置常见关系数据库连接**。这里保持默认，必填参数如**表3-96**所示。

表 3-96 MySQL 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	mysqllink
数据库服务器	MySQL数据库的IP地址或域名。	192.168.1.110
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
单次请求行数	指定每次请求获取的行数。	1000
单次提交行数	支持通过agent从源端提取数据	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 RDS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“云数据库 MySQL”后单击“下一步”，配置连接参数：

- 名称：用户自定义连接名称，例如：“rds_link”。
- 数据库服务器、端口：配置为RDS上MySQL数据库的连接地址、端口。
- 数据库名称：配置为RDS上MySQL数据库的名称。
- 用户名、密码：登录数据库的用户和密码。

说明

- 创建RDS连接时，“使用本地API”设置为“是”时，可以使用MySQL的LOAD DATA功能加快数据导入，提高导入数据到MySQL的性能。
- 由于RDS上的MySQL默认没有开启LOAD DATA功能，所以同时需要修改MySQL实例的参数组，将“local_infile”设置为“ON”，开启该功能。
- 如果“local_infile”参数组不可编辑，则说明是默认参数组，需要先创建一个新的参数组，再修改该参数值，并应用到RDS的MySQL实例上。

步骤3 单击“保存”回到连接管理界面。

----结束

创建整库迁移作业

步骤1 两个连接创建完成后，选择“整库迁移 > 新建作业”，开始创建迁移任务，如图3-78所示。

图 3-78 创建整库迁移作业

- 作业名称：用户自定义整库迁移的任务名称。
- 源端作业配置
  - 源连接名称：选择创建MySQL连接中的“mysql_link”。
  - 模式或表空间：选择从本地MySQL的哪个数据库导出数据。

- 目的端作业配置
  - 目的连接名称：选择**创建RDS连接**中的“rds_link”。
  - 模式或表空间：选择将数据导入到RDS的哪个数据库。
  - 自动创表：选择“不存在时创建”，当RDS数据库中没有本地MySQL数据库里的表时，CDM会自动在RDS数据库中创建那些表。
  - 导入前清空数据：选择“是”，当RDS数据库中存在与本地MySQL数据库重名的表时，CDM会清除RDS中重名表里的数据。
  - 高级属性里的可选参数保持默认即可。

步骤2 单击“下一步”，进入选择待迁移表的界面，您可以选择全部或者部分表进行迁移。

步骤3 单击“保存并运行”，CDM会立即开始执行整库迁移任务。

作业任务启动后，每个待迁移的表都会生成一个子任务，单击整库迁移的作业名称，可查看子任务列表。

步骤4 单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

整库迁移的作业没有日志，子作业才有。在子作业的历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.7 Oracle 数据迁移到云搜索服务

操作场景

云搜索服务（Cloud Search Service）为用户提供结构化、非结构化文本的多条件检索、统计、报表，本章节介绍如何通过CDM将数据从Oracle迁移到云搜索服务中，流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建云搜索服务连接**
3. **创建Oracle连接**
4. **创建迁移作业**

前提条件

- 已经开通了云搜索服务，且获取云搜索服务集群的IP地址和端口。
- 已获取Oracle数据库的IP、数据库名、用户名和密码。
- 如果Oracle数据库是在本地数据中心或第三方云上，需要确保Oracle可通过公网IP访问，或者已经建立好了企业内部数据中心到的VPN通道或专线。
- 用户已参考**管理驱动**，上传了Oracle数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。

- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

步骤2 CDM集群创建完成后，在集群管理界面选择“绑定弹性IP”，CDM通过EIP访问Oracle数据源。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“云搜索服务”后单击“下一步”，配置云搜索服务连接参数。

- 名称：用户自定义连接名称，例如“csslink”。
- Elasticsearch服务器列表：配置为云搜索服务集群（支持5.X以上版本）的连接地址、端口，格式为“ip:port”，多个地址之间使用分号（；）分隔，例如192.168.0.1:9200;192.168.0.2:9200。
- 用户名、密码：配置为访问云搜索服务集群的用户，需要拥有数据库的读写权限。

步骤3 单击“保存”回到连接管理界面。

----结束

创建 Oracle 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“Oracle”后单击“下一步”，配置Oracle连接参数：

- 名称：用户自定义连接名称，例如“oracle_link”。
- 数据库服务器地址、端口：配置为Oracle服务器的地址、端口。
- 数据库名称：选择要导出数据的Oracle数据库名称。
- 用户名、密码：Oracle数据库的登录用户名和密码，该用户需要拥有Oracle元数据的读取权限。

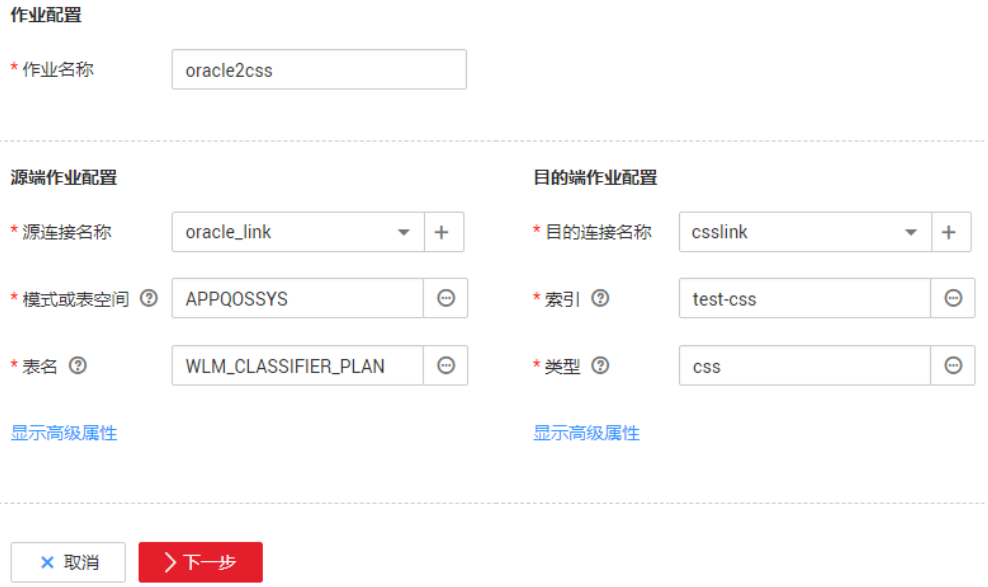
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从Oracle导出数据到云搜索服务的任务。

图 3-79 创建 Oracle 到云搜索服务的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建Oracle连接**中的“oracle_link”。
  - 模式或表空间：待迁移数据的数据库名称。
  - 表名：待迁移数据的表名。
  - 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置常见关系数据库源端参数**。
- 目的端作业配置
  - 目的连接名称：选择**创建云搜索服务连接**中的“csslink”。
  - 索引：待写入数据的Elasticsearch索引，也可以输入一个新的索引，CDM会自动在云搜索服务中创建。
  - 类型：待写入数据的Elasticsearch类型，可输入新的类型，CDM支持在目的端自动创建类型。
  - 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置云搜索服务目的端参数**。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图3-80**所示。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- 如果选择在目的端自动创建类型，这里还需要配置每个类型的字段类型、字段名称。
- CDM支持迁移过程中转换字段内容。

图 3-80 云搜索服务的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。这里保持默认值“1”。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.8 Oracle 数据迁移到 DWS

操作场景

CDM支持表到表的迁移，本章节介绍如何通过CDM将数据从Oracle迁移到数据仓库服务（Data Warehouse Service，简称DWS）中，流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建Oracle连接**
3. **创建DWS连接**
4. **创建迁移作业**

前提条件

- 已创建DWS集群，并且已获取DWS数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有DWS数据库的读、写和删除权限。

- 已获取Oracle数据库的IP、数据库名、用户名和密码。
- 如果Oracle数据库是在本地数据中心或第三方云上，需要确保Oracle可通过公网IP访问，或者已经建立好了企业内部数据中心到云的VPN通道或专线。
- 用户已参考**管理驱动**，上传了Oracle数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群所在VPC、子网、安全组，选择与DWS集群所在的网络一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

步骤2 CDM集群创建完成后，在集群管理界面选择“绑定弹性IP”，CDM通过EIP访问Oracle数据源。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建 Oracle 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

图 3-81 选择连接器类型

步骤2 连接器类型选择“Oracle”后单击“下一步”，配置Oracle连接参数，参数说明如**表3-97**所示。

图 3-82 创建 Oracle 连接

表 3-97 Oracle 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	oracle_link
数据库服务器	数据库服务器域名或IP地址。	192.168.0.1
端口	Oracle数据库的端口。	3306
数据库连接类型	Oracle数据库连接类型。	Service Name
数据库名称	要连接的数据库。	db_user
用户名	拥有Oracle数据库的读取权限的用户。	admin
密码	Oracle数据库的登录密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**连接Agent**中已创建的Agent。	-
ORACLE版本	默认使用最新版本驱动，若不兼容请尝试其他版本。	高于12.1
驱动版本	需要适配的驱动。	-
一次请求行数	指定每次请求获取的行数。	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'

步骤3 单击“保存”回到连接管理界面。

----结束

创建 DWS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“数据仓库服务（DWS）”后单击“下一步”配置DWS连接参数，必填参数如**表3-98**所示，可选参数保持默认即可。

表 3-98 DWS 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	dwslink
数据库服务器	DWS数据库的IP地址或域名。	192.168.0.3

参数名	说明	取值样例
端口	DWS数据库的端口。	8000
数据库名称	DWS数据库的名称。	db_demo
用户名	拥有DWS数据库的读、写和删除权限的用户。	dbadmin
密码	用户的密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择连接Agent中已创建的Agent。	-
导入模式	COPY模式：将源数据经过DWS管理节点后拷贝到数据节点。如果需要通过Internet访问DWS，只能使用COPY模式。	COPY

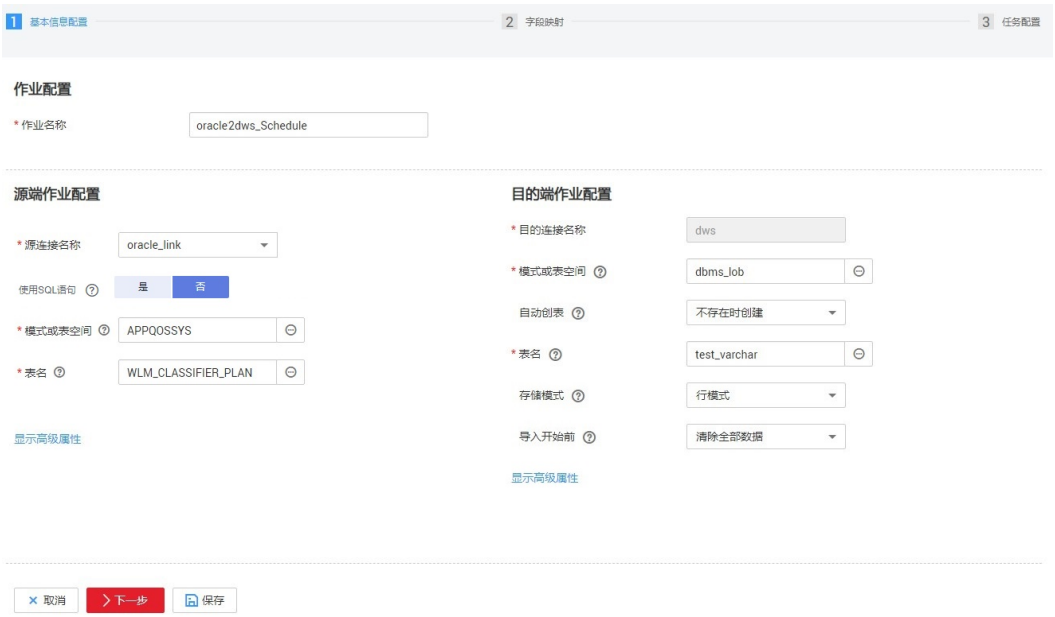
步骤3 单击“保存”完成创建连接。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从Oracle导出数据到DWS的任务。

图 3-83 创建 Oracle 到 DWS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置

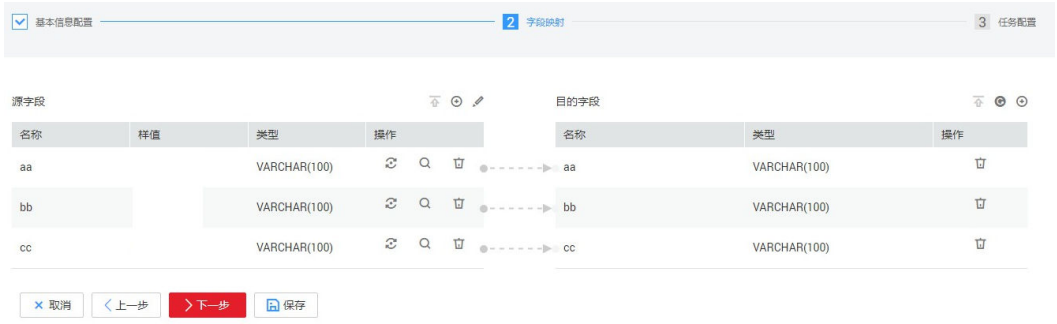
- 源连接名称：选择**创建Oracle连接**中的“oracle_link”。
- 模式或表空间：待迁移数据的数据库名称。
- 表名：待迁移数据的表名。
- 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置常见关系数据库源端参数**。

- 目的端作业配置
  - 目的连接名称：选择**创建DWS连接**中的连接“dwslink”。
  - 模式或表空间：选择待写入数据的DWS数据库。
  - 自动创表：只有当源端和目的端都为关系数据库时，才有该参数。
  - 表名：待写入数据的表名，可以手动输入一个不存在表名，CDM会在DWS中自动创建该表。
  - 存储模式：可以根据具体应用场景，建表的时候选择行存储还是列存储表。一般情况下，如果表的字段比较多（大宽表），查询中涉及到的列不多的情况下，适合列存储。如果表的字段个数比较少，查询大部分字段，那么选择行存储比较好。
  - 扩大字符字段长度：当目的端和源端数据编码格式不一样时，自动建表的字符字段长度可能不够用，配置此选项后CDM自动建表时会将字符字段扩大3倍。
  - 导入前清空数据：任务启动前，是否清除目的表中数据，用户可根据实际需要选择。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图3-84**所示。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- 单击，可批量映射字段。
- CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换。

图 3-84 表到表的字段映射



步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。

- 抽取并发数：设置同时执行的抽取任务数。可适当调大参数，提升迁移效率。
- 是否写入脏数据：表到表的迁移容易出现脏数据，建议配置脏数据归档。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

说明

如遇目的端写太久导致迁移超时，请减少Oracle连接器中“一次请求行数”参数值的设置。

3.3.8.9 OBS 数据迁移到云搜索服务

操作场景

CDM支持在云上各服务之间相互迁移数据，本章节介绍如何通过CDM将数据从OBS迁移到云搜索服务中，流程如下：

1. **创建CDM集群**
2. **创建云搜索服务连接**
3. **创建OBS连接**
4. **创建迁移作业**

前提条件

- 已获取OBS的访问域名、端口，以及AK、SK。
- 已经开通了云搜索服务，且获取云搜索服务集群的IP地址和端口。

创建 CDM 集群

参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“云搜索服务”后单击“下一步”，配置云搜索服务连接参数。

- 名称：用户自定义连接名称，例如“csslink”。
- Elasticsearch服务器列表：配置为云搜索服务集群（支持5.X以上版本）的连接地址、端口，格式为“ip:port”，多个地址之间使用分号（；）分隔，例如192.168.0.1:9200;192.168.0.2:9200。
- 用户名、密码：配置为访问云搜索服务集群的用户，需要拥有数据库的读写权限。

步骤3 单击“保存”回到连接管理界面。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

图 3-85 选择连接器类型

步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数。

- 名称：用户自定义连接名称，例如“obslink”。
- OBS服务器、端口：配置为OBS实际的地址信息。
- 访问标识（AK）、密钥（SK）：登录OBS的AK、SK。

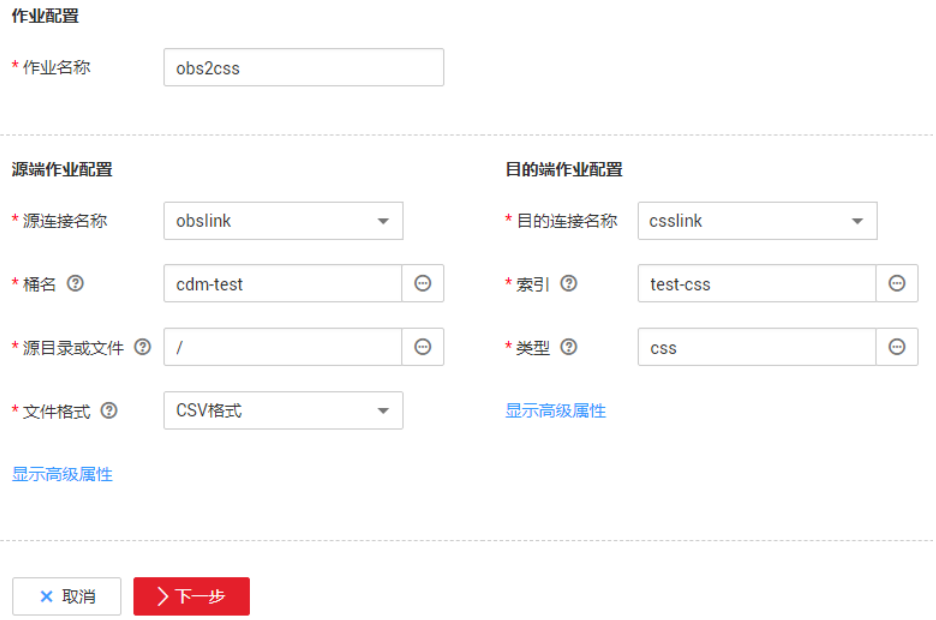
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从OBS导出数据到云搜索服务的任务。

图 3-86 创建 OBS 到云搜索服务的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建OBS连接**中的“obslink”。
  - 桶名：待迁移数据的桶。
  - 源目录或文件：待迁移数据的路径，也可以迁移桶下的所有目录、文件。
  - 文件格式：迁移文件到数据表时，文件格式选择“CSV格式”。
  - 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置OBS源端参数**。
- 目的端作业配置
  - 目的连接名称：选择**创建云搜索服务连接**中的“csslink”。
  - 索引：待写入数据的Elasticsearch索引，也可以输入一个新的索引，CDM会自动在云上搜索服务中创建。
  - 类型：待写入数据的Elasticsearch类型，可输入新的类型，CDM支持在目的端自动创建类型。
  - 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置云搜索服务目的端参数**。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图3-87**所示。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- 如果选择在目的端自动创建类型，这里还需要配置每个类型的字段类型、字段名称。
- CDM支持迁移过程中转换字段内容。

图 3-87 云搜索服务的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。这里保持默认值“1”。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.10 OBS 数据迁移到 DLI 服务

操作场景

数据湖探索（Data Lake Insight，简称DLI）提供大数据查询服务，本章节介绍使用CDM将OBS的数据迁移到DLI，使用流程如下：

1. **创建CDM集群**
2. **创建DLI连接**
3. **创建OBS连接**
4. **创建迁移作业**

前提条件

- 已经开通了OBS和DLI，并且当前用户拥有OBS的读取权限。

- 已经在DLI服务中创建好资源队列、数据库和表。

创建 CDM 集群

参考**创建集群**创建CDM集群。

该场景下，如果CDM集群只是用于迁移OBS数据到DLI，不需要迁移其他数据源，则CDM集群所在的VPC、子网、安全组选择任一个即可，没有要求，CDM通过内网访问DLI和OBS。主要是选择CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。

创建 DLI 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“数据湖探索（DLI）”后单击“下一步”，配置DLI连接参数，如**图3-88**所示。

- 名称：用户自定义连接名称，例如“dlilink”。
- 访问标识（AK）、密钥（SK）：访问DLI数据库的AK、SK。
- 项目ID：DLI所属区域的项目ID。

图 3-88 创建 DLI 连接

步骤3 单击“保存”回到连接管理界面。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

图 3-89 选择连接器类型

步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数。

- 名称：用户自定义连接名称，例如“obslink”。
- OBS服务器、端口：配置为OBS实际的地址信息。
- 访问标识（AK）、密钥（SK）：登录OBS的AK、SK。

步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从OBS迁移数据到DLI的任务，如图3-90所示。

图 3-90 创建 OBS 到 DLI 的迁移任务

- 作业名称：用户自定义作业名称。
- 源连接名称：选择**创建OBS连接**中的“obslink”。
  - 桶名：待迁移数据所属的桶。
  - 源目录或文件：待迁移数据的具体路径。
  - 文件格式：传输文件到数据表时，这里选择“CSV格式”或“JSON格式”。
  - 高级属性里的可选参数保持默认，详细说明请参见**配置OBS源端参数**。
- 目的连接名称：选择**创建DLI连接**中的“dlilink”。
  - 资源队列：选择目的表所属的资源队列。
  - 数据库名称：写入数据的数据库名称。
  - 表名：写入数据的目的表。CDM暂不支持在DLI中自动创表，这里的表需要先在DLI中创建好，且该表的字段类型和格式，建议与待迁移数据的字段类型、格式保持一致。
  - 导入前清空数据：导入数据前，选择是否清空目的表中的数据，这里保持默认“否”。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- CDM支持迁移过程中转换字段内容。

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。

- 抽取并发数：设置同时执行的抽取任务数。这里保持默认值“1”。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.11 MRS HDFS 数据迁移到 OBS

操作场景

CDM支持文件到文件类数据的迁移，本章节以MRS HDFS-->OBS为例，介绍如何通过CDM将文件类数据迁移到文件中。流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建MRS HDFS连接**
3. **创建OBS连接**
4. **创建迁移作业**

前提条件

- 已获取OBS的访问域名、端口，以及AK、SK。
- 已经了MRS。
- 拥有EIP配额。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群所在VPC、子网、安全组，选择与MRS集群所在的网络一致。

步骤2 CDM集群创建完成后，选择集群操作列的“绑定弹性IP”，CDM通过EIP访问MRS HDFS。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建 MRS HDFS 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 连接器类型选择“MRS HDFS”后单击“下一步”，配置MRS HDFS链接参数。

- 名称：用户自定义连接名称，例如“mrs_hdfs_link”。
- Manage IP：MRS Manager的IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。
- 用户名：选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。

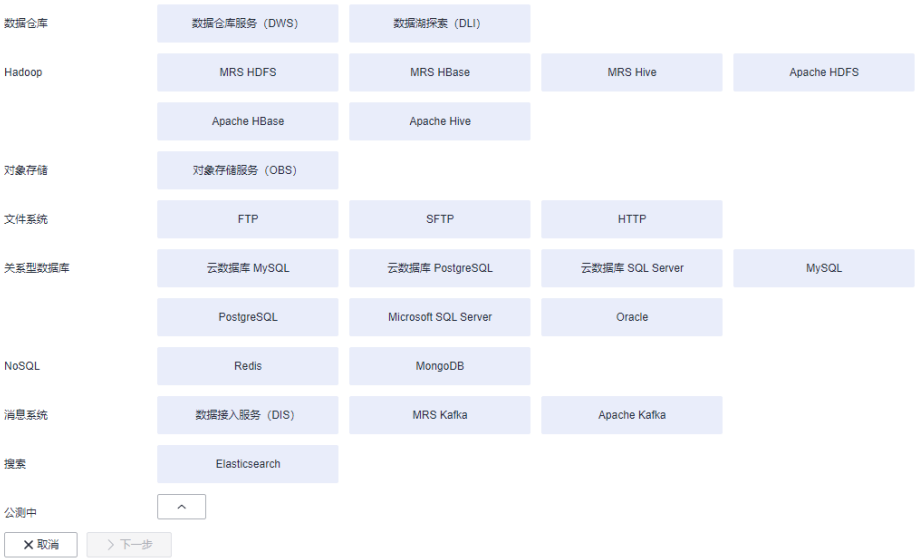
  从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。
- 密码：访问MRS Manager的用户密码。
- 认证类型：访问MRS的认证类型。
- 运行模式：选择HDFS连接的运行模式。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

图 3-91 选择连接器类型



步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数。

- 名称：用户自定义连接名称，例如“obslink”。
- OBS服务器、端口：配置为OBS实际的地址信息。
- 访问标识（AK）、密钥（SK）：登录OBS的AK、SK。

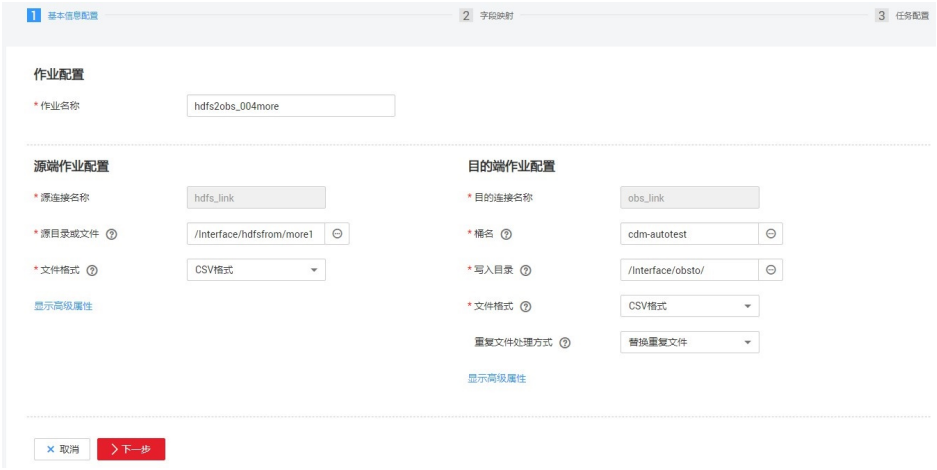
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从MRS HDFS导出数据到OBS的任务。

图 3-92 创建 MRS HDFS 到 OBS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建MRS HDFS连接**中的“hdfs_llink”。
  - 源目录或文件：待迁移数据的目录或单个文件路径。
  - 文件格式：传输数据时所用的文件格式，这里选择“二进制格式”。不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。
  - 其他可选参数一般情况下保持默认即可，详细说明请参见**配置HDFS源端参数**。
- 目的端作业配置
  - 目的连接名称：选择**创建OBS连接**中的“obs_link”。
  - 桶名：待迁移数据的桶。
  - 写入目录：写入数据到OBS服务器的目录。
  - 文件格式：迁移文件类数据到文件时，文件格式选择“二进制格式”。
  - 高级属性里的可选参数一般情况下保持默认既可，详细说明请参见**配置OBS目的端参数**。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换。

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。

- 是否定时执行：如果需要配置作业定时自动执行，请参见**配置定时任务**。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。CDM支持多个文件的并发抽取，调大参数有利于提高迁移效率
- 是否写入脏数据：否，文件到文件属于二进制迁移，不存在脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。根据使用场景，也可配置为“删除”，防止迁移作业堆积。

步骤4 单击“保存并运行”，回到作业管理界面，在作业管理界面可查看作业执行进度和结果。

步骤5 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.8.12 Elasticsearch 整库迁移到云搜索服务

操作场景

云搜索服务（Cloud Search Service）为用户提供结构化、非结构化文本的多条件检索、统计、报表，本章节介绍如何通过CDM将本地Elasticsearch整库迁移到云搜索服务中，流程如下：

1. **创建CDM集群并绑定EIP**
2. **创建云搜索服务连接**
3. **创建Elasticsearch连接**
4. **创建整库迁移作业**

前提条件

- 拥有EIP配额。
- 已经开通了云搜索服务，且获取云搜索服务集群的IP地址和端口。
- 已获取本地Elasticsearch数据库的服务器IP、端口、用户名和密码。

  如果Elasticsearch服务器是在本地数据中心或第三方云上，需要确保Elasticsearch可通过公网IP访问，或者是已经建立好了企业内部数据中心到的VPN通道或专线。

创建 CDM 集群并绑定 EIP

步骤1 参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

步骤2 CDM集群创建完成后，在集群管理界面选择“绑定弹性IP”，CDM通过EIP访问本地Elasticsearch。

说明

如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

----结束

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面。

步骤2 连接器类型选择“云搜索服务”后单击“下一步”，配置云搜索服务连接参数。

- 名称：用户自定义连接名称，例如“csslink”。
- Elasticsearch服务器列表：配置为云搜索服务集群（支持5.X以上版本）的连接地址、端口，格式为“ip:port”，多个地址之间使用分号（；）分隔，例如192.168.0.1:9200;192.168.0.2:9200。
- 用户名、密码：配置为访问云搜索服务集群的用户，需要拥有数据库的读写权限。

步骤3 单击“保存”回到连接管理界面。

----结束

创建 Elasticsearch 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面。

步骤2 连接器类型选择“Elasticsearch”后单击“下一步”，配置Elasticsearch连接参数，Elasticsearch连接参数与云搜索服务的连接参数一样：

- 名称：用户自定义连接名称，例如“es_link”。
- Elasticsearch服务器列表：配置为本地Elasticsearch数据库的IP地址、端口，多个地址之间使用分号（；）分隔。

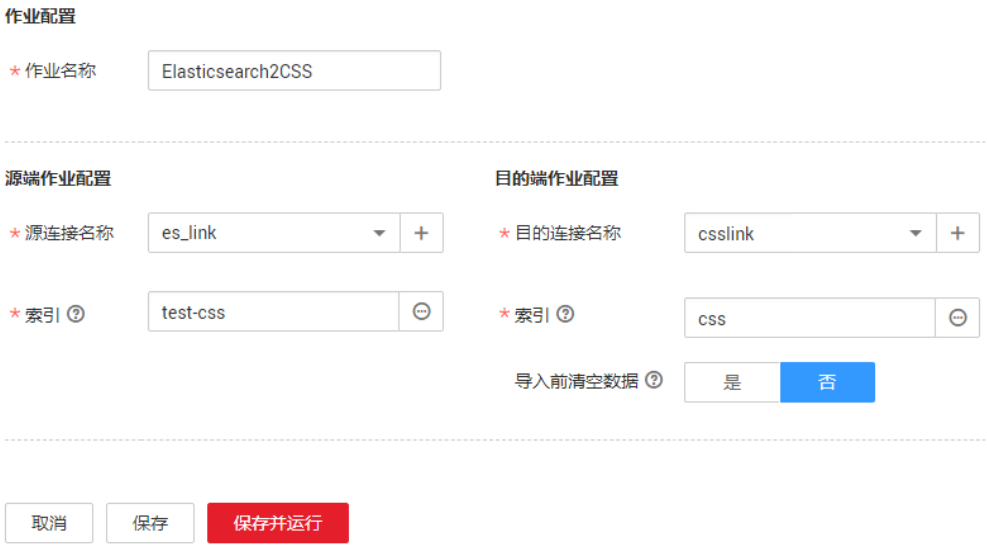
步骤3 单击“保存”回到连接管理界面。

----结束

创建整库迁移作业

步骤1 选择“整库迁移 > 新建作业”，开始创建Elasticsearch整库迁移到云搜索服务的任务。

图 3-93 创建 Elasticsearch 整库迁移作业



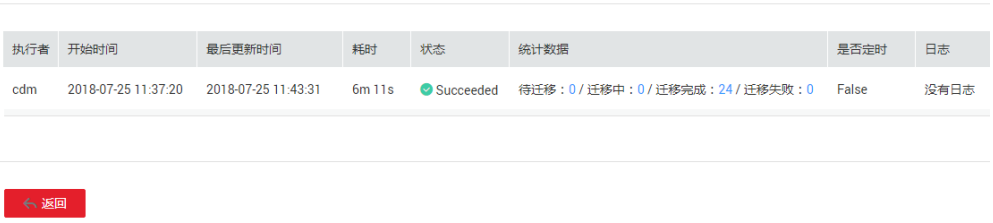
- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建Elasticsearch连接**中的“es_link”。
  - 索引：单击输入框后面的按钮，可选择本地Elasticsearch数据库中的一个索引，也可以手动输入索引名称，名称只能全部小写。需要一次迁移多个索引时，这里可配置为通配符，CDM会迁移所有符合通配符条件的索引。例如这里配置为cdm*时，CDM将迁移所有名称为cdm开头的索引：cdm01、cdmB3、cdm_45……
- 目的端作业配置
  - 目的连接名称：选择**创建云搜索服务连接**中的“csslink”。
  - 索引：待写入数据的索引，这里可以选择一个云搜索服务中已存在的索引，也可以手动输入一个不存在的索引名称，名称只能全部小写，CDM会自动在云搜索服务中创建该索引。一次迁移多个索引时，该参数将被禁止配置，CDM自动在目的端创建索引。
  - 导入前清空数据：如果上面选择的索引，在云搜索服务中已存在，这里可以选择导入数据前是否清空该索引中的数据。如果选择不清空，则数据追加写入该索引。

步骤2 作业配置完成后，单击“保存并运行”，回到作业管理界面，在整库迁移的作业管理界面可查看执行进度和结果。

本地Elasticsearch索引中的每个类型都会生成一个子作业并发执行，可以单击作业名查看子作业进度。

步骤3 作业执行完成后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据，以及日志信息（子作业才有日志）。

图 3-94 作业执行记录



----结束

3.3.9 进阶实践

3.3.9.1 增量迁移原理介绍

3.3.9.1.1 文件增量迁移

CDM支持对文件类数据源进行增量迁移，全量迁移完成之后，第二次运行作业时可以导出全部新增的文件，或者只导出特定的目录/文件。

目前CDM支持以下文件增量迁移方式：

1. **增量导出指定目录的文件**
   - 适用场景：源端数据源为文件类型（OBS/HDFS/FTP/SFTP）。这种增量迁移方式，只追加写入文件，不会更新或删除已存在的记录。
   - 关键配置：**文件/路径过滤器**+定时执行作业。
   - 前提条件：源端目录或文件名带有时间字段。
2. **增量导出指定时间以后的文件**
   - 适用场景：源端数据源为文件类型（OBS/HDFS/FTP/SFTP）。这里的指定时间，是指文件的修改时间，当文件的修改时间晚于指定的时间，CDM才迁移该文件。
   - 关键配置：**时间过滤**+定时执行作业。
   - 前提条件：无。

文件/路径过滤器

- 参数位置：在创建表/文件迁移作业时，如果源端数据源为文件类型，那么源端作业参数的高级属性中可以看到“过滤类型”参数，该参数可选择：通配符或正则表达式。
- 参数原理：“过滤类型”选择“通配符”时，CDM就可以通过用户配置的通配符过滤文件或路径，CDM只迁移满足指定条件的文件或路径。
- 配置样例：

  例如源端文件名带有时间字段“2017-10-15 20:25:26”，这个时刻生成的文件为“/opt/data/file_20171015202526.data”，则在创建作业时，参数配置如下：

  a. 过滤类型：选择“通配符”。

  b. 文件过滤器：配置为**“*${dateformat(yyyyMMdd,-1,DAY)}*”**（这是CDM支持的日期宏变量格式，详见**时间宏变量使用解析**）。

  c. 配置作业定时自动执行，“重复周期”为1天。

这样每天就可以把昨天生成的文件都导入到目的端目录，实现增量同步。

文件增量迁移场景下，“路径过滤器”的使用方法同“文件过滤器”一样，需要路径名称里带有时间字段，这样可以定期增量同步指定目录下的所有文件。

时间过滤

- 参数位置：在创建表/文件迁移作业时，如果源端数据源为文件类型，那么源端作业配置下的高级属性中，“时间过滤”参数选择“是”。
- 参数原理：“起始时间”和“终止时间”参数中输入时间值后，只有介于起始时间和终止时间的文件才会被CDM迁移。
- 配置样例：

  例如需要CDM只同步2021年1月1日~2022年1月1日生成的文件到目的端，则参数配置如下：

  a. 时间过滤器：选择为“是”。
  b. 起始时间：配置为**2021-01-01 00:00:00**（格式要求为yyyy-MM-dd HH:mm:ss）。
  c. 终止时间：配置为**2022-01-01 00:00:00**（格式要求为yyyy-MM-dd HH:mm:ss）

图 3-95 时间过滤

这样CDM作业就只迁移2021年1月1日~2022年1月1日时间段内生成的文件，下次作业再启动时就可以实现增量同步。

3.3.9.1.2 关系数据库增量迁移

CDM支持对关系型数据库进行增量迁移，全量迁移完成之后，可以增量迁移指定时间段内的数据（例如每天晚上0点导出前一天新增的数据）。

- **增量迁移指定时间段内的数据**
  - 适用场景：源端为关系型数据库，目的端没有要求。
  - 关键配置：**Where子句**+定时执行作业。
  - 前提条件：数据表中有时间日期字段或时间戳字段。

关系数据库增量迁移方式，只对数据表追加写入，不会更新或删除已存在的记录。

Where 子句

- 参数位置：在创建表/文件迁移作业时，如果源端为关系型数据库，那么在源端作业参数的高级属性下面可以看到“Where子句”参数。
- 参数原理：通过“Where子句”参数可以配置一个SQL语句（例如：age > 18 and age <= 60），CDM只导出该SQL语句指定的数据；不配置时导出整表。

  Where子句支持配置为**时间宏变量**，当数据表中有时间日期字段或时间戳字段时，配合定时执行作业，能够实现抽取指定日期的数据。
- 配置样例：

  假设数据库表中存在表示时间的列DS，类型为“varchar(30)”，插入的时间格式类似于“2017-xx-xx”，如**图3-96**所示，参数配置如下：

  图 3-96 表数据

  SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

  a. Where子句：配置为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**。

  b. 配置定时任务：重复周期为1天，每天的凌晨0点自动执行作业。

  这样就可以每天0点导出前一天产生的所有数据。Where子句支持配置多种**时间宏变量**，结合CDM定时任务的重复周期：分钟、小时、天、周、月，可以实现自动导出任意指定日期内的数据。

3.3.9.1.3 时间宏变量使用解析

在创建表/文件迁移作业时，CDM支持在源端和目的端的以下参数中配置时间宏变量：

- 源目录
- 源端的表名
- 目的端的写入目录
- 目的端的表名

- Where子句

支持通过宏定义变量表示符“${}”来完成时间类型的宏定义，当前支持两种类型：dateformat和timestamp。

通过时间宏变量+定时执行作业，可以实现数据库增量同步和文件增量同步。

dateformat

dateformat支持两种形式的参数：

- dateformat(format)

  format表示返回日期的格式，格式定义参考“java.text.SimpleDateFormat.java”中的定义。

  例如当前日期为“2017-10-16 09:00:00”，则“yyyy-MM-dd HH:mm:ss”表示“2017-10-16 09:00:00”。

- dateformat(format, dateOffset, dateType)
  - format表示返回日期的格式。
  - dateOffset表示日期的偏移量。
  - dateType表示日期的偏移量的类型。

    目前dateType支持以下几种类型：SECOND（秒），MINUTE（分钟），HOUR（小时），DAY（天）。

  例如当前日期为“2017-10-16 09:00:00”，则：
  - “dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)”表示当前时间的前一天，也就是“2017-10-15 09:00:00”。
  - “dateformat(yyyy-MM-dd HH:mm:ss, -1, HOUR)”表示当前时间的前一小时，也就是“2017-10-16 08:00:00”。
  - “dateformat(yyyy-MM-dd HH:mm:ss, -1, MINUTE)”表示当前时间的前一分钟，也就是“2017-10-16 08:59:00”。
  - “dateformat(yyyy-MM-dd HH:mm:ss, -1, SECOND)”表示当前时间的前一秒，也就是“2017-10-16 08:59:59”。

timestamp

timestamp支持两种形式的参数：

- timestamp()

  返回当前时间的时间戳，即从1970年到现在的毫秒数，如1508078516286。

- timestamp(dateOffset, dateType)

  返回经过时间偏移后的时间戳，“dateOffset”和“dateType”表示日期的偏移量以及偏移量的类型。

  例如当前日期为“2017-10-16 09:00:00”，则“timestamp(-10, MINUTE)”返回当前时间点10分钟前的时间戳，即“1508115000000”。

时间变量宏定义具体展示

假设当前时间为“2017-10-16 09:00:00”，时间变量宏定义具体如**表3-99**所示。

表 3-99 时间变量宏定义具体展示

宏变量	含义	实际显示效果
${dateformat(yyyy-MM-dd)}	以yyyy-MM-dd格式返回当前时间。	2017-10-16
${dateformat(yyyy/MM/dd)}	以yyyy/MM/dd格式返回当前时间。	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	以yyyy_MM_dd HH:mm:ss格式返回当前时间。	2017_10_16 09:00:00
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天。	2017-10-15 09:00:00
${timestamp()}	返回当前时间的时间戳，即1970年1月1日（00:00:00 GMT）到当前时间的毫秒数。	1508115600000
${timestamp(-10, MINUTE)}	返回当前时间点10分钟前的时间戳。	1508115000000
${timestamp(dateformat(yyyyMMdd))}	返回今天0点的时间戳。	1508083200000
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	返回昨天0点的时间戳。	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	返回当前整小时的时间戳。	1508115600000

路径和表名的时间宏变量

如图3-97所示，如果将：

- 源端的“表名”配置为“CDM_/${dateformat(yyyy-MM-dd)}”。
- 目的端的“写入目录”配置为“/opt/ttxx/${timestamp()}”。

经过宏定义转换，这个作业表示：将Oracle数据库的“SQOOP.CDM_20171016”表中数据，迁移到HDFS的“/opt/ttxx/1508115701746”目录中。

图 3-97 源表名和写入目录配置为时间宏变量

目前也支持一个表名或路径名中有多个宏定义变量，例如“/opt/ttxx/${dateformat(yyyy-MM-dd)}/${timestamp()}”，经过转换后为“/opt/ttxx/2017-10-16/1508115701746”。

Where 子句中的时间宏变量

以SQOOP.CDM_20171016表为例，该表中存在表示时间的列DS，如图3-98所示。

图 3-98 表数据

SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

假设当前时间为“2017-10-16”，要导出前一天的数据（即DS=‘2017-10-15’），则可以在创建作业时配置“Where子句”为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**，即可将符合DS=‘2017-10-15’条件的数据导出。

时间宏变量和定时任务配合完成增量同步

这里列举两个简单的使用场景：

- 数据库表中存在表示时间的列DS，类型为“varchar(30)”，插入的时间格式类似于“2017-xx-xx”。

  定时任务中，重复周期为1天，每天的凌晨0点执行定时任务。配置“Where子句”为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**，这样就可以在每天的凌晨0点导出前一天产生的所有数据。

- 数据库表中存在表示时间的列time，类型为“Number”，插入的时间格式为时间戳。

  定时任务中，重复周期为1天，每天的凌晨0点执行定时任务。配置“Where子句”为**time between ${timestamp(-1,DAY)} and ${timestamp()}**，这样就可以在每天的凌晨0点导出前一天产生的所有数据。

其它的配置方式原理相同。

3.3.9.1.4 HBase/CloudTable 增量迁移

使用CDM导出HBase（包括MRS HBase、FusionInsight HBase、Apache HBase）或者表格存储服务（CloudTable）的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现HBase/CloudTable的增量迁移。

在创建CDM表/文件迁移的作业，源连接选择为HBase连接或CloudTable连接时，高级属性的可选参数中可以配置时间区间。

图 3-99 HBase 时间区间

- 起始时间（包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间及以后的数据。
- 终止时间（不包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间以前的数据。

这2个参数支持配置为**时间宏变量**，例如：

- 起始时间配置为**${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}**时，表示只导出昨天以后的数据。
- 终止时间配置为**${dateformat(yyyy-MM-dd HH:mm:ss)}**时，表示只导出当前时间以前的数据。

这2个参数同时配置后，CDM就只导出前一天内的数据，再将该作业配置为每天0点执行一次，就可以增量同步每天新生成的数据。

3.3.9.2 事务模式迁移

CDM的事务模式迁移，是指当CDM作业执行失败时，将数据回滚到作业开始之前的状态，自动清理目的表中的数据。

- 参数位置：创建表/文件迁移的作业时，如果目的端为关系型数据库，在目的端作业配置的高级属性中，可以通过“先导入阶段表”参数选择是否启用事务模式。
- 参数原理：如果启用，在作业执行时CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中；导入失败则将目的表回滚到作业开始之前的状态。

图 3-100 事务模式迁移

说明

如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。

3.3.9.3 迁移文件时加解密

在迁移文件到文件系统时，CDM支持对文件加解密，目前支持以下加密方式：

- **AES-256-GCM加密**
- **KMS加密**

AES-256-GCM 加密

目前只支持AES-256-GCM（NoPadding）。该加密算法在目的端为加密，在源端为解密，支持的源端与目的端数据源如下。

- 源端支持的数据源：OBS、FTP、SFTP、HDFS（使用二进制格式传输时支持）、HTTP（适用于OBS共享文件的下载场景）。
- 目的端支持的数据源：OBS、FTP、SFTP、HDFS（使用二进制格式传输时支持）。

下面分别以OBS导出加密文件时解密、导入文件到OBS时加密为例，介绍AES-256-GCM加解密的使用方法。其它数据源的使用方法一样。

- **源端配置解密**

  创建从OBS导出文件的CDM作业时，源端数据源选择OBS后，在“源端作业配置”的“高级属性”中，配置如下参数。

  a. 加密方式：选择“AES-256-GCM”。
  b. 数据加密密钥：这里的密钥必须与**加密**时配置的密钥一致，否则解密出来的数据会错误，且系统不会提示异常。
  c. 初始化向量：这里的初始化向量必须与**加密**时配置的初始化向量一致，否则解密出来的数据会错误，且系统不会提示异常。

  这样CDM从OBS导出加密过的文件时，写入目的端的文件便是解密后的明文文件。

- **目的端配置加密**

  创建CDM导入文件到OBS的作业时，目的端数据源选择OBS后，在“目的端作业配置”的“高级属性”中，配置如下参数。

  a. 加密方式：选择“AES-256-GCM”。
  b. 数据加密密钥：用户自定义密钥，密钥由长度64的十六进制数组成，不区分大小写但必须64位，例如“DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B”。
  c. 初始化向量：用户自定义初始化向量，初始化向量由长度32的十六进制数组成，不区分大小写但必须32位，例如“5C91687BA886EDCD12ACBC3FF19A3C3F”。

  这样在CDM导入文件到OBS时，目的端OBS上的文件便是经过AES-256-GCM算法加密后的文件。

KMS 加密

说明

源端解密不支持KMS。

CDM目前只支持导入文件到OBS时，目的端使用KMS加密，表/文件迁移和整库迁移都支持。在“目的端作业配置”的“高级属性”中配置。

当启用KMS加密功能后，用户上传对象时，数据会加密成密文存储在OBS。用户从OBS下载加密对象时，存储的密文会先在OBS服务端解密为明文，再提供给用户。

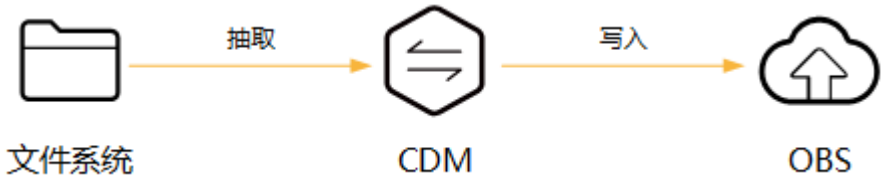
说明

- 如果选择使用KMS加密，则无法**使用MD5校验一致性**。
- 如果这里使用其它项目的KMS ID，则需要修改“项目ID”参数为KMS ID所属的项目ID；如果KMS ID与CDM在同一个项目下，“项目ID”参数保持默认即可。
- 使用KMS加密后，OBS上对象的加密状态不可以修改。
- 使用中的KMS密钥不可以删除，如果删除将导致加密对象不能下载。

3.3.9.4 MD5 校验文件一致性

CDM数据迁移以抽取-写入模式进行，CDM首先从源端抽取数据，然后将数据写入到目的端。在迁移文件到OBS时，迁移模式如**图3-101**所示。

图 3-101 迁移文件到 OBS



在这个过程中，CDM支持使用MD5检验文件一致性。

- **抽取时**
  - 该功能支持源端为OBS、HDFS、FTP、SFTP、HTTP。可校验CDM抽取的文件，是否与源文件一致。
  - 该功能由源端作业参数“MD5文件名后缀”控制（“文件格式”为“二进制格式”时生效），配置为源端文件系统中的MD5文件名后缀。
  - 当源端数据文件同一目录下有对应后缀的保存md5值的文件，例如build.sh和build.sh.md5在同一目录下。若配置了“MD5文件名后缀”，则只迁移有MD5值的文件至目的端，没有MD5值或者MD5不匹配的数据文件将迁移失败，MD5文件自身不被迁移。
  - 若未配置“MD5文件名后缀”，则迁移所有文件。
- **写入时**
  - 该功能目前只支持目的端为OBS。可校验写入OBS的文件，是否与CDM抽取的文件一致。
  - 该功能由目的端作业参数“校验MD5值”控制，读取文件后写入OBS时，通过HTTP Header将MD5值提供给OBS做写入校验，并将校验结果写入OBS桶（该桶可以不是存储迁移文件的桶）。如果源端没有MD5文件则不校验。

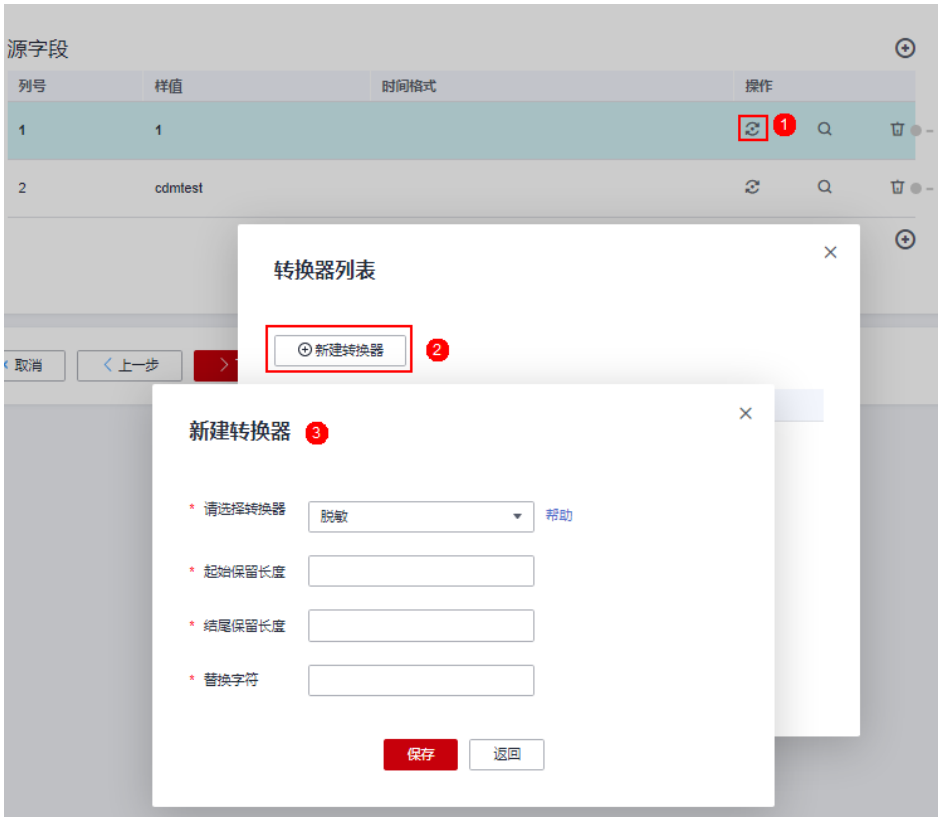
说明

- 迁移文件到文件系统时，目前只支持校验CDM抽取的文件是否与源文件一致（即只校验抽取的数据）。
- 迁移文件到OBS时，支持抽取和写入文件时都校验。
- 如果选择使用MD5校验，则无法**使用KMS加密**。

3.3.9.5 字段转换

在创建表/文件迁移作业的字段映射界面，可新建字段转换器，如**图3-102**所示。

图 3-102 新建字段转换器



说明

当使用二进制格式进行文件到文件的迁移时，没有字段映射这一步。

CDM可以在迁移过程中对字段进行转换，目前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**
- **去换行**
- **表达式转换**

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“*”。

图 3-103 字段脱敏

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。

去换行

将字段中的换行符（\n、\r、\r\n）删除。

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。

表达式支持以下工具类：

- StringUtils：字符串处理类，参考Java SDK代码的包结构“org.apache.commons.lang.StringUtils”。

- DateUtils：日期工具类。
- CommonUtils：公共工具类。
- NumberUtils：字符串转数值类。
- HttpsUtils：读取网络文件类。

应用举例：

1. 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.lowerCase(value)
2. 将当前字段的字符串全部转为大写。

   表达式：StringUtils.upperCase(value)
3. 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,"-")
4. 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
5. 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"
6. 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
7. 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")
8. 获取一个36位的UUID（Universally Unique Identifier，通用唯一识别码）。

   表达式：CommonUtils.randomUUID()
9. 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)
10. 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)
11. 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*)
12. 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)
13. 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*")
14. 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*")

15. 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*")

16. 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*")

17. 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*)

18. 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*)

19. 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*")

20. 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*")

21. 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*")

22. 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*)

23. 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*")

24. 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)

25. 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)

26. 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)

27. 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)

28. 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)

29. 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)

30. 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)

31. 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*)

32. 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*)

33. 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“yzyzybat”。

    表达式：StringUtils.leftPad(value,*8*,"*yz*")

34. 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“batyzyzy”。

    表达式：StringUtils.rightPad(value,*8*,"*yz*")

35. 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)

36. 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*")

37. 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*")

38. 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.*")

39. 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“zbz”。

    表达式：StringUtils.replace(value,"*a*","*z*")

40. 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*")

41. 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*")

42. 如果当前字段为字符串类型，去除字段中所有指定的字符，例如去除“abcyx”中所有的“x”、“y”和“z”，转换后为“abc”。

    表达式：StringUtils.strip(value,"*xyz*")

43. 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所有空格。

表达式：StringUtils.stripEnd(value,*null*)

44. 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所有空格。

    表达式：StringUtils.stripStart(value,*null*)

45. 如果当前字段为字符串类型，获取字符串指定位置后（不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。例如获取“abcde”第2个字符后的字符串，则转换后为“cde”。

    表达式：StringUtils.substring(value,*2*)

46. 如果当前字段为字符串类型，获取字符串指定区间的子字符串，区间位置如果为负数，则从末尾往前计算位置。例如获取“abcde”第2个字符后、第5个字符前的字符串，则转换后为“cd”。

    表达式：StringUtils.substring(value,*2*,*5*)

47. 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*")

48. 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*")

49. 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*")

50. 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*")

51. 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*")

52. 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（char≤32），例如删除字符串前后的空格。

    表达式：StringUtils.trim(value)

53. 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)

54. 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*)

55. 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)

56. 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*)

57. 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)

58. 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*)

59. 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)

60. 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*)

61. 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)

62. 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*)

63. 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)

64. 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*)

65. 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)

66. 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*")

67. 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*"))

68. 取出缓存的IP与地址映射对象。

    表达式：CommonUtils.getCache("*ipList*")

69. 判断是否有IP与地址映射缓存。

    表达式：CommonUtils.cacheExists("*ipList*")

70. 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*)

3.3.9.6 指定文件名迁移

从FTP/SFTP/OBS导出文件时，CDM支持指定文件名迁移，用户可以单次迁移多个指定的文件（最多50个），导出的多个文件只能写到目的端的同一个目录。

在创建表/文件迁移作业时，如果源端数据源为FTP/SFTP/OBS，CDM源端的作业参数“源目录或文件”支持输入多个文件名（最多50个），文件名之间默认使用“|”分隔，您也可以自定义文件分隔符，从而实现文件列表迁移。

说明

1. 迁移文件或对象时支持文件级增量迁移（通过配置跳过重复文件实现），但不支持断点续传。

   例如要迁移3个文件，第2个文件迁移到一半时由于网络原因失败，再次启动迁移任务时，会跳过第1个文件，从第2个文件开始重新传，但不能从第2个文件失败的位置重新传。
2. 文件迁移时，单个任务支持千万数量的文件，如果待迁移目录下文件过多，建议拆分到不同目录并创建多个任务。

3.3.9.7 正则表达式分隔半结构化文本

在创建表/文件迁移作业时，对简单CSV格式的文件，CDM可以使用字段分隔符进行字段分隔。但是对于一些复杂的半结构化文本，由于字段值也包含了分隔符，所以无法使用分隔符进行字段分隔，此时可以使用正则表达式分隔。

正则表达式参数在源端作业参数中配置，要求源连接为对象存储或者文件系统，且“文件格式”必须选择“CSV格式”。

图 3-104 正则表达式参数

在迁移CSV格式的文件时，CDM支持使用正则表达式分隔字段，并按照解析后的结果写入目的端。正则表达式语法请参考对应的相关资料，这里举例下面几种日志文件的正则表达式的写法：

- **Log4J日志**
- **Log4J审计日志**
- **Tomcat日志**
- **Django日志**

- Apache server日志

Log4J 日志

- 日志样例：
```
2018-01-11 08:50:59,001 INFO
[org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)]
Adding jars to current classloader from property: org.apache.sqoop.classpath.extra
```
- 正则表达式为：
```
^(\d.*\d) (\w*) \[(.*)\] (\w.*).*
```
- 解析出的结果如下：

表 3-100 Log4J 日志解析结果

列号	样值
1	2018-01-11 08:50:59,001
2	INFO
3	org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)
4	Adding jars to current classloader from property: org.apache.sqoop.classpath.extra

Log4J 审计日志

- 日志样例：
```
2018-01-11 08:51:06,156 INFO
[org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)]
user=sqoop.anonymous.user ip=189.xxx.xxx.75 op=show obj=version objId=x
```
- 正则表达式为：
```
^(\d.*\d) (\w*) \[(.*)\] user=(\w.*) ip=(\w.*) op=(\w.*) obj=(\w.*) objId=(.*).*
```
- 解析结果如下：

表 3-101 Log4J 审计日志解析结果

列号	样值
1	2018-01-11 08:51:06,156
2	INFO
3	org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)
4	sqoop.anonymous.user
5	189.xxx.xxx.75
6	show
7	version

列号	样值
8	x

Tomcat 日志

- 日志样例：
  11-Jan-2018 09:00:06.907 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log OS Name: Linux
- 正则表达式为：
  ^(\d.*\d) (\w*) \[(.*)\] ([\w\.]*) (\w.*).*
- 解析结果如下：

表 3-102 Tomcat 日志解析结果

列号	样值
1	11-Jan-2018 09:00:06.907
2	INFO
3	main
4	org.apache.catalina.startup.VersionLoggerListener.log
5	OS Name:Linux

Django 日志

- 日志样例：
  [08/Jan/2018 20:59:07] settings INFO Welcome to Hue 3.9.0
- 正则表达式为：
  ^\[(.*)\] (\w*) (\w*) (.*).*
- 解析结果如下：

表 3-103 Django 日志解析结果

列号	样值
1	08/Jan/2018 20:59:07
2	settings
3	INFO
4	Welcome to Hue 3.9.0

Apache server 日志

- 日志样例：
  [Mon Jan 08 20:43:51.854334 2018] [mpm_event:notice] [pid 36465:tid 140557517657856] AH00489: Apache/2.4.12 (Unix) OpenSSL/1.0.1t configured -- resuming normal operations

- 正则表达式为：
  ```
  ^\[(.*)\] \[(.*)\] \[(.*)\] (.*).*
  ```
- 解析结果如下：

表 3-104 Apache server 日志解析结果

列号	样值
1	Mon Jan 08 20:43:51.854334 2018
2	mpm_event:notice
3	pid 36465:tid 140557517657856
4	AH00489: Apache/2.4.12 (Unix) OpenSSL/1.0.1t configured -- resuming normal operations

3.3.9.8 记录数据迁移入库时间

CDM在创建表/文件迁移的作业，支持连接器源端为关系型数据库时，在表字段映射中使用时间宏变量增加入库时间字段，用以记录关系型数据库的入库时间等用途。

前提条件

已创建连接器源端为关系型数据库，以及目的端数据连接。

创建表/文件迁移作业

步骤1 在创建表/文件迁移作业时，选择已创建的源端连接器、目的端连接器。

图 3-105 配置作业

步骤2 单击“下一步”，进入“字段映射”配置页面后，单击源字段图标。

图 3-106 配置字段映射

步骤3 选择“自定义字段”页签，填写字段名称及字段值后单击“确认”按钮，例如：

名称：InputTime。

值：${timestamp()}，更多时间宏变量请参见表3-105。

图 3-107 添加字段

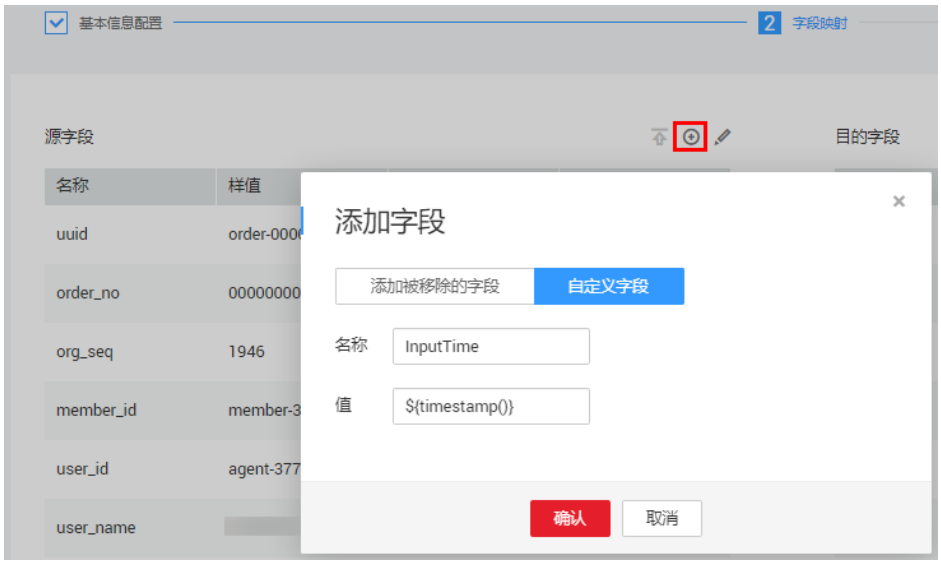


表 3-105 时间变量宏定义具体展示

宏变量	含义	实际显示效果
${dateformat(yyyy-MM-dd)}	以yyyy-MM-dd格式返回当前时间。	2017-10-16
${dateformat(yyyy/MM/dd)}	以yyyy/MM/dd格式返回当前时间。	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	以yyyy_MM_dd HH:mm:ss格式返回当前时间。	2017_10_16 09:00:00
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天。	2017-10-15 09:00:00
${timestamp()}	返回当前时间的时间戳，即1970年1月1日（00:00:00 GMT）到当前时间的毫秒数。	1508115600000
${timestamp(-10, MINUTE)}	返回当前时间点10分钟前的时间戳。	1508115000000
${timestamp(dateformat(yyyyMMdd))}	返回今天0点的时间戳。	1508083200000

宏变量	含义	实际显示效果
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	返回昨天0点的时间戳。	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	返回当前整小时的时间戳。	1508115600000

说明

- 添加完字段后，新增的字段在界面不显示样值，不会影响字段值的传输，CDM会将字段值直接写入目的端。
- 这里“添加字段”中“自定义字段”的功能，要求源端连接器为JDBC连接器、HBase连接器、MongoDB连接器、ElasticSearch连接器、Kafka连接器，或者目的端为HBase连接器。

步骤4 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行，这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。这里保持默认值“1”。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。

步骤5 单击“保存并运行”，回到作业管理的表/文件迁移界面，在作业管理界面可查看作业执行进度和结果。

步骤6 作业执行成功后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据。

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

3.3.9.9 文件格式介绍

在创建CDM作业时，有些场景下源端、目的端的作业参数中需要选择“文件格式”，这里分别介绍这几种文件格式的使用场景、子参数、公共参数、使用示例等。

- **CSV格式**
- **JSON格式**
- **二进制格式**

- 文件格式的公共参数
- 文件格式问题解决方法

CSV 格式

如果想要读取或写入某个CSV文件，请在选择“文件格式”的时候选择“CSV格式”。CSV格式的主要有以下使用场景：

- 文件导入到数据库、NoSQL。
- 数据库、NoSQL导出到文件。

选择了CSV格式后，通常还可以配置以下可选子参数：

1.换行符

2.字段分隔符

3.编码类型

4.使用包围符

5.使用正则表达式分隔字段

6.首行为标题行

7.写入文件大小

1. **换行符**

   用于分隔文件中的行的字符，支持单字符和多字符，也支持特殊字符。特殊字符可以使用URL编码输入，例如：

   表 3-106 特殊字符对应的 URL 编码

特殊字符	URL编码
空格	%20
Tab	%09
%	%25
回车	%0d
换行	%0a
标题开头\u0001 (SOH)	%01

2. **字段分隔符**

   用于分隔CSV文件中的列的字符，支持单字符和多字符，也支持特殊字符，详见表3-106。

3. **编码类型**

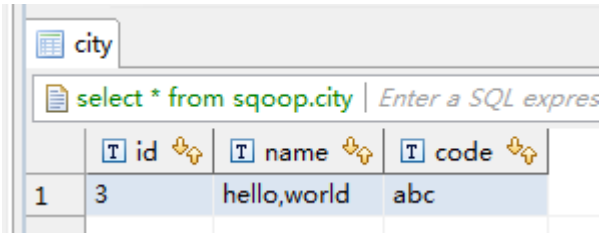
   文件的编码类型，默认是UTF-8。

   如果源端指定该参数，则使用指定的编码类型去解析文件；目的端指定该参数，则写入文件的时候，以指定的编码类型写入。

4. **使用包围符**

- 数据库、NoSQL导出到CSV文件（“使用包围符”在目的端）：当源端某列数据的字符串中出现字段分隔符时，目的端可以通过开启“使用包围符”，将该字符串括起来，作为一个整体写入CSV文件。CDM目前只使用双引号（""）作为包围符。如图3-108所示，数据库的name字段的值中包含了字段分隔符逗号：

  图 3-108 包含字段分隔符的字段值

  不使用包围符的时候，导出的CSV文件，数据会显示为：

  ```
  3.hello,world,abc
  ```

  如果使用包围符，导出的数据则为：

  ```
  3,"hello,world",abc
  ```

  如果数据库中的数据已经包含了双引号（""），那么使用包围符后，导出的CSV文件的包围符会是三个双引号（"""）。例如字段的值为：a"hello,world"c，使用包围符后导出的数据为：

  ```
  """a"hello,world"c"""
  ```

- CSV文件导出到数据库、NoSQL（“使用包围符”在源端）：CSV文件为源，并且其中数据是被包围符括起来的时候，如果想把数据正确的导入到数据库，就需要在源端开启“使用包围符”，这样包围符内的值的，会写入一个字段内。

5. **使用正则表达式分隔字段**

   这个功能是针对一些复杂的半结构化文本，例如日志文件的解析，详见：**使用正则表达式分隔半结构化文本**。

6. **首行为标题行**

   这个参数是针对CSV文件导出到其它地方的场景，如果源端指定了该参数，CDM在抽取数据时将第一行作为标题行。在传输CSV文件的时候会跳过标题行，这时源端抽取的行数，会比目的端写入的行数多一行，并在日志文件中进行说明跳过了标题行。

7. **写入文件大小**

   这个参数是针对数据库导出到CSV文件的场景，如果一张表的数据量比较大，那么导出到CSV文件的时候，会生成一个很大的文件，有时会不方便下载或查看。这时可以在目的端指定该参数，这样会生成多个指定大小的CSV文件，避免导出的文件过大。该参数的数据类型为整型，单位为MB。

JSON 格式

这里主要介绍JSON文件格式的以下内容：

- **CDM支持解析的JSON类型**
- **记录节点**
- **从JSON文件复制数据**

1. **CDM支持解析的JSON类型：JSON对象、JSON数组。**
   - JSON对象：JSON文件包含单个对象，或者以行分隔/串连的多个对象。
     i. 单一对象JSON：
     ```
     {
        "took" : 190,
        "timed_out" : false,
        "total" : 1000001,
        "max_score" : 1.0
     }
     ```
     ii. 行分隔的JSON对象：
     ```
     {"took" : 188, "timed_out" : false, "total" : 1000003, "max_score" : 1.0 }
     {"took" : 189, "timed_out" : false, "total" : 1000004, "max_score" : 1.0 }
     ```
     iii. 串连的JSON对象：
     ```
     {
       "took": 190,
       "timed_out": false,
       "total": 1000001,
       "max_score": 1.0
     }
     {
       "took": 191,
       "timed_out": false,
       "total": 1000002,
       "max_score": 1.0
     }
     ```
   - JSON数组：JSON文件是包含多个JSON对象的数组。
     ```
     [{
        "took" : 190,
        "timed_out" : false,
        "total" : 1000001,
        "max_score" : 1.0
     },
     {
        "took" : 191,
        "timed_out" : false,
        "total" : 1000001,
        "max_score" : 1.0
     }]
     ```
2. **记录节点**

   记录数据的根节点。该节点对应的数据为JSON数组，CDM会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。
3. **从JSON文件复制数据**

   a. 示例一：从行分隔/串连的多个对象中提取数据。JSON文件包含了多个JSON对象，例如：
   ```
   {
     "took": 190,
     "timed_out": false,
     "total": 1000001,
     "max_score": 1.0
   }
   {
     "took": 191,
     "timed_out": false,
     "total": 1000002,
     "max_score": 1.0
   }
   {
     "took": 192,
     "timed_out": false,
     "total": 1000003,
     "max_score": 1.0
   }
   ```

如果您想要从该JSON对象中提取数据，使用以下格式写入到数据库，只需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，然后在作业第二步进行字段匹配即可。

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0
192	false	1000003	1.0

b. 示例二：从记录节点中提取数据。JSON文件包含了单个的JSON对象，但是其中有效的数据在一个数据节点下，例如：

```
{
   "took": 190,
   "timed_out": false,
   "hits": {
      "total": 1000001,
      "max_score": 1.0,
      "hits":
       [{
          "_id": "650612",
          "_source": {
              "name": "tom",
              "books": ["book1","book2","book3"]
          }
       },
       {
           "_id": "650616",
           "_source": {
               "name": "tom",
               "books": ["book1","book2","book3"]
           }
       },
     {
           "_id": "650618",
           "_source": {
               "name": "tom",
               "books": ["book1","book2","book3"]
           }
       }]
   }
}
```

如果想以如下格式写入到数据库，则需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，并且指定记录节点为“hits.hits”，然后在作业第二步进行字段匹配。

ID	SourceName	SourceBooks
650612	tom	["book1","book2","book3"]
650616	tom	["book1","book2","book3"]
650618	tom	["book1","book2","book3"]

c. 示例三：从JSON数组中提取数据。JSON文件是包含了多个JSON对象的JSON数组，例如：

```
[{
    "took" : 190,
    "timed_out" : false,
    "total" : 1000001,
```

```
    "max_score" : 1.0
},
{
    "took" : 191,
    "timed_out" : false,
    "total" : 1000002,
    "max_score" : 1.0
}]
```

如果想以如下格式写入到数据库，需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON数组”，然后在作业第二步进行字段匹配。

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0

d. 示例四：在解析JSON文件的时候搭配转换器。在示例二前提下，想要把hits.max_score字段附加到所有记录中，即以如下格式写入到数据库中：

ID	SourceName	SourceBooks	MaxScore
650612	tom	["book1","book2","book3"]	1.0
650616	tom	["book1","book2","book3"]	1.0
650618	tom	["book1","book2","book3"]	1.0

则需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，并且指定记录节点为“hits.hits”，然后在作业第二步添加转换器，操作步骤如下：

i. 单击⊕添加字段，新增一个字段。

图 3-109 添加字段

ii. 在添加的新字段后面，单击 添加字段转换器。

图 3-110 添加字段转换器

iii. 创建“表达式转换”的转换器，表达式输入”1.0”，然后保存。

图 3-111 配置字段转换器

二进制格式

如果想要在文件系统间按原样复制文件，则可以选择二进制格式。二进制格式传输文件到文件的速率、性能都最优，且不需要在作业第二步进行字段匹配。

- **文件传输的目录结构**

  CDM的文件传输，支持单文件，也支持一次传输目录下所有的文件。传输到目的端后，目录结构会保持原样。

- **增量迁移文件**

  使用CDM进行二进制传输文件时，目的端有一个参数“重复文件处理方式”，可以用作文件的增量迁移，具体请参见**文件增量迁移**。

增量迁移文件的时候，选择“重复文件处理方式”为“跳过重复文件”，这样如果源端有新增的文件，或者是迁移过程中出现了失败，只需要再次运行任务，已经迁移过的文件就不会再次迁移。

- **写入到临时文件**

  二进制迁移文件时候，可以在目的端指定是否写入到临时文件。如果指定了该参数，在文件复制过程中，会将文件先写入到一个临时文件中，迁移成功后，再进行rename或move操作，在目的端恢复文件。

- **生成文件MD5值**

  对每个传输的文件都生成一个MD5值，并将该值记录在一个新文件中，新文件以“.md5”作为后缀，并且可以指定MD5值生成的目录。

文件格式的公共参数

- **源文件处理方式**

  CDM在文件复制成功后，可以对源端文件进行操作，包括：不处理、重命名源文件或者删除源文件。

- **启动作业标识文件**

  这个主要用于自动化场景中，CDM配置了定时任务，周期去读取源端文件，但此时源端的文件正在生成中，CDM此时读取会造成重复写入或者是读取失败。所以，可以在源端作业参数中指定启动作业标识文件为“ok.txt”，在源端生成文件成功后，再在文件目录下生成“ok.txt”，这样CDM就能读取到完整的文件。

  另外，可以设置超时时间，在超时时间内，CDM会周期去查询标识文件是否存在，超时后标识文件还不存在的话，则作业任务失败。

  启动作业标识文件本身不会被迁移。

- **作业成功标识文件**

  文件系统为目的端的时候，当任务成功时，在目的端的目录下，生成一个空的文件，标识文件名由用户来指定。一般和“启动作业标识文件”搭配使用。

  这里需要注意的是，不要和传输的文件混淆，例如传输文件为finish.txt，但如果作业成功标识文件也设置为finish.txt，这样会造成这两个文件相互覆盖。

- **过滤器**

  使用CDM迁移文件的时候，可以使用过滤器来过滤文件。支持通过通配符或时间过滤器来过滤文件。

  - 选择通配符时，CDM只迁移满足过滤条件的目录或文件。
  - 选择时间过滤器时，只有文件的修改时间晚于输入的时间才会被传输。

  例如：用户的“/table/”目录下存储了很多数据表的目录，并且按天进行了划分：DRIVING_BEHAVIOR_20180101～DRIVING_BEHAVIOR_20180630，保存了DRIVING_BEHAVIOR从1月到6月的所有数据。如果只想迁移DRIVING_BEHAVIOR的3月份的表数据。那么需要在作业第一步指定源目录为“/table”，过滤类型选择“通配符”，然后指定“路径过滤器”为“DRIVING_BEHAVIOR_201803*”。

文件格式问题解决方法

1. 数据库的数据导出到CSV文件，由于数据中含有分隔符逗号，造成导出的CSV文件中数据混乱。

   CDM提供了以下几种解决方法：

   a. 指定字段分隔符

使用数据库中不存在的字符，或者是极少见的不可打印字符来作为字段分隔符。例如：可以在目的端指定“字段分隔符”为“%01”，这样导出的字段分隔符就是“\u0001”，详情可见**表3-106**。

b. 使用包围符

在目的端作业参数中开启“使用包围符”，这样数据库中如果字段包含了字段分隔符，在导出到CSV文件的时候，CDM会使用包围符将该字段括起来，使之作为一个字段的值写入CSV文件。

2. 数据库的数据包含换行符

场景：使用CDM先将MySQL中的某张表（表的某个字段值中包含了换行符\n）导出到CSV格式的文件中，然后再使用CDM将导出的CSV文件导入到MRS HBase，发现导出的CSV文件中出现了数据被截断的情况。

解决方法：指定换行符。

在使用CDM将MySQL的表数据导出到CSV文件时，指定目的端的换行符为“%01”（确保这个值不会出现在字段值中），这样导出的CSV文件中换行符就是“%01”。然后再使用CDM将CSV文件导入到MRS HBase时，指定源端的换行符为“%01”，这样就避免了数据被截断的问题。

3.4 数据开发

3.4.1 数据开发概述

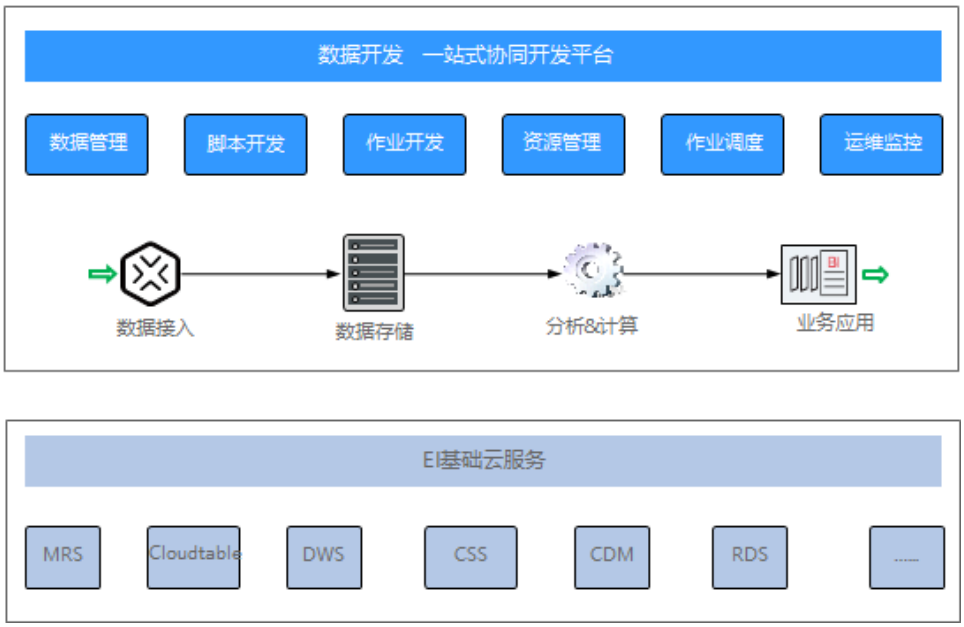
数据开发是一个一站式的大数据协同开发平台，提供全托管的大数据调度能力。它可管理多种大数据服务，极大降低用户使用大数据的门槛，帮助您快速构建大数据处理中心。

数据开发模块曾被称为数据湖工厂（Data Lake Factory，后简称DLF）服务，因此在本文中，“数据湖工厂”、“DLF”均可用于指代“数据开发”模块。

数据开发简介

使用数据开发模块，用户可进行数据管理、脚本开发、作业开发、作业调度、运维监控等操作，轻松完成整个数据的处理分析流程。

图 3-112 数据开发模块架构



数据开发的主要功能

表 3-107 数据开发的主要功能

支持的功能	说明
数据管理	• 支持管理DWS、DLI、MRS Hive等多种数据仓库。 • 支持可视化和DDL方式管理数据库表。
脚本开发	• 提供在线脚本编辑器，支持多人协作进行SQL、Shell、Python脚本在线代码开发和调测。 • 支持使用变量和函数。
作业开发	• 提供图形化设计器，支持拖拉拽方式快速构建数据处理工作流。 • 预设数据集成、SQL、Shell等多种任务类型，通过任务间依赖完成复杂数据分析处理。 • 支持导入和导出作业。
资源管理	支持统一管理在脚本开发和作业开发使用到的file、jar、archive类型的资源。
作业调度	支持单次调度、周期调度和事件驱动调度，周期调度支持分钟、小时、天、周、月多种调度周期。
运维监控	• 支持对作业进行运行、暂停、恢复、终止等多种操作。 • 支持查看作业和其内各任务节点的运行详情。 • 支持配置多种方式报警，作业和任务发生错误时可及时通知相关人，保证业务正常运行。

数据开发中的对象

- 数据连接：定义访问数据实体存储（计算）空间所需信息的集合，包括连接类型、名称和登录信息等。
- 解决方案：解决方案为用户提供便捷的、系统的方式管理作业，更好地实现业务需求和目标。每个解决方案可以包含一个或多个业务相关的作业，一个作业可以被多个解决方案复用。
- 作业：作业由一个或多个节点组成，共同执行以完成对数据的一系列操作。
- 脚本：脚本（Script）是一种批处理文件的延伸，是一种纯文本保存的程序，一般来说的计算机脚本程序是确定的一系列控制计算机进行运算操作动作的组合，在其中可以实现一定的逻辑分支等。
- 节点：定义对数据执行的操作。
- 资源：用户可以上传自定义的代码或文本文件作为资源，以便在节点运行时调用。
- 表达式：数据开发作业中的节点参数可以使用表达式语言（Expression Language，简称EL），根据运行环境动态生成参数值。数据开发EL表达式包含简单的算术和逻辑计算，引用内嵌对象，包括作业对象和一些工具类对象。
- 环境变量：环境变量是在操作系统中一个具有特定名字的对象，它包含了一个或者多个应用程序所将使用到的信息。
- 补数据：手工触发周期方式调度的作业任务，生成某时间段内的实例。

3.4.2 数据管理

3.4.2.1 数据管理流程

数据管理功能可以协助用户快速建立数据模型，为后续的脚本和作业开发提供数据实体。通过数据管理，您可以：

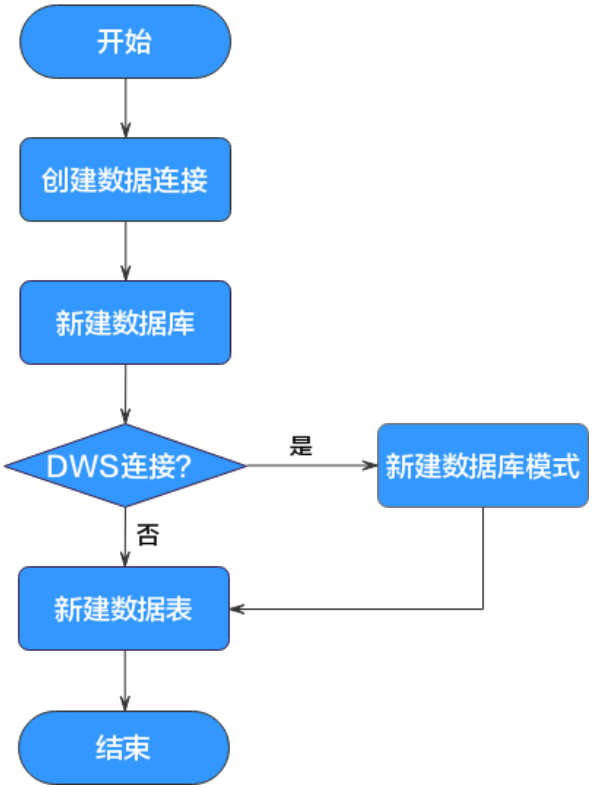
- 支持管理DWS、MRS Hive等多种数据湖。
- 支持可视化和DDL方式管理数据库表。

说明

如果您在使用数据开发前，已参考**使用DataArts Studio前的准备**创建了数据连接和对应的数据库和数据表，则可跳过数据管理操作，直接进入**脚本开发**或**作业开发**。

数据管理的使用流程如下：

图 3-113 数据管理流程



1. 创建数据连接，连接相关数据湖底座服务。具体请参见**新建数据连接**。
2. 基于相应服务，新建数据库。具体请参见**新建数据库**。
3. 如果是DWS连接，则需要新建数据库模式；否则直接新建数据表。具体请参见**（可选）新建数据库模式**。
4. 新建数据表。具体请参见**新建数据表**。

3.4.2.2 新建数据连接

通过创建数据连接，您可以在数据开发模块中对相应服务进行更多数据操作，例如：管理数据库、管理命名空间、管理数据库模式、管理数据表。

在同一个数据连接下，可支持多个作业运行和多个脚本开发，当数据连接保存的信息发生变化时，您只需在连接管理中编辑修改该数据连接的信息。

新建数据连接

数据开发模块的数据连接，是基于管理中心的数据连接完成的，创建方法请参考**创建数据连接**。

查看连接引用

当用户需要查看某个连接被引用的情况时，可以参考如下操作查看引用。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-114 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 单击，进入连接目录列表。
4. 在连接目录中，右键单击对应的连接，选择“查看引用”，弹出“引用列表”窗口。
5. 在引用列表窗口，可以查看该连接被引用的情况。

3.4.2.3 新建数据库

数据连接创建完成后，您可以基于数据连接，通过可视化模式或SQL脚本方式新建数据库。

- （推荐）可视化模式：您可以直接在DataArts Studio数据开发模块通过No Code方式，新建数据库。
- SQL脚本方式：您也可以在DataArts Studio数据开发模块或对应数据湖产品的SQL编辑器上，开发并执行用于创建数据库的SQL脚本，从而创建数据库。

本章节以可视化模式为例，介绍如何在数据开发模块新建数据库。

前提条件

- 已开通相应的云服务。
- 已新建数据连接，请参见**新建数据连接**。
- MRS API方式连接不支持通过可视化模式管理数据库，建议通过SQL脚本方式进行创建。
- 删除数据库时，请确保该数据库未被使用，且没有关联数据表。

新建数据库（可视化模式）

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-115 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。
3. 在左侧菜单选择  ，右键单击数据连接名称，选择“新建数据库”，配置如**表 3-108**所示的参数。

表 3-108 新建数据库

参数	是否必选	说明
数据库名称	是	数据库的名称，命名要求如下： • DLI：数据库名称只能包含数字、英文字母和下划线，但不能是纯数字，且不能以下划线开头。 • DWS：数据库名称只能包含数字、英文字母和下划线，但不能是纯数字，且不能以下划线开头。 • MRS Hive：只能包含英文字母、数字、“_”，只能以数字和字母开头，不能全部为数字，且长度为1~128个字符。
描述	否	数据库的描述信息，填写要求如下： • DLI：最大长度为256个字符。 • DWS：最大长度为1024个字符。 • MRS Hive：最大长度为1024个字符。

4. 单击“确定”，新建数据库。

编辑数据库

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。

2. 在左侧菜单选择，展开创建的数据连接，并右键单击数据库名称，选择“修改”。
3. 在弹出的页面中修改数据库的信息。
4. 单击“确定”，保存修改。

删除数据库

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。
2. 在左侧菜单选择，展开创建的数据连接，并右键单击数据连接名称，选择“删除”。
3. 在弹出的数据连接列表页面，单击“删除”。
4. 单击“确定”，保存修改。

3.4.2.4 （可选）新建数据库模式

DWS数据连接创建完成后，用户可以在右侧区域中管理DWS数据连接的数据库模式。

前提条件

- 已新建DWS数据连接，请参见**新建数据连接**。
- 已新建DWS数据库，请参见**新建数据库**。

新建数据库模式

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-116 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。

3. 在左侧菜单选择  ，单击DWS数据连接名称，选择需配置的数据库，展开目录层级至“schemas”，右键单击“schemas”，选择“新建模式”。
4. 在弹出的“新建模式”页面，配置如表3-109所示的参数。

表 3-109 新建模式

参数	是否必选	说明
模式名称	是	数据库模式的名称。
描述	否	数据库模式的描述信息。

5. 单击“确定”，新建数据库模式。

修改数据库模式

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。
2. 在左侧菜单选择  ，单击数据连接名称，选择数据库，目录层级展开至需要修改的数据库模式，右键单击数据库模式名称，选择“修改”。
3. 在弹出的“修改模式”页面，修改数据库模式的描述信息。
4. 单击“确定”，保存修改。

删除数据库模式

说明

- 默认的数据库模式不可删除。
- 删除操作不可撤销，请谨慎操作。

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。
2. 在左侧菜单选择  ，单击数据连接名称，选择数据库，目录层级展开至需要删除的数据库模式，右键单击数据库模式名称，选择“删除”。
3. 在弹出的“删除模式”页面，单击“确定”，删除数据库模式。

3.4.2.5 新建数据表

您可以通过可视化模式、DDL模式或SQL脚本方式新建数据表。

- （推荐）可视化模式：您可以直接在DataArts Studio数据开发模块通过No Code方式，新建数据表。
- （推荐）DDL模式：您可以在DataArts Studio数据开发模块，通过选择DDL方式，通过SQL语句新建数据表。
- SQL脚本方式：您也可以在DataArts Studio数据开发模块或对应数据湖产品的SQL编辑器上，开发并执行用于创建数据表的SQL脚本，从而创建数据表。

本章节以可视化模式和DDL模式为例，介绍如何在数据开发模块新建数据表。

前提条件

- 已在云服务中创建数据库。
- 已在数据开发模块中创建与数据表类型匹配的数据连接，请参见**新建数据连接**。

新建数据表（可视化模式）

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-117 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”，进入“右侧区域”页面。
3. 在左侧菜单选择  ，单击“数据连接”，目录层级展开至“tables”，右键单击“新建数据表”。
4. 在弹出的对话框中，显示“配置基本属性”页面，选择“数据表连接类型”，并参见**表3-110**配置相关参数。

   表 3-110 基本属性

数据连接类型	参数说明
DLI	请见**表3-114**的“基本属性”部分
DWS	请见**表3-115**的“基本属性”部分
MRS Hive	请见**表3-116**的“基本属性”部分

5. 单击“下一步”，在“配置表结构”页面配置如**表3-111**所示的参数。

表 3-111 表结构

数据连接类型	参数说明
DLI	请见表3-114的“表结构”部分
DWS	请见表3-115的“表结构”部分
MRS Hive	请见表3-116的“表结构”部分

6. 单击“保存”，新建数据表。

新建数据表（DDL 模式）

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-118 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”/“数据开发 > 作业开发”，进入“右侧区域”页面。
3. 在左侧菜单选择  ，单击“数据连接”，目录层级展开至“tables”，右键单击“新建数据表”。
4. 单击“DDL模式建表”，选择如表3-112所示的参数，并在下方的编辑器中输入SQL语句。

表 3-112 数据表参数

参数	说明
数据连接类型	选择数据表所属的数据连接类型。 ● DLI ● DWS ● HIVE
数据连接	选择数据表所属的数据连接。
数据库	选择数据表所属的数据库。

5. 单击“确定”，新建数据表。

查看表详情

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”，进入“右侧区域”页面。
2. 在左侧菜单选择  ，单击“数据连接”，目录层级展开至数据表的名称，右键单击“查看表详情”。
3. 进入数据表详情页面，查看如表3-113所示的数据表信息。

表 3-113 表详情页面

页签名称	说明
表信息	显示数据表的基本信息和存储信息。
字段信息	显示数据表的字段信息。
数据预览	预览数据表的10条记录。
DDL	显示DLI/DWS/MRS Hive数据表的DDL。

查看数据表列详情

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。
2. 在左侧菜单选择  ，展开数据连接目录，在数据表下查看对应的列信息。

删除表详情

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”，进入“右侧区域”页面。
2. 在左侧菜单选择  ，单击“数据连接”，目录层级展开至数据表的名称，右键单击“删除”。

3. 在弹出的“删除数据表”页面，单击“确定”，删除数据表。

参数说明

表 3-114 DLI 数据表

参数	是否必选	说明
基本属性		
表名	是	数据表的名称。只能包含英文小写字母、数字、“_”，不能为纯数字，不能以“_”开头，且长度为1~63个字符。
别名	否	数据表的别名，只能包含中文字符、英文字母、数字、“_”，不能为纯数字，不能以“_”开头，且长度为1~63个字符。
数据连接	是	选择数据表所属的数据连接。
数据库	是	选择数据表所属的数据库。
数据位置	是	选择数据存储的位置： ● OBS ● DLI
数据格式	是	选择数据的格式。“数据位置”为“OBS”时，配置该参数。 ● parquet：支持读取不压缩、snappy压缩、gzip压缩的parquet数据。 ● csv：支持读取不压缩、gzip压缩的csv数据。 ● orc：支持读取不压缩、snappy压缩的orc数据。 ● json：支持读取不压缩、gzip压缩的json数据。
路径	是	选择数据存储的OBS路径。“数据位置”为“OBS”时，配置该参数。
表描述	否	数据表的描述信息。
表结构		
列名	是	填写列名，列名不能重复。
类型	是	选择数据类型。
列描述	否	填写列的描述信息。
操作	否	单击  ，增加列。

表 3-115 DWS 数据表

参数	是否必选	说明
基本属性		
表名	是	数据表的名称。只能包含英文字母、数字、“_”，不能为纯数字，不能以“_”开头，且长度为1~63个字符。
别名	否	数据表的别名，只能包含中文字符、英文字母、数字、“_”，不能为纯数字，不能以“_”开头，且长度为1~63个字符。
数据连接	是	选择数据表所属的数据连接。
数据库	是	选择数据表所属的数据库。
模式	是	选择数据库的模式。
表描述	否	数据表的描述信息。
高级选项	否	提供以下高级选项： ● 选择数据表的存储方式 – 行存模式 – 列存模式 ● 选择数据表的压缩级别 – 行存模式：压缩级别的有效值为YES/NO。 – 列存模式：压缩级别的有效值为YES/NO/LOW/MIDDLE/HIGH，还可以配置列存模式同一压缩级别下不同的压缩水平0-3（数值越大，表示同一压缩级别下压缩比越大）。
表结构		
列名	是	填写列名，列名不能重复。

参数	是否必选	说明
数据分类	是	选择数据类型的类别： ● 数值类型 ● 货币类型 ● 布尔类型 ● 二进制类型 ● 字符类型 ● 时间类型 ● 几何类型 ● 网络地址类型 ● 位串类型 ● 文本搜索类型 ● UUID类型 ● JSON类型 ● 对象标识符类型
类型	是	选择数据类型。
列描述	否	填写列的描述信息。
是否建ES索引	否	单击复选框时，表示需要建立ES索引。建立ES索引时，请同时在“CloudSearch集群名”中选择建立好的CSS集群。如何创建CSS集群，请参见《云搜索服务用户指南》。
ES索引数据类型	否	选择ES索引的数据类型： ● text ● keyword ● date ● long ● integer ● short ● byte ● double ● boolean ● binary
操作	否	单击  ，增加列。

表 3-116 MRS Hive 数据表

参数	是否必选	说明
基本属性		
表名	是	数据表的名称。只能包含英文小写字母、数字、“_”，不能为纯数字，不能以“_”开头，且长度为1~63个字符。
别名	否	数据表的别名，只能包含中文字符、英文字母、数字、“_”，不能为纯数字，不能以“_”开头，且长度为1~63个字符。
数据连接	是	选择数据表所属的数据连接。
数据库	是	选择数据表所属的数据库。
表描述	否	数据表的描述信息。
表结构		
列名	是	填写列名，列名不能重复。
数据分类	是	选择数据类型的类别： ● 原始类型 ● ARRAY ● MAP ● STRUCT ● UNION
类型	是	选择数据类型，具体说明请参见 **LanguageManual DDL**。
列描述	否	填写列的描述信息。
操作	否	单击  ，增加列。

3.4.3 脚本开发

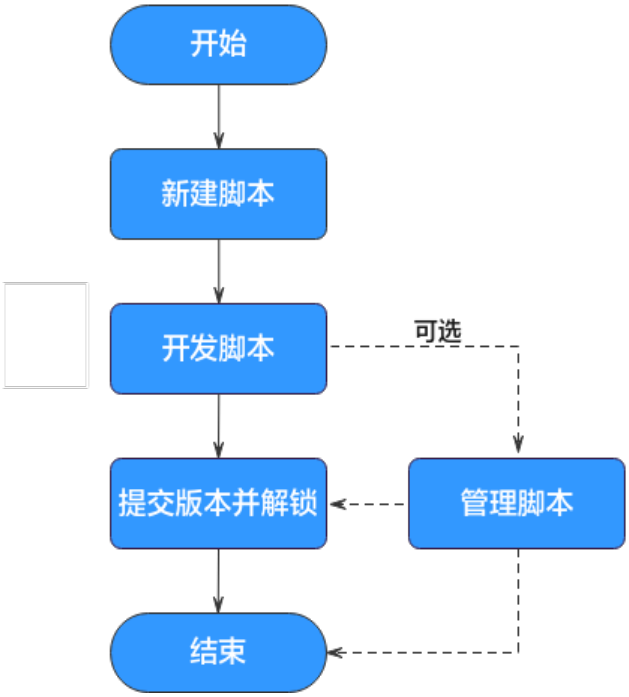
3.4.3.1 脚本开发流程

脚本开发功能提供如下能力：

- 提供在线脚本编辑器，支持进行SQL、Shell、Python等脚本在线代码开发和调测。
- 支持导入和导出脚本。
- 支持使用变量和函数。
- 提供编辑锁定能力，支持多人协同开发场景。
- 支持脚本的版本管理能力。

脚本开发的使用流程如下：

图 3-119 脚本开发流程



1. 新建脚本：新建相应类型的脚本。具体请参见**新建脚本**。
2. 开发脚本：基于新建的脚本，进行脚本的在线开发、调试和执行。具体请参见**开发脚本**。
3. 提交版本并解锁：脚本开发完成后，您需要提交版本并解锁，提交版本并解锁后才能正式地被作业调度运行，便于其他开发者修改。具体请参见**提交版本并解锁**。
4. （可选）管理脚本：脚本开发完成后，您可以根据需要，进行脚本管理。具体请参见**（可选）管理脚本**。

3.4.3.2 新建脚本

数据开发模块的脚本开发功能支持在线编辑、调试、执行脚本，开发脚本前请先新建脚本。

数据开发模块目前支持新建以下几种脚本，用户可根据需要新建相应的脚本。

- DLI SQL脚本
- Hive SQL脚本
- DWS SQL脚本
- Spark SQL脚本
- Flink SQL脚本
- RDS SQL脚本
- Presto SQL脚本
- Shell脚本

- Python脚本

前提条件

已完成**新建数据连接**和**新建数据库**等操作。

操作步骤

新建目录（可选，如果已存在可用的目录，可以不用新建目录）

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-120 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中，右键单击目录名称，选择“新建目录”。
4. 在弹出的“新建目录”页面，配置如**表3-117**所示的参数。

   表 3-117 脚本目录参数

参数	说明
目录名称	脚本目录的名称，只能包含英文字母、数字、中文字符、“_”、“-”，且长度为1~64个字符。
选择目录	选择该脚本目录的父级目录，父级目录默认为根目录。

5. 单击“确定”，新建目录。

新建脚本

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
2. 新建脚本的方式有如下两种：
   方式一：在“右侧区域”，选择并单击相应的脚本类型，新建脚本。

方式二：在脚本目录中，右键单击目录名称，选择新建相应的脚本。

3. 进入脚本开发页面，具体操作请参见**开发SQL脚本**、**开发Shell脚本**、**开发Python脚本**。

说明

当前最多支持创建5个同类型的临时脚本。当关闭了临时未保存的脚本，再次新建同类型的脚本时，会打开上次未保存的临时脚本。

3.4.3.3 开发脚本

3.4.3.3.1 开发 SQL 脚本

对SQL脚本进行在线开发、调试和执行，开发完成的脚本也可以在作业中执行（请参见**开发作业**）。

前提条件

- 已开通相应的云服务并在云服务中创建数据库。Flink SQL脚本不涉及该操作。
- 已创建与脚本的数据连接类型匹配的数据连接，请参见**新建数据连接**。Flink SQL脚本不涉及该操作。
- 当前用户已锁定该脚本，否则需要通过“抢锁”锁定脚本后才能继续开发脚本。新建或导入脚本后默认被当前用户锁定，详情参见**编辑锁定功能**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-121 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中，双击脚本名称，进入脚本开发页面。
4. 在编辑器上方，选择如**表3-118**所示的属性。创建Flink SQL脚本时请跳过此步骤。

表 3-118 SQL 脚本属性

属性	说明
数据连接	选择数据连接。
数据库	选择数据库。
资源队列	选择执行DLI作业的资源队列。当脚本为DLI SQL时，配置该参数。 如需新建资源队列，请参考以下方法： ● 单击⊕，进入DLI的“队列管理”页面新建资源队列。 ● 前往DLI管理控制台进行新建。 **说明** DLI提供默认资源队列“default”，该资源队列不支持insert、load、cat命令。 如需以“key/value”的形式设置提交SQL作业的属性，请单击⚙。最多可设置10个属性，属性说明如下： ● dli.sql.autoBroadcastJoinThreshold（自动使用BroadcastJoin的数据量阈值） ● dli.sql.shuffle.partitions（指定Shuffle过程中Partition的个数） ● dli.sql.cbo.enabled（是否打开CBO优化策略） ● dli.sql.cbo.joinReorder.enabled（开启CBO优化时，是否允许重新调整join的顺序） ● dli.sql.multiLevelDir.enabled（OBS表的指定目录或OBS表分区表的分区目录下有子目录时，是否查询子目录的内容；默认不查询） ● dli.sql.dynamicPartitionOverwrite.enabled（在动态分区模式时，只会重写查询中的数据涉及的分区，未涉及的分区不删除）

5. 在编辑器中输入SQL语句（支持输入多条SQL语句）。

说明

- 需要注意，使用SQL语句获取的系统日期和通过数据库工具获取的系统日期是不一样，查询结果存到数据库是以YYYY-MM-DD格式，而页面显示查询结果是经过转换后的格式。
- SQL语句之间以“;”分隔。如果其它地方使用“;”，请通过“\”进行转义。例如：

```
select 1;
select * from a where b="dsfa\;";  --example 1\;example 2.
```

为了方便脚本开发，数据开发模块提供了如下能力：

– 脚本编辑器支持使用如下快捷键，以提升脚本开发效率。

  - Ctrl + /：注释或解除注释光标所在行或代码块
  - Ctrl + S：保存
  - Ctrl + Z：撤销

- Ctrl + Y：重做
- Ctrl + F：查找
- Ctrl + Shift + R：替换
- Ctrl + X：剪切，光标未选中时剪切一行
- Alt + 鼠标拖动：列模式编辑，修改一整块内容
- Ctrl + 鼠标点选：多列模式编辑，多行缩进
- Shift + Ctrl + K：删除当前行
- Ctrl + →或Ctrl + ←：向右或向左按单词移动光标
- Ctrl + Home或Ctrl + End：移至当前文件的最前或最后
- Home或End：移至当前行最前或最后
- Ctrl + Shift + L：鼠标双击相同的字符串后，为所有相同的字符串添加光标，实现批量修改

– 支持系统函数功能（当前Flink SQL、Spark SQL、ClickHouse SQL、Presto SQL不支持该功能）。

  单击编辑器右侧的“系统函数”，显示该数据连接类型支持的函数，您可以双击函数到编辑器中使用。

– 支持可视化读取数据表生成SQL语句功能（当前Flink SQL、Spark SQL、ClickHouse SQL、Presto SQL不支持该功能）。

  单击编辑器右侧的“数据表”，显示当前数据库或schema下的所有表，可以根据您的需要勾选数据表和对应的列名，在右下角点击“生成SQL语句”，生成的SQL语句需要您手动格式化。

– 支持脚本参数（当前仅Flink SQL不支持该功能）。

  在SQL语句中直接写入脚本参数，调试脚本时可以在脚本编辑器下方输入参数值。如果脚本被作业引用，在作业开发页面可以配置参数值，参数值支持使用EL表达式（参见**表达式概述**）。

  脚本示例如下，其中str1是参数名称，只支持英文字母、数字、“-”、“_”、“<”和“>”，最大长度为16字符，且参数名称不允许重名。

```
select ${str1} from data;
```

  另外，对于MRS Spark SQL和MRS Hive SQL脚本的运行程序参数，除了在SQL脚本中参考语句“set hive.exec.parallel=true;”配置参数，也可以在对应作业节点属性的“运行程序参数”中配置该参数。

图 3-122 运行程序参数

- 支持设置脚本责任人

  单击编辑器右侧的“脚本基本信息”，可设置脚本的责任人和描述信息。

6. （可选）在编辑器上方，单击“格式化”，格式化SQL语句。创建Flink SQL脚本请跳过此步骤。
7. 在编辑器上方，单击“运行”。如需单独执行某部分SQL语句，请选中SQL语句再运行。SQL语句运行完成后，在编辑器下方可以查看脚本的执行历史、执行结果。Flink SQL脚本不涉及，请跳过该步骤。

   说明

   - 对于执行结果支持如下操作：
     - 重命名：可通过双击执行结果页签的名称进行重命名，也可通过右键单击执行结果页签的名称，单击重命名。重命名不能超过16个字符。
     - 可通过右键单击执行结果页签的名称关闭当前页签、关闭左侧页签、关闭右侧页签、关闭其它页签、关闭所有页签。
   - MRS集群为非安全集群、且未限制命令白名单时，在Hive SQL执行过程中，添加application name信息后，则可以方便的根据脚本名称与执行时间在MRS的Yarn管理界面中根据job name找到对应任务。需要注意若默认引擎为tez，则要显式配置引擎为mr，使tez引擎下不生效。

8. 在编辑器上方，单击，保存脚本。

   如果脚本是新建且未保存过的，请配置如表3-119所示的参数。

表 3-119 保存脚本

参数	是否必选	说明
脚本名称	是	脚本的名称，只能包含字符：英文字母、数字、中文、中划线、下划线和点号，且长度小于等于128个字符。
责任人	否	为该脚本指定责任人。默认为创建脚本的人为责任人。
描述	否	脚本的描述信息。
选择目录	是	选择脚本所属的目录，默认为根目录。

> **说明**
>
> 如果脚本未保存，重新打开脚本时，可以从本地缓存中恢复脚本内容。

下载或转储脚本执行结果

约束限制：转储脚本执行结果功能依赖于OBS服务，如无OBS服务，则不支持该功能。

脚本运行成功后，您可以在执行结果页签下载或转储执行结果，仅支持具有拥有DAYU Administrator或Tenant Administrator权限的用户下载和转储。

- 下载结果：下载CSV格式的结果文件到本地。
- 转储结果：转储CSV格式的结果文件到OBS中，请参见**表3-120**。

> **说明**
>
> Flink SQL脚本、RDS SQL脚本、Shell脚本的执行结果，不支持转储。

表 3-120 转储结果

参数	是否必选	说明
数据格式	是	目前仅支持导出CSV格式的结果文件。
资源队列	否	选择执行导出操作的DLI队列。当脚本为DLI SQL时，配置该参数。
压缩格式	否	选择压缩格式。当脚本为DLI SQL时，配置该参数。 • none • bzip2 • deflate • gzip
存储路径	是	设置结果文件的OBS存储路径。选择OBS路径后，您需要在选择的路径后方自定义一个文件夹名称，系统将在OBS路径下创建文件夹，用于存放结果文件。
覆盖类型	否	如果“存储路径”中，您自定义的文件夹在OBS路径中已存在，选择覆盖类型。当脚本为DLI SQL时，配置该参数。 • 覆盖：删除OBS路径中已有的重名文件夹，重新创建自定义的文件夹。 • 存在即报错：系统返回错误信息，退出导出操作。

3.4.3.3.2 开发 Shell 脚本

对Shell脚本进行在线开发、调试和执行，开发完成的脚本也可以在作业中执行（请参见**开发作业**）。

前提条件

- 已新增Shell脚本，请参见**新建脚本**。
- 已新建主机连接，该主机用于执行Shell脚本，请参见**表3-11**。
- 当前用户已锁定该脚本，否则需要通过“抢锁”锁定脚本后才能继续开发脚本。新建或导入脚本后默认被当前用户锁定，详情参见**编辑锁定功能**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-123 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中，双击脚本名称，进入脚本开发页面。
4. 在编辑器上方，配置如**表3-121**所示的属性。

   表 3-121 Shell 脚本属性

参数	说明	示例
主机连接	选择执行Shell脚本的主机。	-

参数	说明	示例
参数	填写执行Shell脚本时，向脚本传递的参数。多个参数之间使用空格分隔，例如：a b c。 此处的“参数”需要在Shell脚本中使用位置变量（如$1，$2，$3）引用，否则配置无效。位置变量由0开始，其中0变量预留用来保存实际脚本的名字，1变量对应脚本的第1个参数，依次类推。如$1、$2、$3分别引用参数a、参数b和参数c。 注意：shell脚本中若引用变量请直接使用$args格式，不要使用${args}格式，否则会导致被作业中同名参数替换。	例如参数输入为“a b c”，执行如下shell脚本，执行结果显示为“b”。 `echo $2`

交互式输入	填写交互式参数，即执行Shell脚本的过程中，需要用户输入的交互式信息（例如密码）。	例如执行如下交互式shell脚本， 交互参数1、2、3 分别对应 begin、end、exit。
		● 当交互参数输入1时，执行结果显示为“start something”。
		● 当交互参数输入2时，执行结果显示为“stop something”。
		● 当交互参数输入3时，执行结果显示为“exit”。
		`#!/bin/bash`
		`select Actions in "begin" "end" "exit"`
		`do`
		`case $Actions in`
		`"begin")`
		`echo "start something"`
		`break`
		`;;`
		`"end")`
		`echo "stop something"`
		`break`
		`;;`
		`"exit")`
		`echo "exit"`
		`break`
		`;;`
		`*)`
		`echo "Ignorant"`
		`;;`
		`esac`
		`done`

5. 在编辑器中编辑Shell语句。为了方便脚本开发，数据开发模块提供了如下能力：
   - 脚本编辑器支持使用如下快捷键，以提升脚本开发效率。
     - Ctrl + /：注释或解除注释光标所在行或代码块

- Ctrl + S：保存
- Ctrl + Z：撤销
- Ctrl + Y：重做
- Ctrl + F：查找
- Ctrl + Shift + R：替换
- Ctrl + X：剪切，光标未选中时剪切一行
- Alt + 鼠标拖动：列模式编辑，修改一整块内容
- Ctrl + 鼠标点选：多列模式编辑，多行缩进
- Shift + Ctrl + K：删除当前行
- Ctrl + →或Ctrl + ←：向右或向左按单词移动光标
- Ctrl + Home或Ctrl + End：移至当前文件的最前或最后
- Home或End：移至当前行最前或最后
- Ctrl + Shift + L：鼠标双击相同的字符串后，为所有相同的字符串添加光标，实现批量修改

– 支持脚本参数功能，使用方法如下：

i. 在Shell语句中直接写入脚本参数名称和参数值。当Shell脚本被作业引用时，如果作业配置的参数名称与Shell脚本的参数名称相同，Shell脚本的参数值将被作业的参数值替换。

脚本示例如下：

```
a=1
echo ${a}
```

其中，a是参数名称，只支持英文字母、数字、“-”、“_”、“<”和“>”，最大长度为16字符，且参数名称不允许重名。

ii. 在编辑器上方配置参数，在执行Shell脚本时，参数会向脚本传递。参数之间使用空格分隔，例如：a b c。此处的“参数”需要在Shell脚本中引用，否则配置无效。

注意：shell脚本中若引用变量请直接使用$args格式，不要使用${args}格式，否则会导致被作业中同名参数替换。

– 支持设置脚本责任人

单击编辑器右侧的“脚本基本信息”，可设置脚本的责任人和描述信息。

6. 在编辑器上方，单击“运行”。Shell语句运行完成后，在编辑器下方可以查看脚本的执行历史和执行结果。

说明

对于执行结果支持如下操作：

- 重命名：可通过双击执行结果页签的名称进行重命名，也可通过右键单击执行结果页签的名称，单击重命名。重命名不能超过16个字符。
- 可通过右键单击执行结果页签的名称关闭当前页签、关闭左侧页签、关闭右侧页签、关闭其它页签、关闭所有页签。

7. 在编辑器上方，单击，保存脚本。
   如果脚本是新建且未保存过的，请配置如表3-122所示的参数。

   表 3-122 保存脚本

参数	是否必选	说明
脚本名称	是	脚本的名称，只能包含字符：英文字母、数字、中文、中划线、下划线和点号，且长度小于等于128个字符。
描述	否	脚本的描述信息。
选择目录	是	选择脚本所属的目录，默认为根目录。

   说明

   如果脚本未保存，重新打开脚本时，可以从本地缓存中恢复脚本内容。

3.4.3.3.3 开发 Python 脚本

对Python脚本进行在线开发、调试和执行，开发完成的脚本也可以在作业中执行（请参见开发作业）。

前提条件

- 已新增Python脚本，请参见新建脚本。
- 已新建主机连接，该主机配有用于执行Python脚本的环境。新建主机连接请参见表3-11。
- 当前用户已锁定该脚本，否则需要通过“抢锁”锁定脚本后才能继续开发脚本。新建或导入脚本后默认被当前用户锁定，详情参见编辑锁定功能。

约束限制

Python脚本暂不支持脚本参数及作业参数。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-124 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中，双击脚本名称，进入脚本开发页面。
4. 在编辑器上方，配置执行Python脚本的主机连接。
5. 在编辑器中编辑Python语句。为了方便脚本开发，数据开发模块提供了如下能力：
   - 脚本编辑器支持使用如下快捷键，以提升脚本开发效率。
     - Ctrl + /：注释或解除注释光标所在行或代码块
     - Ctrl + S：保存
     - Ctrl + Z：撤销
     - Ctrl + Y：重做
     - Ctrl + F：查找
     - Ctrl + Shift + R：替换
     - Ctrl + X：剪切，光标未选中时剪切一行
     - Alt + 鼠标拖动：列模式编辑，修改一整块内容
     - Ctrl + 鼠标点选：多列模式编辑，多行缩进
     - Shift + Ctrl + K：删除当前行
     - Ctrl + →或Ctrl + ←：向右或向左按单词移动光标
     - Ctrl + Home或Ctrl + End：移至当前文件的最前或最后
     - Home或End：移至当前行最前或最后
     - Ctrl + Shift + L：鼠标双击相同的字符串后，为所有相同的字符串添加光标，实现批量修改

- 支持设置脚本责任人

  单击编辑器右侧的“脚本基本信息”，可设置脚本的责任人和描述信息。

6. 在编辑器上方，单击“运行”。Python语句运行完成后，在编辑器下方可以查看脚本的执行历史和执行结果。

   说明

   对于执行结果支持如下操作：

   - 重命名：可通过双击执行结果页签的名称进行重命名，也可通过右键单击执行结果页签的名称，单击“重命名”。重命名不能超过16个字符。
   - 可通过右键单击执行结果页签的名称关闭当前页签、关闭左侧页签、关闭右侧页签、关闭其它页签、关闭所有页签。

7. 在编辑器上方，单击  ，保存脚本。

   如果脚本是新建且未保存过的，请配置如表3-123所示的参数。

   表 3-123 保存脚本

参数	是否必选	说明
脚本名称	是	脚本的名称，只能包含字符：英文字母、数字、中文、中划线、下划线和点号，且长度小于等于128个字符。
描述	否	脚本的描述信息。
选择目录	是	选择脚本所属的目录，默认为根目录。

   说明

   如果脚本未保存，重新打开脚本时，可以从本地缓存中恢复脚本内容。

3.4.3.4 提交版本并解锁

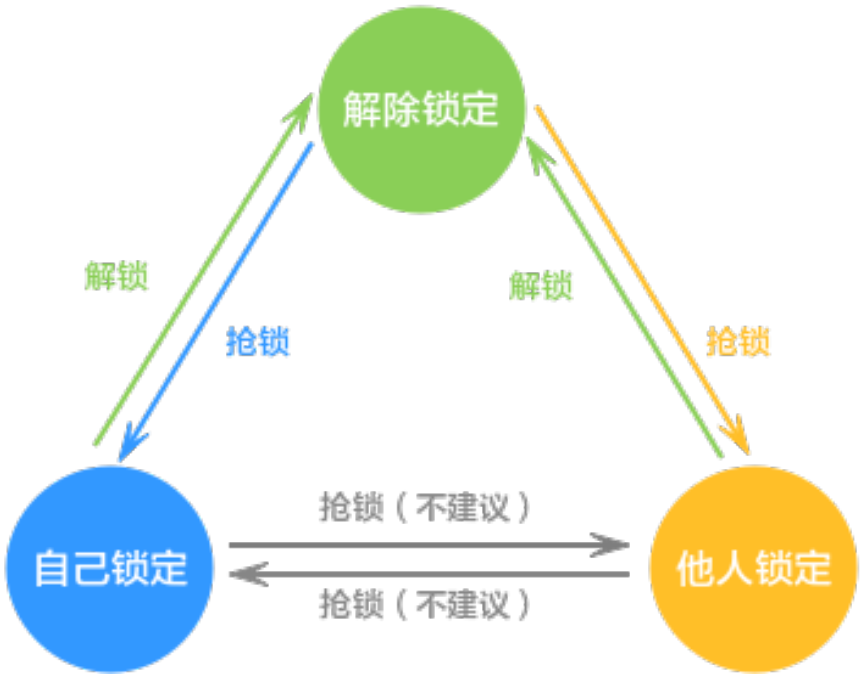
提交版本并解锁，涉及到数据开发的版本管理和编辑锁定功能。

- 版本管理：用于追踪脚本/作业的变更情况，支持版本对比和回滚。系统最多保留最近10条的版本记录，更早的版本记录会被删除。另外，版本管理还可用于区分开发态和生产态，这两种状态隔离，互不影响。
  - 开发态：未提交版本的脚本/作业为开发态，仅用于个人调试开发。在开发态下，可以随意编辑、保存、运行脚本/作业，不会影响调度中的脚本/作业；另外在作业关联脚本、配置作业依赖时，被关联的脚本/作业均会读取开发态的配置。
  - 生产态：提交后版本的脚本/作业为生产态，用于正式调度。在正式调度中，调用脚本、实例重跑、作业依赖、补数据等场景均是关联脚本/作业最新的已提交版本。
- 编辑锁定：用于避免多人协同开发脚本/作业时产生的冲突。新建或导入脚本/作业后，默认当前用户锁定脚本/作业，只有当前用户自己锁定的脚本/作业才可以直接编辑、保存或提交，通过“解锁”功能可解除锁定；处于解除锁定或他人锁定状态的脚本/作业，必须通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。

须知

- 当前脚本/作业的锁定状态可以通过脚本/作业的目录树查看。
- 对于已被他人锁定状态的脚本/作业，您需要通过重新打开该脚本/作业，查看最近的保存/提交时的内容。已打开的脚本/作业内容不会实时刷新。
- 在DataArts Studio更新编辑锁定功能前已经创建的脚本/作业，在更新后默认为解除锁定状态。您需要通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。
- 抢锁的操作依赖于软硬锁的处理策略。配置软硬锁的策略请参见**配置默认项**。
  - 软锁：忽略当前作业或脚本是否被他人锁定，可以进行抢锁或解锁。
  - 硬锁：若作业或脚本被他人锁定，则需锁定的用户解锁之后，当前使用人方可抢锁，空间管理员或DAYU Administrator可以任意抢锁或解锁。
- 不建议直接抢锁处于他人锁定状态的脚本/作业，这会导致他人的修改丢失。如果您有修改需求，请先联系锁定人将脚本/作业解锁，然后再抢锁。

图 3-125 锁定状态转换图



前提条件

已完成脚本开发任务。

提交版本并解锁

“提交”会将当前开发态的最新脚本保存并提交为版本，并覆盖之前的脚本版本。为了便于后续其他开发者对此脚本进行修改，建议您在“提交”后通过“解锁”解除该脚本锁定。

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

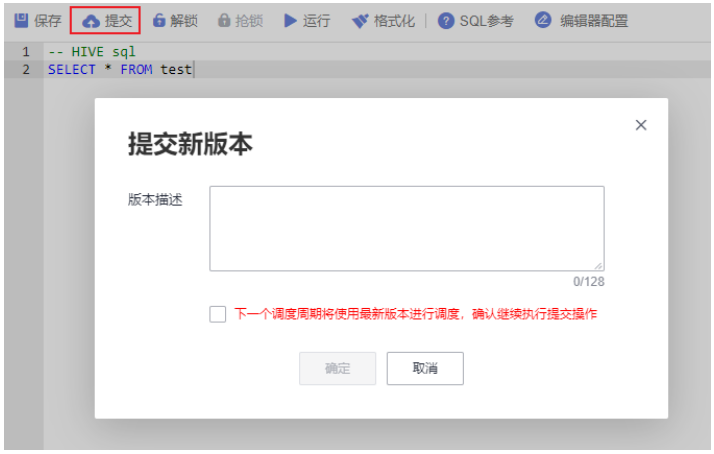
图 3-126 选择数据开发

步骤2 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。

步骤3 在脚本目录中，双击已开发完成的脚本名称，进入脚本开发页面。

步骤4 在脚本编辑器上方单击“提交”，提交版本描述内容长度最多为128个字符，并勾选是否在下个调度周期使用新版本，不勾选则无法点击确认。

图 3-127 提交



步骤5 “提交”后在脚本编辑器上方单击“解锁”，解除锁定，便于后续其他开发者对此脚本进行修改更新。

图 3-128 解锁

----结束

版本回滚

提交版本后，可以在版本列表中看到已经提交过的版本信息（当前最多保存最近10条版本信息）。点击“回滚”，可以回退到任意一个已提交的版本。

回滚内容包括：

- DLI：数据连接、数据库、资源队列、脚本内容。
- DWS：数据连接、数据库、脚本内容。
- HIVE：数据连接、数据库、资源队列、脚本内容。
- SPARK：数据连接、数据库、脚本内容。
- SHELL：主机连接、参数、交互式参数、脚本内容。
- RDS：数据连接、数据库、脚本内容。
- PRESTO：数据连接、模式、脚本内容。
- PYTHON：主机连接、参数、交互式参数、脚本内容。
- FLINK：脚本内容。

操作如下：

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-129 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中，双击脚本名称，进入脚本开发页面。
4. 在页面右侧单击“版本”，查看版本提交记录，找到需要回滚的版本单击“回滚”即可。

   如果当前有开发态的编辑内容没有提交，将会被覆盖。回滚之后需要重新提交才能生效，调度默认使用最新提交的版本进行调度。

图 3-130 版本回滚

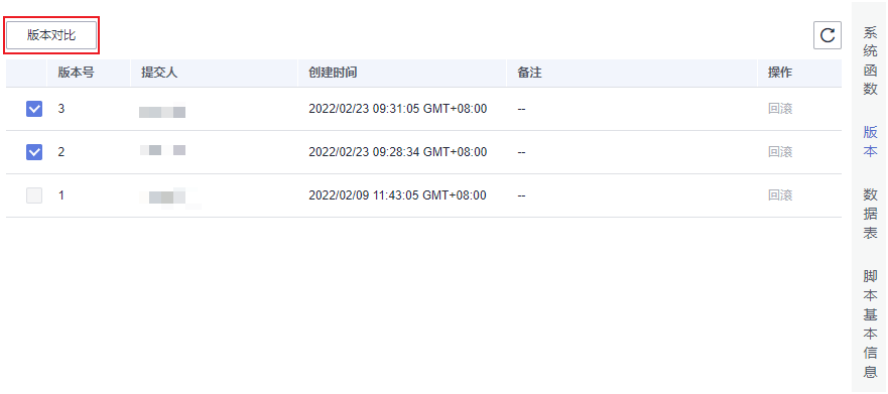
版本对比

支持对比两个不同版本的脚本内容。如果只勾选一个版本，则对比该版本和开发态的脚本内容；如果勾选两个版本，则对比选中的两个版本的脚本内容。

操作如下：

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
2. 在脚本目录中，双击脚本名称，进入脚本开发页面。
3. 在页面右侧单击“版本”，查看版本提交记录，勾选需要对比的版本，单击“版本对比”。

图 3-131 对比版本



4. 单击“版本对比”后，将会打开新窗口，左右两边分别展示出不同版本的脚本内容。两个版本的不同之处将会被标识出来以便用户查看，右上角有上一个不同和下一个不同两个按钮，可以直接跳到上一个或者下一个修改的地方。

图 3-132 版本对比详情

3.4.3.5 （可选）管理脚本

3.4.3.5.1 复制脚本

本章节主要介绍如何复制一个脚本。

前提条件

已完成脚本开发。如何开发脚本，请参见**开发脚本**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-133 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中选择需要要复制的脚本，右键单击脚本名称，选择“拷贝另存为”。
4. 在弹出的“另存为”页面，配置如**表3-124**所示的参数。

表 3-124 脚本目录参数

参数	说明
脚本名称	脚本的名称，只能包含字符：英文字母、数字、中文、中划线、下划线和点号，且长度小于等于128个字符。 **说明** 复制后的脚本名称不能和原脚本名称相同。
选择目录	选择该脚本目录的父级目录，父级目录默认为根目录。

5. 单击“确定”，复制脚本。

3.4.3.5.2 复制名称与重命名脚本

您可以通过复制名称功能复制当前脚本名称，通过重命名功能修改当前脚本名称。

前提条件

已完成脚本开发。如何开发脚本，请参见**开发脚本**。

复制名称

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-134 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中选择需要复制名称的脚本，右键单击脚本名称，选择“复制名称”，即可复制名称到剪贴板。

重命名脚本

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-135 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。

3. 在脚本目录中选择需要重命名的脚本，右键单击脚本名称，选择“重命名”。

   说明

   已经打开了的脚本文件不支持重命名。

4. 在弹出的“重命名脚本名称”页面，配置新脚本名称。

   图 3-136 重命名脚本名称

   表 3-125 重命名脚本参数

参数	说明
脚本名称	脚本的名称，只能包含字符：英文字母、数字、中文、中划线、下划线和点号，且长度小于等于128个字符。

5. 单击“确定”，重命名脚本。

3.4.3.5.3 移动脚本/脚本目录

您可以通过移动功能把脚本文件从当前目录移动到另一个目录，也可以把当前脚本目录移动到另一个目录中。

前提条件

已完成脚本开发。如何开发脚本，请参见**开发脚本**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-137 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 移动脚本或脚本目录。

   方式一：通过右键的“移动”功能。

   a. 在脚本目录中选择需要移动的脚本或脚本文件夹，右键单击脚本或脚本文件夹名称，选择“移动”。
   b. 在弹出的“移动脚本”或“移动目录”页面，配置参数。

      表 3-126 移动脚本/移动目录参数

参数	说明
选择目录	选择脚本或脚本目录要移动到的目录，父级目录默认为根目录。

   c. 单击“确定”，移动脚本/移动目录。

   方式二：通过拖拽的方式。

   单击选中待移动的脚本或脚本文件夹，拖拽至需要移动的目标文件夹松开鼠标即可。

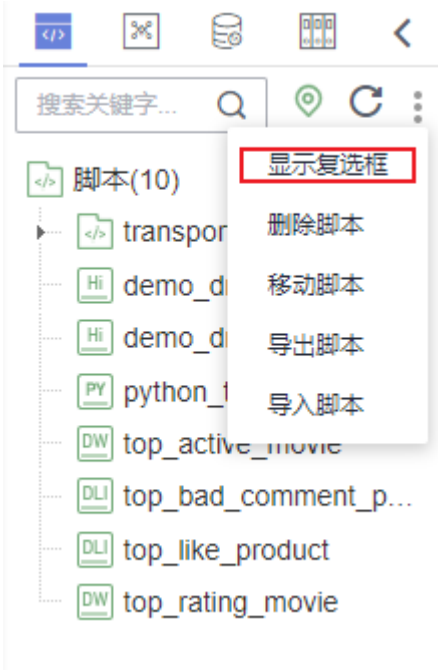
3.4.3.5.4 导出导入脚本

导出脚本

您可以在脚本目录中导出一个或多个脚本文件，导出的为开发态最新的已保存内容。

1. 单击脚本目录中的 ，选择“显示复选框”。

图 3-138 显示脚本复选框



2. 勾选需要导出的脚本，单击 > 导出脚本。导出完成后，即可通过浏览器下载地址，获取到导出的zip文件。

图 3-139 选择并导出脚本

导入脚本

导入脚本功能依赖于OBS服务，如无OBS服务，可从本地导入。

您可以在脚本目录中导入一个或多个脚本文件。导入会覆盖开发态的内容，并自动提交一个新版本。

1. 单击作业目录中的 > 导入脚本，选择已上传至OBS的脚本文件，以及重名处理策略。

   说明

   在硬锁策略下，如果锁在其他人手中，重名策略选择了覆盖，则会覆盖失败。软硬锁策略请参考**配置软硬锁策略**。

   图 3-140 导入脚本

2. 单击“下一步”，根据提示导入脚本。

3.4.3.5.5 查看脚本引用

当用户需要查看某个脚本或者某个文件夹下的所有脚本被引用的情况时，可以参考如下操作查看引用。

前提条件

已完成脚本开发。如何开发脚本，请参见**开发脚本**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-141 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 如要查看某个脚本引用情况，右键单击待查看的脚本，选择“查看引用”，弹出“引用列表”窗口。

   如要查看文件夹下的所有脚本引用情况，右键单击待查看的文件夹，选择“查看引用”，弹出“查看引用”窗口。
4. 在弹出的窗口，可以查看该脚本或该文件夹下所有脚本被引用的情况。

3.4.3.5.6 删除脚本

当用户不需要使用某个脚本时，可以参考如下操作删除该脚本。

删除脚本时会检查脚本被哪个作业引用，引用列表中显示“版本”，表示此脚本被哪些作业版本引用。点击删除时，会删除对应的作业和这个作业的所有版本信息。

说明

如果某一个待删除的脚本正在被作业关联，请确保强制删除脚本后，不影响业务使用。如果希望作业能继续正常使用，请前往作业开发页面，重新关联可用的脚本。

前提条件

删除脚本前，请确保该脚本未被作业使用。

普通删除

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-142 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录中，右键单击脚本名称，选择“删除”。
4. 在弹出的“删除脚本”页面，单击“确认”，删除脚本。

批量删除

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-143 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录顶部，单击，选择“显示复选框”，在脚本目录前出现复选框。
4. 选择需要删除的脚本，再次单击，选择“删除脚本”。
5. 在弹出的“删除脚本”页面，单击“确认”，批量删除脚本。

3.4.3.5.7 迁移脚本责任人

数据开发模块提供了迁移脚本责任人的功能，您可以将责任人A的所有脚本一键迁移到责任人B名下。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-144 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在脚本目录顶部，单击 ，选择“责任人配置”。

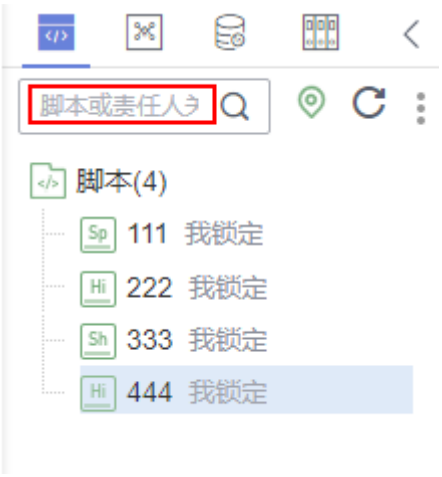
图 3-145 责任人配置

4. 分别设置“当前责任人”和“目标责任人”，单击“迁移”。
5. 提示迁移成功后，单击“关闭”。

相关操作

您还可以根据脚本责任人筛选脚本，在脚本目录上方的搜索框输入责任人，单击放大镜图标，如下图所示。

图 3-146 根据脚本责任人筛选脚本



3.4.3.5.8 批量解锁

数据开发模块提供了批量解锁脚本的功能，您可参照本节内容对锁定的脚本进行批量解锁。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-147 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 单击脚本目录中的 ，选择“显示复选框”。

图 3-148 显示脚本复选框

4. 勾选需要解锁的脚本，单击 > 批量解锁。弹出“解锁成功”提示。

图 3-149 批量解锁

3.4.4 作业开发

3.4.4.1 作业开发流程

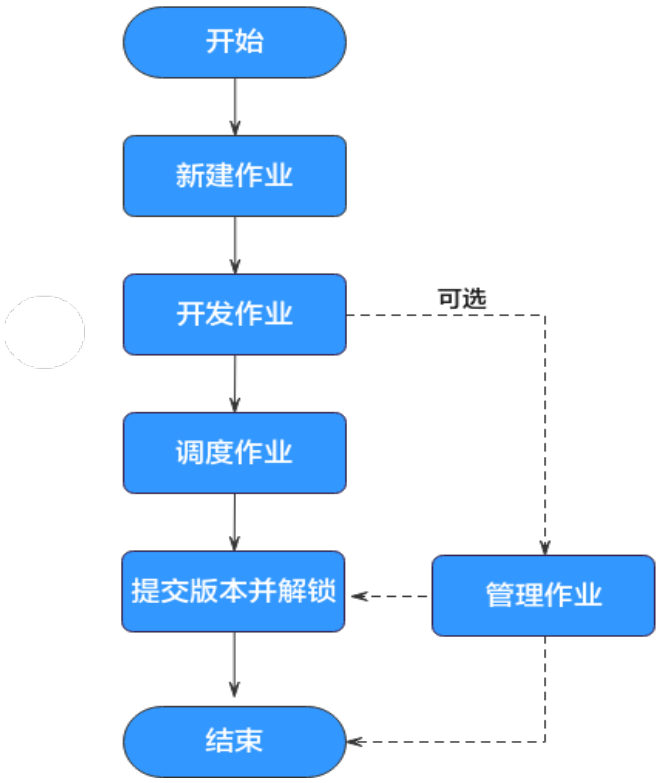
作业开发功能提供如下能力：

- 提供图形化设计器，支持拖拉拽方式快速构建数据处理工作流。
- 预设数据集成、计算&分析、资源管理、数据监控、其他等多种任务类型，通过任务间依赖完成复杂数据分析处理。
- 支持多种作业调度方式。

- 支持导入和导出作业。
- 支持作业状态运维监控和作业结果通知。
- 提供编辑锁定能力，支持多人协同开发场景。
- 支持作业的版本管理能力。

开发作业前，您可以通过**图3-150**了解数据开发模块作业开发的基本流程。

图 3-150 作业开发流程



1. 新建作业：当前提供两种作业类型：批处理和实时处理，分别应用于批量数据处理和实时连接性数据处理，具体请参见**新建作业**。
2. 开发作业：基于新建的作业，进行作业开发，您可以进行编排、配置节点。具体请参见**开发作业**。
3. 调度作业：配置作业调度任务。具体请参见**调度作业**。
   - 如果您的作业是批处理作业，您可以配置作业级别的调度任务，即以作业为一个整体进行调度，支持单次调度、周期调度、事件驱动调度三种调度方式。具体请参见**配置作业调度任务（批处理作业）**。
   - 如果您的作业是实时处理作业，您可以配置节点级别的调度任务，即每一个节点可以独立调度，支持单次调度、周期调度、事件驱动调度三种调度方式。具体请参见**配置节点调度任务（实时作业）**。
4. 提交版本并解锁：作业调度配置完成后，您需要提交版本并解锁，提交版本并解锁后才能用于调度运行，便于其他开发者修改。具体请参见**提交版本并解锁**。
5. （可选）管理作业：作业开发完成后，您可以根据需要，进行作业管理。具体请参见**（可选）管理作业**。

3.4.4.2 新建作业

作业由一个或多个节点组成，共同执行以完成对数据的一系列操作。开发作业前请先新建作业。

前提条件

作业在每工作空间的最大配额为10000，请确保当前作业的数量未达到最大配额。

新建目录（可选）

如果已存在可用的目录，可以不用新建目录。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-151 选择数据开发

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”。
3. 在作业目录中，右键单击目录名称，选择“新建目录”。
4. 在弹出的“新建目录”页面，配置如表3-127所示的参数。

   表 3-127 作业目录参数

参数	说明
目录名称	作业目录的名称，只能包含英文字母、数字、中文字符、“_”、“-”，且长度为1~64个字符。
选择目录	选择该作业目录的父级目录，父级目录默认为根目录。

5. 单击“确定”，新建目录。

新建作业

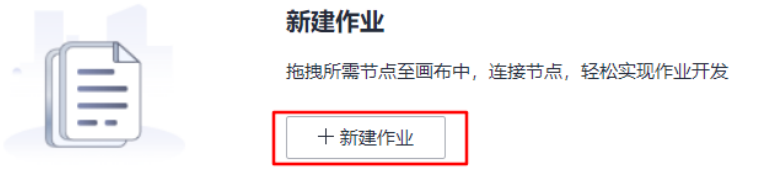
默认作业的最大配额是10000，请确保当前作业的数量未达到最大配额。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-152 选择数据开发

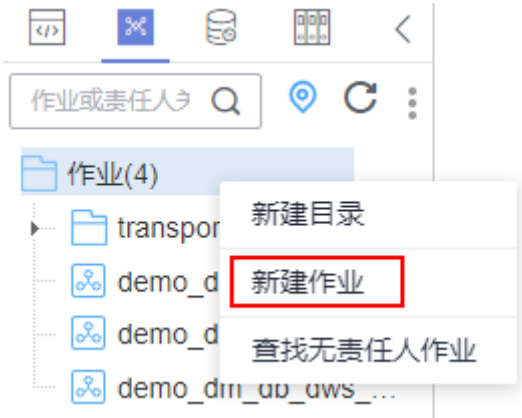
2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
3. 新建作业的方式有如下两种：
   方式一：在“作业开发”界面中，单击“新建作业”。

图 3-153 新建作业（方式一）



方式二：在作业目录中，右键单击目录名称，选择“新建作业”。

图 3-154 新建作业（方式二）



4. 在弹出的“新建作业”页面，配置如**表3-128**所示的参数。

表 3-128 作业参数

参数	说明
作业名称	自定义作业的名称，只能包含英文字母、数字、中文、“-”、“_”、“.”，且长度为1～128个字符。
作业类型	选择作业的类型。 ● 批处理作业：按调度计划定期处理批量数据，主要用于实时性要求低的场景。批作业是由一个或多个节点组成的流水线，以流水线作为一个整体被调度。被调度触发后，任务执行一段时间必须结束，即任务不能无限时间持续运行。 批处理作业可以配置作业级别的调度任务，即以作业为一整体进行调度，具体请参见**配置作业调度任务（批处理作业）**。 ● 实时处理作业：处理实时的连续数据，主要用于实时性要求高的场景。实时作业是由一个或多个节点组成的业务关系，每个节点可单独被配置调度策略，而且节点启动的任务可以永不下线。在实时作业里，带箭头的连线仅代表业务上的关系，而非任务执行流程，更不是数据流。 实时处理作业可以配置节点级别的调度任务，即每一个节点可以独立调度，具体请参见**配置节点调度任务（实时作业）**。
创建方式	选择作业的创建方式。 ● 创建空作业：创建一个空的作业。 ● 基于模板创建：使用数据开发模块提供的模板来创建。
选择目录	选择作业所属的目录，默认为根目录。
责任人	填写该作业的责任人。
作业优先级	选择作业的优先级，提供高、中、低三个等级。
委托配置	配置委托后，作业执行过程中，以委托的身份与其他服务交互。若该工作空间已配置过委托，参见**配置工作空间级委托**，则新建的作业默认使用该工作空间级委托。您也可参见**配置作业级委托**，修改为作业级委托。 **说明** 作业级委托优先于工作空间级委托。
日志路径	选择作业日志的OBS存储路径。日志默认存储在以dlf-log-{Projectid}命名的桶中。 **说明** ● 若您想自定义存储路径，请参见**（可选）修改作业日志存储路径**选择您已在OBS服务侧创建的桶。 ● 请确保您已具备该参数所指定的OBS路径的读、写权限，否则系统将无法正常写日志或显示日志。

5. 单击“确定”，创建作业。

3.4.4.3 开发作业

对已新建的作业进行开发和配置。

前提条件

- 已**新建作业**。
- 当前用户已锁定该作业，否则需要通过“抢锁”锁定作业后才能继续开发作业。新建或导入作业后默认被当前用户锁定，详情参见**编辑锁定功能**。

编排作业节点

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-155 选择数据开发

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”。
3. 在作业目录中，双击Pipeline模式批处理作业或实时处理作业的名称，进入作业开发页面。
4. 拖动所需的节点至画布，鼠标移动到节点图标上，选中连线图标并拖动，连接到下一个节点上。

   说明

   每个作业建议最多包含200个节点。

图 3-156 编排作业

5. 配置节点功能。右键单击画布中的节点图标，根据实际需要选择如表3-129所示的功能。

表 3-129 右键节点功能

功能	说明
配置	进入该节点的“节点属性”页面。
删除	支持删除一个节点或同时删除多个节点。 • 单节点删除：右键单击画布中的节点图标，选择删除或按快捷键**Delete**。 • 多节点删除：按下键盘中的**Ctrl**，单击画布中需要删除的节点图标，在当前作业画布空白处单击右键，选择删除或按快捷键**Delete**。
复制	支持复制一个或多个节点至任意作业中： • 单节点复制：右键单击画布中的节点图标，选择复制或按快捷键**Ctrl+C**，在作业画布空白处粘贴节点或按快捷键**Ctrl+V**，复制后的节点携带原节点的配置信息。 • 多节点复制：按下键盘中的**Ctrl**，单击画布中需要复制的节点图标，在当前作业画布空白处单击右键选择复制或按快捷键**Ctrl+C**，在目标作业画布空白处粘贴或按快捷键**Ctrl+V**。复制后的节点携带原节点的配置信息，但不包含节点间的连接关系。
测试运行	测试运行该节点。
从当前节点测试运行	仅在批作业下显示该选项。选择“从当前节点测试运行”，则测试运行当前节点以及后续节点。
添加/删除连线	可以选择为两个不同的节点添加或删除连线，
编辑CDM作业	仅CDM Job节点显示该选项。选择CDM集群和作业后，可以跳转到CDM作业编辑页面，进行作业修改。

功能	说明
查看CDM作业日志	仅CDM Job节点显示该选项。当CDM作业运行后，右键选中CDM Job节点，单击“查看CDM日志”，可以跳转到作业监控页面，查看作业日志打印的详细信息，帮助开发者定界定位作业运行异常原因。
编辑脚本	仅关联了脚本的节点显示该选项。跳转到脚本编辑页面，对关联的脚本进行编辑。
添加便签	为该节点添加便签，每个节点可以有多个便签。

6. （可选）配置连线功能。右键单击画布中的节点间连线，显示“删除”和“设置条件”功能，您可以根据实际需要进行选择。
   - 删除：可以删除节点间的连线。
   - 设置条件：在弹出的窗口中，您可以通过EL表达式语法填写三元表达式。当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

     如下图所示，是一个典型的三元表达式。当“DQM”节点的运行结果为true时，才会执行连线后的节点。当运行结果为false时，如果失败策略为“跳过所有节点”，则该连线后面的节点A以及A后的所有节点均会被跳过。

     关于EL表达式的语法，您可以查看**EL表达式参考**。
7. 请参见**节点概述**配置具体节点的属性。
8. 配置节点属性。单击画布中的节点，在右侧显示“节点属性”页签，默认展开此配置页面，请参见**节点概述**配置具体节点的属性。

配置作业基本信息

为作业配置责任人、优先级信息后，用户可根据责任人、优先级来检索相应的作业。操作方法如下：

单击画布右侧“作业基本信息”页签，展开配置页面，配置如**表3-130**所示的参数。

表 3-130 作业基本信息

参数	说明
作业责任人	自动匹配创建作业时配置的作业责任人，此处支持修改。
执行用户	执行作业的用户。如果输入了执行用户，则作业以执行用户身份执行；如果没有输入执行用户，则以提交作业启动的用户身份执行。
作业委托	配置委托后，作业执行过程中，以委托的身份与其他服务交互。
作业优先级	自动匹配创建作业时配置的作业优先级，此处支持修改。
实例超时时间	配置作业实例的超时时间，设置为0或不配置时，该配置项不生效。如果您为作业设置了异常通知，当作业实例执行时间超过超时时间，将触发异常通知，发送消息给用户。
自定义字段	配置自定义字段的参数名称和参数值。
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。

配置作业参数

作业参数为全局参数，可用于作业中的任意节点。操作方法如下：

单击画布的空白处，在右侧显示“作业参数配置”页签，单击此页签，展开配置页面，配置如**表3-131**所示的参数。

表 3-131 作业参数配置

功能	说明
参数	
新增	单击“新增”，在文本框中填写作业参数的名称和参数值。 • 参数名称 名称只能包含字符：英文字母、数字、中划线和下划线。 • 参数值 – 字符串类的参数直接填写字符串，例如：str1 – 数值类的参数直接填写数值或运算表达式。 参数配置完成后，在作业中的引用格式为：${参数名称}
修改	在参数名和参数值的文本框中直接修改。
掩码显示	在参数值为密钥等情况下，从安全角度，请单击  将参数值掩码显示。
删除	在参数值文本框后方，单击，删除作业参数。

功能	说明
常量	
新增	单击“新增”，在文本框中填写作业常量的名称和参数值。 • 参数名称 名称只能包含字符：英文字母、数字、中划线和下划线。 • 参数值 – 字符串类的参数直接填写字符串，例如：str1 – 数值类的参数直接填写数值或运算表达式。 参数配置完成后，在作业中的引用格式为：${参数名称}
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击 ，删除作业常量。

调测并保存作业

作业编排和配置完成后，请执行以下操作：

批处理作业

步骤1 单击画布上方的测试运行按钮  ，测试作业。

步骤2 测试完成后，单击画布上方的保存按钮  ，保存作业的配置信息。如果测试未通过请按照提示修改后再次运行。

----结束

实时处理作业

步骤1 单击画布上方的保存按钮  ，保存作业的配置信息。

----结束

3.4.4.4 调度作业

对已编排好的作业设置调度方式。

- 如果您的作业是批处理作业，您可以配置作业级别的调度任务，即以作业为一个整体进行调度，支持单次调度、周期调度、事件驱动调度三种调度方式。具体请参见**配置作业调度任务（批处理作业）**。
- 如果您的作业是实时处理作业，您可以配置节点级别的调度任务，即每一个节点可以独立调度，支持单次调度、周期调度、事件驱动调度三种调度方式。具体请参见**配置节点调度任务（实时作业）**。

前提条件

- 已**开发作业**。

- 当前用户已锁定该作业，否则需要通过“抢锁”锁定作业后才能继续开发作业。新建或导入作业后默认被当前用户锁定，详情参见**编辑锁定功能**。

约束限制

- 调度周期需要合理设置，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，从而出现计划时间和开始时间相差大。例如CDM、ETL作业的调度周期至少应在5分钟以上，并根据作业表的数据量、源端表更新频次等调整。
- 如果通过DataArts Studio数据开发调度CDM迁移作业，CDM迁移作业处也配置了定时任务，则两种调度均会生效。为了业务运行逻辑统一和避免调度冲突，推荐您启用数据开发调度即可，无需配置CDM定时任务。

配置作业调度任务（批处理作业）

配置批处理作业的作业调度任务，支持单次调度、周期调度、事件驱动调度三种方式。操作方法如下：

单击画布右侧“调度配置”页签，展开配置页面，配置如**表3-132**所示的参数。

表 3-132 作业调度配置

参数	说明
调度方式	选择作业的调度方式： ● 单次调度：手动触发作业单次运行。 ● 周期调度：周期性自动运行作业，参数说明请参见**表3-133**。 ● 事件驱动调度：根据外部条件触发作业运行，参数说明请参见**表3-134**。
空跑	如果勾选了空跑，任务不会实际执行，将直接返回成功。

表 3-133 “周期调度”的参数配置

参数	说明
生效时间	调度任务的生效时间段。

参数	说明
调度周期	选择调度任务的执行周期，并配置相关参数。 调度周期需要合理设置，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，从而出现计划时间和开始时间相差大。例如CDM、ETL作业的调度周期至少应在5分钟以上，并根据作业表的数据量、源端表更新频次等调整。 ● 分钟：支持在小时整点开始调度运行，调度周期可按间隔时间配置为分钟级别，在当天结束时间结束调度后第二天再自动开始调度。 ● 小时：支持在某一时刻开始调度运行，调度周期可按间隔时间配置为小时级别，在当天结束时间结束调度后第二天再自动开始调度。 ● 天：支持在某天的某一时刻开始调度运行，调度周期为1天。 ● 周：支持在一周中选择一天或多天的某一时刻开始调度运行。 ● 月：支持在一月中选择一天或多天的某一时刻开始调度运行。
依赖作业	选择周期调度作业作为依赖作业，则仅当依赖的作业在某段时间内有实例运行完成时，才开始执行当前作业。当前仅支持通过搜索作业名来选择符合条件的作业为依赖作业。关于设置依赖作业的条件，以及设置依赖作业后的作业运行原理请参见**作业依赖详解**。 另外，依赖作业可以配置为多个作业，对于多个依赖作业，需等到某时间区间（详见**设置依赖作业后的作业运行原理**）内所有依赖作业实例运行完成后，才能开始执行。 约束条件如下： ● 作业A的调度周期不能比依赖作业B小。例如，作业A和作业B同为分钟/小时调度，A的间隔时间小于B的间隔时间，则作业A不能设置作业B为依赖作业；作业A为分钟调度，作业B为小时调度，则作业A不能设置作业B为依赖作业。 ● 作业A和依赖作业B的不能有任一调度周期为周。例如，作业A的调度周期为周或作业B的调度周期为周，则作业A不能设置作业B为依赖作业。 ● 调度周期为月的作业只能依赖调度周期为天的作业。例如，作业A的调度周期为月，则作业A只能设置调度周期为天的作业为依赖作业。

参数	说明
依赖的作业失败后，当前作业处理策略	当依赖的作业在当前作业周期内存在运行失败实例后，选择当前作业的处理策略： ● 挂起 挂起当前作业，挂起的作业会阻塞后续作业的执行。您可以手动将依赖的作业强制成功，解决阻塞问题。 ● 继续执行 继续执行当前作业。 ● 终止执行 终止执行当前作业，当前作业的状态为“取消”。 例如，当前作业调度周期为1小时， 依赖作业调度周期为5分钟。 ● 如果当前参数配置的是终止执行，依赖的作业12个实例中只要有一个失败的，当前作业就终止执行。 ● 如果当前参数配配置的是继续执行，只要依赖的作业12个实例跑完了，当前作业就继续执行。 **说明** 依赖的作业失败后，当前作业处理策略可通过配置默认项进行批量设置，无需每个作业单独设置。具体请参见**配置默认项**。
等待依赖作业的上一周期结束，才能运行	当作业依赖其他作业时，默认情况下等待某时间区间（详见**设置依赖作业后的作业运行原理**）内是否有依赖的作业实例运行完成，然后才执行当前作业。如果依赖的作业实例未成功运行结束，则当前作业为等待运行状态。 当勾选此选项后，检查此时间区间的上一周期区间内是否有作业实例运行完，然后再执行当前作业。
跨周期依赖	选择作业实例之间的依赖关系。 ● 不依赖上一调度周期。此处可以配置并发数，表示多个作业实例并行执行的个数。如果并发数配置为1，前一个批次执行完成后(包括成功、取消、或失败)，下一批次才开始执行。 ● 自依赖（等待上一调度周期结束才能继续运行）。

表 3-134 “事件驱动调度”的参数配置

参数	说明
触发事件类型	选择触发作业运行的事件类型。 ● **“KAFKA”**
“KAFKA”触发事件类型的参数	
连接名称	选择数据连接，需先在“管理中心”创建kafka数据连接。
Topic	选择需要发往kafka的消息Topic。
事件处理并发数	选择作业并行处理的数量，最大并发数为128。

参数	说明
事件检测间隔	配置时间间隔，检测通道下是否有新的消息。时间间隔单位可以配置为秒或分钟。
读取策略	选择数据的读取位置： ● 从上次位置读取：首次启动时，从最新的位置读取数据。后续启动时，则从前一次记录的位置读取数据。 ● 从最新位置读取：每次启动都会从最新的位置读取数据。
失败策略	选择调度失败后的策略： ● 挂起 ● 忽略失败，读取下一个

配置节点调度任务（实时作业）

配置实时处理作业的节点调度任务，支持单次调度、周期调度、事件驱动调度三种方式。操作方法如下：

单击画布中的节点，在右侧显示“调度配置”页签，单击此页签，展开配置页面，配置如表3-135所示的参数。

表 3-135 节点调度配置

参数	说明
调度方式	选择作业的调度方式： ● 单次调度：手动触发作业单次运行。 ● 周期调度：周期性自动运行作业。 ● 事件驱动调度：根据外部条件触发作业运行。
“周期调度”的参数	
生效时间	调度任务的生效时间段。
调度周期	选择调度任务的执行周期，并配置相关参数： ● 分钟 ● 小时 ● 天 ● 周 ● 月 调度周期需要合理设置，如CDM、ETL作业的调度周期至少应在5分钟以上，并根据作业表的数据量、源端表更新频次等调整。
跨周期依赖	选择作业下实例之间的依赖关系。 ● 不依赖上一调度周期 ● 自依赖（等待上一调度周期结束才能继续运行）

参数	说明
“事件驱动调度”的参数	
触发事件类型	选择触发作业运行的事件类型。
连接名称	选择数据连接，需先在“管理中心”创建kafka数据连接。
Topic	选择需要发往kafka的消息Topic。
消费组	消费者组是kafka提供的可扩展且具有容错性的消费者机制。 它是一个组，所以内部有可以有多个消费者，这些消费者共用一个ID，一个组内的所有消费者共同协作，完成对订阅的主题的所有分区进行消费。其中一个主题中的一个分区只能由一个消费者消费。 **说明** 1. 一个消费者组可以有多个消费者。 2. Group ID是一个字符串，在一个kafka集群中，它标识唯一的一个消费者组。 3. 每个消费者组订阅的所有主题中，每个主题的每个分区只能由一个消费者消费。消费者组之间不影响。 当触发事件类型选择了DIS或KAFKA时，会自动关联出消费组的ID，用户也可以手动修改。
事件处理并发数	选择作业并行处理的数量，最大并发数为10。
事件检测间隔	配置时间间隔，检测通道下是否有新的消息。时间间隔单位可以配置为秒或分钟。
失败策略	选择节点执行失败后的策略： ● 挂起 ● 忽略失败，继续调度

3.4.4.5 提交版本并解锁

提交版本并解锁，涉及到数据开发的版本管理和编辑锁定功能。

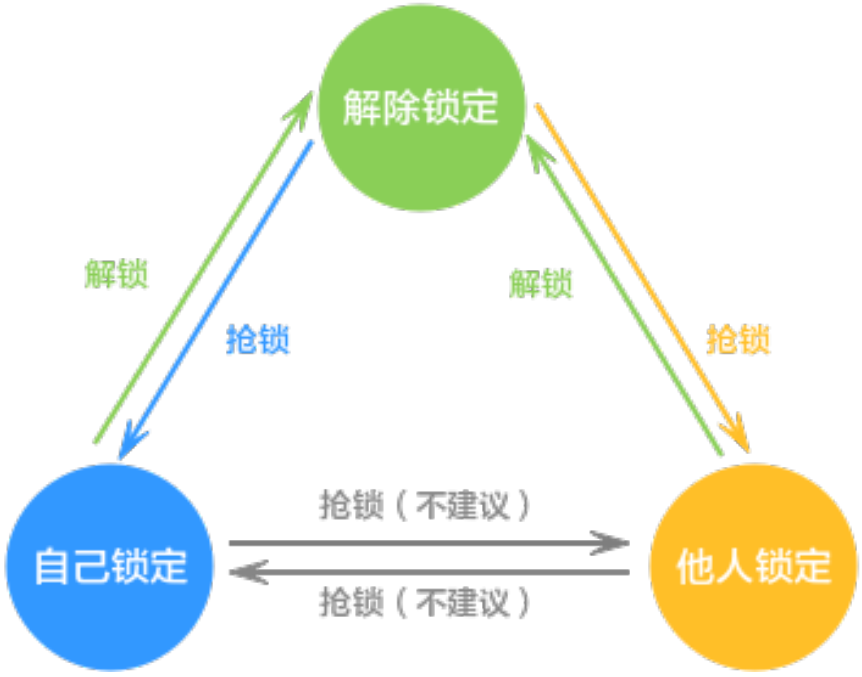
- 版本管理：用于追踪脚本/作业的变更情况，支持版本对比和回滚。系统最多保留最近10条的版本记录，更早的版本记录会被删除。另外，版本管理还可用于区分开发态和生产态，这两种状态隔离，互不影响。
  - 开发态：未提交版本的脚本/作业为开发态，仅用于个人调试开发。在开发态下，可以随意编辑、保存、运行脚本/作业，不会影响调度中的脚本/作业；另外在作业关联脚本、配置作业依赖时，被关联的脚本/作业均会读取开发态的配置。
  - 生产态：提交后版本的脚本/作业为生产态，用于正式调度。在正式调度中，调用脚本、实例重跑、作业依赖、补数据等场景均是关联脚本/作业最新的已提交版本。
- 编辑锁定：用于避免多人协同开发脚本/作业时产生的冲突。新建或导入脚本/作业后，默认当前用户锁定脚本/作业，只有当前用户自己锁定的脚本/作业才可以直接编辑、保存或提交，通过“解锁”功能可解除锁定；处于解除锁定或他人锁定状

态的脚本/作业，必须通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。

须知

- 当前脚本/作业的锁定状态可以通过脚本/作业的目录树查看。
- 对于已被他人锁定状态的脚本/作业，您需要通过重新打开该脚本/作业，查看最近的保存/提交时的内容。已打开的脚本/作业内容不会实时刷新。
- 在DataArts Studio更新编辑锁定功能前已经创建的脚本/作业，在更新后默认为解除锁定状态。您需要通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。
- 抢锁的操作依赖于软硬锁的处理策略。配置软硬锁的策略请参见**配置默认项**。
  - 软锁：忽略当前作业或脚本是否被他人锁定，可以进行抢锁或解锁。
  - 硬锁：若作业或脚本被他人锁定，则需锁定的用户解锁之后，当前使用人方可抢锁，空间管理员或DAYU Administrator可以任意抢锁或解锁。
- 不建议直接抢锁处于他人锁定状态的脚本/作业，这会导致他人的修改丢失。如果您有修改需求，请先联系锁定人将脚本/作业解锁，然后再抢锁。

图 3-157 锁定状态转换图



前提条件

已完成作业开发任务。

提交版本并解锁

“提交”会将当前开发态的最新作业保存并提交为版本，并覆盖之前的作业版本。为了便于后续其他开发者对此作业进行修改，建议您在“提交”后通过“解锁”解除该作业锁定。

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-158 选择数据开发

步骤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”。

步骤3 在作业目录中，双击已开发完成的作业名称，进入作业开发页面。

步骤4 在作业画布上方单击“提交”，提交版本。描述内容长度最多为128个字符，并勾选是否在下个调度周期使用新版本，不勾选则无法点击确认。

图 3-159 提交

步骤5 “提交”后在作业画布上方单击“解锁”，解除锁定，便于后续其他开发者对此作业进行修改更新。

图 3-160 解锁

----结束

版本回滚

用户可以在版本列表中看到已经提交过的版本信息（当前最多保存最近10条版本信息）。点击“回滚”，可以回退到任意一个已提交的版本。

回滚内容包括：

- 作业定义（算子属性、连线等）；
- 作业基本信息、作业调度配置、作业参数、血缘关系中的所有内容；

操作如下：

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-161 选择数据开发

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”。
3. 在作业目录中，双击作业名称，进入作业开发页面。
4. 在页面右侧单击“版本”，查看版本提交记录，找到需要回滚的版本单击“回滚”即可。

图 3-162 版本回滚操作界面

版本详情查看

用户可以在版本列表中看到已经提交过的版本信息。

操作如下：

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-163 选择数据开发

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”。
3. 在作业目录中，双击作业名称，进入作业开发页面。
4. 在页面右侧单击“版本”，查看版本提交记录，找到需要查看详情的版本单击“查看”即可。

   点击查看，将会打开一个新窗口，展示出该版本的作业定义。查看窗口仅用于展示某个版本的作业属性，不可修改任何作业属性。

图 3-164 版本详情查看

版本对比

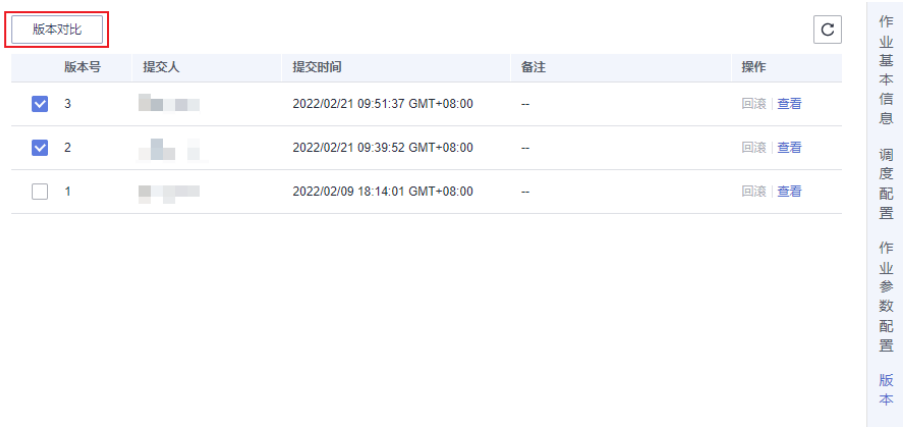
1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-165 选择数据开发

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”。
3. 在作业目录中，双击作业名称，进入作业开发页面。
4. 在页面右侧单击“版本”，查看版本提交记录，勾选需要对比的版本单击“版本对比”即可。

   若只勾选一个版本，则比较选中的版本和开发态的作业属性Json。若勾选两个版本，则比较两个版本的作业属性Json。

图 3-166 对比版本操作界面



3.4.4.6 （可选）管理作业

3.4.4.6.1 复制作业

本章节主要介绍如何复制一份作业。

前提条件

已完成作业开发。如何开发作业，请参见**开发作业**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-167 选择数据开发

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”。
3. 在作业目录中选择需要要复制的作业，右键单击作业名称，选择“拷贝另存为”。
4. 在弹出的“另存为”页面，配置如表3-136所示的参数。

表 3-136 作业目录参数

参数	说明
作业名称	自定义作业的名称，只能包含英文字母、数字、中文、“-”、“_”、“.”，且长度为1～128个字符。
选择目录	选择该作业目录的父级目录，父级目录默认为根目录。

5. 单击“确定”，复制作业。

3.4.4.6.2 复制名称和重命名作业

您可以通过复制名称功能复制当前作业名称，通过重命名功能修改当前作业名称。

前提条件

已完成作业开发。如何开发作业，请参见开发作业。

复制名称

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-168 选择数据开发

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”。
3. 在作业目录中选择需要复制名称的作业，右键单击作业名称，选择“复制名称”，即可复制名称到剪贴板。

重命名作业

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-169 选择数据开发

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”。
3. 在作业目录中选择需要重命名的作业，右键单击作业名称，选择“重命名”。
4. 在弹出的“重命名作业名称”页面，配置新作业名。

表 3-137 重命名作业参数

参数	说明
作业名称	自定义作业的名称，只能包含英文字母、数字、中文、“-”、“_”、“.”，且长度为1～128个字符。

5. 单击“确定”，重命名作业。

3.4.4.6.3 移动作业/作业目录

您可以通过移动功能把作业文件或作业目录从当前目录移动到另一个目录。

前提条件

已完成作业开发。如何开发作业，请参见**开发作业**。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-170 选择数据开发

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”。
3. 移动作业或作业目录。

   方式一：通过右键的“移动”功能。

   a. 在作业目录中选择需要移动的作业或作业文件夹，右键单击作业或作业文件夹名称，选择“移动”。
   b. 在弹出的“移动作业”或“移动目录”页面，配置作业要移动到的目录。

表 3-138 移动作业/作业目录参数

参数	说明
选择目录	选择作业或作业文件夹要移动到的目录，父级目录默认为根目录。

c. 单击“确定”，移动作业。

方式二：通过拖拽的方式。

单击选中待移动的作业或作业文件夹，拖拽至需要移动的目标文件夹松开鼠标即可。

3.4.4.6.4 导出导入作业

- 导出作业，均是导出开发态的最新的已保存内容。
- 导入作业，会覆盖开发态的内容并自动提交一个新版本。

导出作业

方式一：在作业开发页面导出某一个作业

步骤1 双击作业名称，进入某一作业的开发页面，单击画布上方的导出按钮，选择导出作业的类型。

- 只导出作业：导出作业中节点的连接关系，以及各节点的属性配置到本地，不包含密码等敏感信息。导出后，您可以通过浏览器下载内容获取到zip格式的压缩包文件。
- 导出作业及其依赖脚本：导出作业中节点的连接关系、各节点的属性配置以及作业的调度配置、参数配置、依赖的脚本、资源定义到本地，不包含密码等敏感信息。导出后，您可以通过浏览器下载内容获取到zip格式的压缩包文件。

图 3-171 导出作业（方式一）

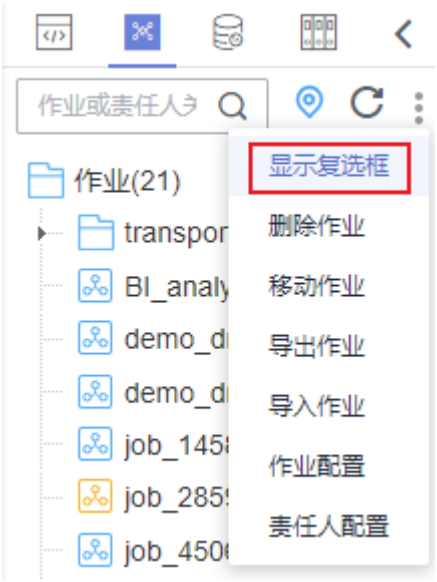
步骤2 单击“确定”，导出所需的作业文件。

----结束

方式二：在作业目录中导出一个或多个作业

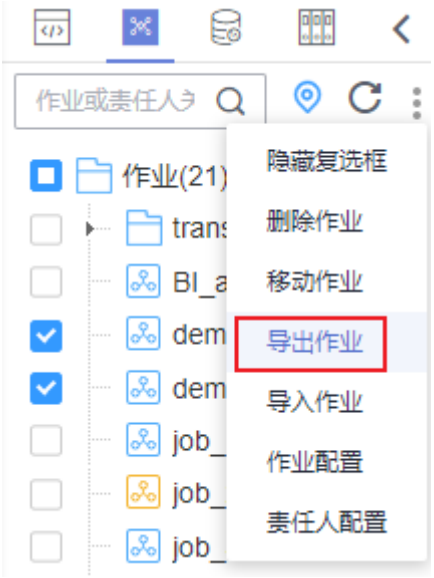
步骤1 单击作业目录中的，选择“显示复选框”。

图 3-172 显示作业复选框



步骤2 勾选需要导出的作业，单击 > 导出作业，可选择“只导出作业”或“导出作业及其依赖脚本和资源定义”。导出完成后，即可通过浏览器下载地址，获取到导出的zip文件。

图 3-173 选择并导出作业



----结束

导入作业

导入作业功能依赖于OBS服务，如无OBS服务，可从本地导入。

在作业目录中导入一个或多个作业

步骤1 单击作业目录中的 > 导入作业，选择已上传至OBS或者本地中的作业文件，以及重名处理策略。

说明

在硬锁策略下，如果锁在其他人手中，重名策略选择了覆盖，则会覆盖失败。软硬锁策略请参考**配置软硬锁策略**。

图 3-174 导入作业定义及依赖

步骤2 单击“下一步”，根据提示导入作业。

说明

在导入作业过程中，若作业关联的数据连接、dli队列、ges图等在数据开发模块系统中不存在时，系统会提示您重新选择。

----结束

操作示例

背景信息：

- 在数据开发模块系统中创建一个DWS的数据连接“doctest”
- 在作业目录中创建实时作业“doc1”，作业中添加节点“DWS SQL”，配置节点的“数据连接”为“doctest”，配置“SQL脚本”和“数据库”。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-175 选择数据开发

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”。
3. 在作业搜索框中搜索作业“doc1”，导出作业到本地，并上传作业至OBS文件夹中。
4. 在数据开发模块系统中删掉作业关联的dws数据连接“doctest”。
5. 单击作业目录中的 > 导入作业，选择上传至OBS文件夹中的作业，并设置重名处理策略。
6. 单击“下一步”，根据导入作业页面的提示重新选择数据连接。
7. 单击“下一步”，再单击“关闭”。

3.4.4.6.5 删除作业

当用户不需要使用某个作业时，可以参考如下操作删除该作业，以减少作业的配额占用。

说明

作业删除后，将无法恢复，请确保删除作业后，不影响业务。

普通删除

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-176 选择数据开发

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”。
3. 在作业目录中，右键单击作业名称，选择“删除”。
4. 在弹出的“删除作业”页面，单击“确定”，删除作业。

批量删除

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-177 选择数据开发

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”。
3. 在作业目录顶部，单击 ，选择“显示复选框”，在作业目录前出现复选框。
4. 选择需要删除的作业，再次单击 ，选择“删除作业”。
5. 在弹出的“删除作业”页面，单击“确定”，批量删除作业。

3.4.4.6.6 迁移作业责任人

数据开发模块提供了迁移作业责任人的功能，您可以将责任人A的所有作业一键迁移到责任人B名下。

操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-178 选择数据开发

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”。
3. 在作业目录顶部，单击，选择“责任人配置”。

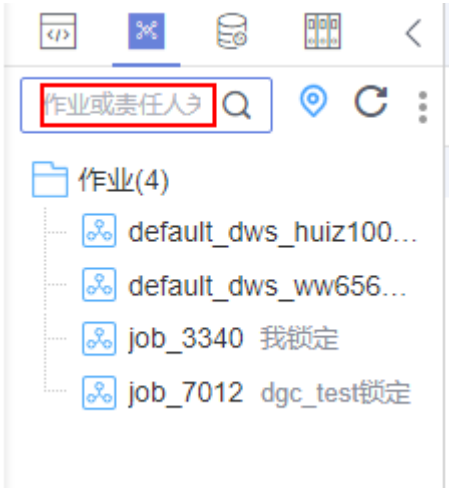
图 3-179 责任人配置

4. 分别设置“当前责任人”和“目标责任人”，单击“迁移”。
5. 提示迁移成功后，单击“关闭”。

相关操作

您还可以根据作业责任人筛选作业，在作业目录上方的搜索框输入责任人，单击放大镜图标，如下图所示。

图 3-180 根据作业责任人筛选作业



3.4.4.6.7 批量解锁

数据开发模块提供了批量解锁作业的功能，您可参照本节内容对锁定的作业进行批量解锁。

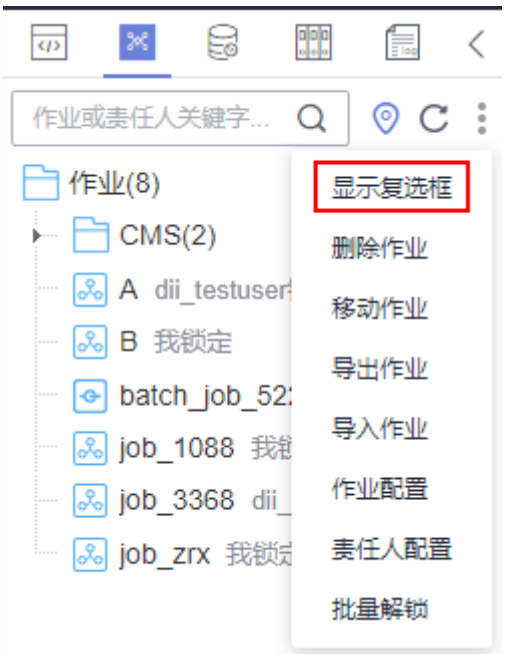
操作步骤

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-181 选择数据开发

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”。
3. 单击作业目录中的，选择“显示复选框”。

图 3-182 显示作业复选框



4. 勾选需要解锁的作业，单击 > 批量解锁。弹出“解锁成功”提示。

图 3-183 批量解锁

3.4.5 解决方案

背景信息

解决方案定位于为用户提供便捷的、系统的方式管理作业，更好地实现业务需求和目标。每个解决方案可以包含一个或多个业务相关的作业，一个作业可以被多个解决方案复用。

数据开发模块目前支持处理以下几种方式的解决方案。

- 新建解决方案
- 编辑解决方案
- 导出解决方案
- 导入解决方案
- 升级解决方案
- 删除解决方案

新建解决方案

在数据开发模块的开发页面，新建一个解决方案，设置解决方案名称并选择业务相关的作业。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-184 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”或“数据开发 > 作业开发”。
3. 在左侧目录上方，单击解决方案图标，显示解决方案目录。
4. 单击解决方案目录上方的，弹出“新建解决方案”页面，配置如表3-139所示的参数。

表 3-139 解决方案参数

参数	说明
名称	自定义解决方案的名称。
选择作业	选择解决方案包含的作业。

5. 单击“确定”，新建的解决方案将在左侧目录中显示。

编辑解决方案

在解决方案目录中，右键单击解决方案名称，选择“编辑”，修改名称和作业。

导出解决方案

在解决方案目录中，右键单击解决方案名称，选择“导出”，导出zip格式的解决方案文件至本地。

导入解决方案

导入解决方案功能依赖于OBS服务，如无OBS服务，可从本地导入。

在解决方案目录中，右键单击根目录“解决方案”，选择“导入解决方案”，导入已上传到OBS或者本地的解决方案文件。

说明

在硬锁策略下，如果锁在其他人手中，重名策略选择了覆盖，则会覆盖失败。软硬锁策略请参考**配置软硬锁策略**。

升级解决方案

在解决方案目录中，右键单击解决方案名称，选择“升级”，导入已上传到OBS中的解决方案文件。升级解决方案时，会停止其中正在运行的作业，系统将依据用户配置的升级重启策略，判断是否在升级完成后重新启动作业。

删除解决方案

在解决方案目录中，右键单击解决方案名称，选择“删除”，删除解决方案。删除的解决方案不可恢复，请谨慎操作。

3.4.6 运行历史

运行历史功能可支持查看脚本、作业和节点的一周（7天）内用户的运行记录。

前提条件

运行历史功能依赖于OBS桶，若要使用该功能，必须先配置OBS桶。请参考**配置OBS桶**进行配置。

脚本运行历史

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-185 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
3. 在左侧目录上方，单击运行历史图标，显示该登录用户历史7天的脚本、作业的运行记录。

4. 在过滤框中选择“脚本”，展示历史7天的脚本运行记录。
5. 单击某一条运行记录，可查看当时的脚本信息和运行结果。

作业运行历史

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-186 选择数据开发

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”。
3. 在左侧目录上方，单击运行历史图标，显示该登录用户历史7天的脚本、作业的运行记录。
4. 在过滤框中选择“作业”，展示历史7天的作业运行记录。
5. 单击某一条运行记录，可查看当时的作业信息和日志信息。

说明

如果该作业当时只有部分节点执行测试，则运行历史只展示参与测试运行的节点信息和日志信息。

3.4.7 运维调度

3.4.7.1 运维概览

在“运维调度 > 运维概览”页面，用户可以通过图表的形式查看作业实例的统计数据，目前支持查看以下四种统计数据。

- 今日作业实例调度情况概览
- 近七天作业实例调度情况概览
- 近30天作业实例执行时长排行TOP 10

  单击作业名称，跳转至“实例监控”页面，查看执行时间长的作业实例的详细运行记录。

- 近30天作业实例运行失败TOP 10

  单击“运行失败次数”列的统计次数，跳转至“实例监控”页面，查看运行异常的作业实例的详细运行记录。

3.4.7.2 作业监控

3.4.7.2.1 批作业监控

批作业监控提供了对批处理作业的状态进行监控的能力。

批处理作业支持作业级别的调度计划，可以定期处理批量数据，主要用于实时性要求低的场景。批作业是是由一个或多个节点组成的流水线，以流水线作为一个整体被调度。被调度触发后，任务执行一段时间必须结束，即任务不能无限时间持续运行。

您可以在“作业监控 > 批作业监控”页面查看批处理作业的调度状态、调度频率、调度开始时间等信息，以及进行如**表3-140**所示的操作。

图 3-187 批作业监控

表 3-140 批作业监控支持的操作项

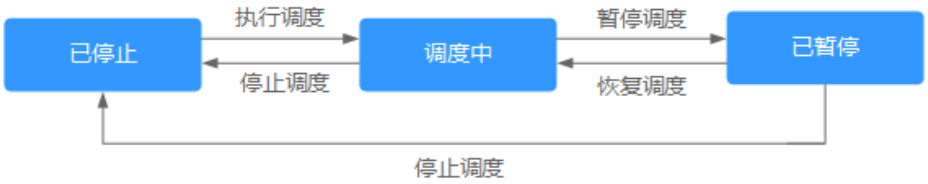
序号	支持的操作项	说明
1	根据“作业名”或“责任人名”搜索作业	-
2	根据“作业是否配置通知”、“调度状态”、“作业标签”或“下次计划时间”范围，筛选作业	-
3	批量配置作业	通过勾选作业名称前的复选框，支持批量执行操作。
4	查看作业实例状态	单击作业名称前方的，显示“最近的实例”页面，查看该作业最近的实例信息。
5	查看作业的节点信息	单击作业名称，在打开的页面中点击作业节点，查看该节点的相关关联作业/脚本与监控信息。
6	调度作业相关	在作业的“操作”列，支持执行调度、暂停调度、恢复调度、停止调度、调度配置等，详情请参见**批作业监控：调度作业**。
7	通知设置	在作业的“操作”列，选择“更多 > 通知设置”，弹出“新建通知”页面，参考**表3-150**配置通知参数。

序号	支持的操作项	说明
8	实例监控	在作业的“操作”列，选择“更多 > 实例监控”，跳转到实例监控页面，查看该作业所有实例的运行记录。
9	补数据	在作业的“操作”列，选择“更多 > 补数据”，弹出“补数据”对话框，详情请参见**批作业监控：补数据**。
10	添加作业标签	在作业的“操作”列，选择“更多 > 添加作业标签”，弹出“添加作业标签”对话框，详情请参见**批作业监控：添加作业标签**。

批作业监控：调度作业

作业开发完成后，用户可以在“作业监控”页面中管理作业的调度任务，例如：执行调度、暂停调度、恢复调度、停止调度。

图 3-188 调度作业



1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

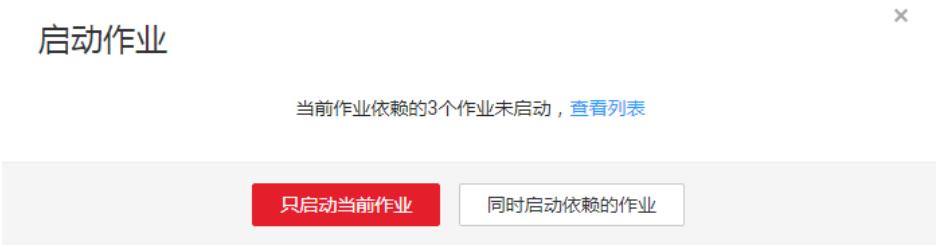
图 3-189 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
3. 单击“批作业监控”页签，进入批作业的监控页面。

4. 在作业的“操作”列，单击“执行调度”/“暂停调度”/“恢复调度”/“停止”。

如果该批处理作业设置有依赖的作业，执行调度该作业时可以设置为只启动当前作业或同时启动依赖的作业。如何配置依赖作业，请参见**配置作业调度任务（批处理作业）**。

图 3-190 启动作业



批作业监控：补数据

补数据是指作业执行一个调度任务，在过去某一段时间里生成一系列的实例。用户可以通过补数据，修正历史中出现数据错误的作业实例，或者构建更多的作业记录以便调试程序等。

只有配置了周期调度的作业，才支持使用该功能。如需查看补数据的执行情况，请参见**补数据监控**。

说明

当作业正在补数据时，请勿修改作业配置，否则会影响补数据过程中生成的作业实例。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-191 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。

3. 单击“批作业监控”页签，进入批作业的监控页面。
4. 在作业的“操作”列，选择“更多 > 补数据”。
5. 弹出“补数据”对话框，配置如表3-141所示的参数。

图 3-192 补数据参数

表 3-141 参数说明

参数	说明
补数据名称	系统自动生成一个补数据的任务名称，允许修改。
作业名称	显示需要补数据的作业名称。
业务日期	选择需要补数据的时间段。 **说明** 一个作业可进行多次补数据。但多次补数据的业务日期需要避免交叉重叠，否则可能导致数据重复或混乱，用户请谨慎操作。
并行周期数	设置同时执行的实例数量，最多可同时执行5个实例。 **说明** 请根据实际情况配置并行周期数，例如CDM作业实例，不可同时执行补数据操作，并行周期数只可设置为1。
需要补数据的下游作业	选择需要补数据的下游作业（指依赖于当前作业的作业），支持多选。

6. 单击“确定”，开始补数据，并进入“补数据监控”页面。

批作业监控：添加作业标签

支持给作业添加标签，便于作业实例的筛选分类。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-193 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
3. 单击“批作业监控”页签，进入批作业的监控页面。
4. 在作业的“操作”列，选择“更多 > 添加作业标签”。
5. 弹出“添加作业标签”对话框，填写需要配置的作业标签。

图 3-194 添加作业标签参数

6. 填写完标签后，单击“确认”，完成作业标签的添加。

3.4.7.2.2 实时作业监控

实时作业监控提供了对实时处理作业的状态进行监控的能力。

实时处理作业处理实时的连续数据，主要用于实时性要求高的场景。实时作业是由一个或多个节点组成的流水线，每个节点配置独立的、节点级别的调度策略，而且节点启动的任务可以永不下线。在实时作业里，带箭头的连线仅代表业务上的关系，而非任务执行流程，更不是数据流。

您可以在“作业监控 > 实时作业监控”页面查看实时处理作业的运行状态、开始执行时间、结束执行时间等信息，以及进行如**表3-142**所示的操作。

图 3-195 实时作业监控

表 3-142 实时作业监控支持的操作项

序号	支持的操作项	说明
1	根据“作业名”或“责任人名”搜索作业	-
2	根据“运行状态”或“作业标签”筛选作业	-
3	批量配置作业	通过勾选作业名称前的复选框，支持批量执行操作。
4	查看作业实例状态	单击作业名称前方的，显示“最近的实例”页面，查看该作业最近的实例信息。
5	作业状态相关	在作业的“操作”列，支持作业级别的启动、暂停、恢复、停止调度等。
6	添加作业标签	在作业的“操作”列，选择“更多 > 添加作业标签”，弹出“添加作业标签”对话框进行配置。
7	查看作业的节点信息	单击作业名称，进入“作业监控”详情页面后，单击某个节点，查看该节点的相关关联作业/脚本与监控信息。 **说明** 当作业中某个节点配置有事件驱动调度时，在单击此节点时会弹出子作业监控页面。
8	“禁用”和“恢复”节点	单击作业名称，进入“作业监控”详情页面后，右键单击某个节点选择“禁用”，禁用后可以再选择“恢复”，恢复运行时可以重新选择运行位置。详情请参见**实时作业监控：禁用节点后恢复**。
9	查看启动日志	单击作业名称，进入“作业监控”详情页面后，右键单击某个节点选择“查看启动日志”，您可以查看该节点的日志信息。
10	调度配置	单击作业名称，进入“作业监控”详情页面后，在“作业监控”详情页面中右键单击配置有事件驱动调度的节点，选择“调度配置”，您可以查看查看和修改节点的调度信息。详情请参见**实时作业监控：事件驱动调度节点调度配置**。
11	子作业监控	单击作业名称，进入“作业监控”详情页面后，单击配置有事件驱动调度的节点，查看子作业监控页面。详情请参见**实时作业监控：子作业监控**。
12	清除通道消息	单击作业名称，进入“作业监控”详情页面后，右键单击配置有事件驱动调度的节点，选择“清除通道消息”，您可以清除通道消息。

实时作业监控：禁用节点后恢复

您可以对实时作业中某个节点配置“禁用”后恢复运行，恢复运行时可以重新选择运行位置。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-196 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
3. 选择“实时作业监控”页签，单击作业名称。
4. 进入“作业监控”详情页面后，右键单击节点，选择“禁用”。
5. 设置禁用后，再右键单击选择“恢复”。弹出“恢复”对话框，配置如表3-143所示的参数。

表 3-143 恢复参数说明

参数	说明
上次暂停时间	节点暂停运行的起始时间。
未运行任务数	节点暂停期间没有运行的任务数量。
运行位置	“运行暂停期间任务”的参数。 表示选择节点暂停运行后，恢复运行时的启动位置。 • 从暂停节点开始运行 • 从子作业第一个节点开始运行
处理并发数	“运行暂停期间任务”的参数。 表示选择任务处理的数量。
任务名称	“运行暂停期间任务”的参数。 表示恢复的任务名称。

实时作业监控：事件驱动调度节点调度配置

当您配置的实时作业中某个节点配置有事件驱动调度时，在“作业监控”详情页面中右键单击配置有事件驱动调度的节点，选择“调度配置”，可以查看和修改节点的调度信息。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-197 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
3. 选择“实时作业监控”页签，单击作业名称。
4. 进入“作业监控”详情页面后，右键单击配置有事件驱动调度的节点，选择“调度配置”，配置如表3-144所示的参数。

图 3-198 调度配置

表 3-144 调度配策略参数说明

参数	说明
事件处理并发数	选择作业并行处理的数量，最大并发数为10。

参数	说明
事件检测间隔	配置事件检测时间间隔。时间间隔单位可以配置为秒或分钟。
失败策略	选择调度失败后的策略： • 结束调度 • 忽略失败，继续调度

实时作业监控：子作业监控

当用户配置的作业中某个节点配置有事件调度时，单击此节点可以查询子作业监控。在“子作业监控”页面可以对子作业设置停止、重跑、继续执行、强制成功、查看事件内容等操作。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-199 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
3. 选择“实时作业监控”页签，单击作业名称。
4. 进入“作业监控”详情页面后，单击配置有事件调度的节点。
   在“子作业监控”页面的“操作”列，提供如表3-145所示的操作。

表 3-145 子作业监控操作

操作项	说明
停止	停止运行状态为“运行中”的子作业实例。
重跑	重新运行状态为“成功”或“失败”的子作业实例。

操作项	说明
继续执行	子作业实例的状态为“运行异常”时，支持继续运行子作业实例中的后续节点。 **说明** 只有节点的“节点属性 > 高级 > 失败策略”设置为“挂起当前作业执行计划”时，才可以执行该操作。
强制成功	强制将状态为“失败”的子作业实例变更为“运行成功”状态。
事件内容	查看子作业的事件内容。

5. 单击“子作业监控”页面“状态”列下方的，显示该子作业节点的运行记录。在节点的“操作”列，提供如表3-146所示的操作。

表 3-146 操作（节点）

操作项	说明
查看日志	查看节点的日志信息。
更多 > 手工重试	节点的状态为“失败”时，支持重新运行节点。 **说明** 只有节点的“节点属性 > 高级 > 失败策略”设置为“挂起当前作业执行计划”时，才可以执行该操作。
更多 > 强制成功	节点的状态为“失败”时，支持将该节点强制变更为“成功”状态，且实例监控中作业实例的状态显示为“强制成功”。 **说明** 只有节点的“节点属性 > 高级 > 失败策略”设置为“挂起当前作业执行计划”时，才可以执行该操作。
更多 > 跳过	节点的状态为“待运行”或“已暂停节点”时，支持跳过该节点。
更多 > 暂停	节点的状态为“待运行”时，支持暂停运行该节点，该暂停节点的后续节点将会被阻塞。
更多 > 恢复	节点的状态为“已暂停”时，支持恢复运行该节点。

3.4.7.3 实例监控

作业每次运行，都会对应产生一次作业实例记录。在数据开发模块控制台的左侧导航栏，选择 “运维调度”，进入实例监控列表页面，用户可以在该页面中查看作业的实例信息，并根据需要对实例进行更多操作。

实例监控支持从“作业名称”、“创建人”、“CDM作业”和“节点类型”等维度搜索实例。其中按照“CDM作业”搜索，是从节点的维度搜索，搜索包含该节点的作业实例列表。

作业实例操作

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-200 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 实例监控”。
3. 当前支持批量停止、重跑、继续执行、强制成功多个实例，使用说明参见**表3-147**。

   其中，批量重跑多个实例时，重跑的顺序如下：

   - 如果作业不依赖上一调度周期，多个实例并行重跑。
   - 如果作业自依赖，多个实例串行重跑，以上一调度周期中实例执行完成的先后顺序为准，先执行完成的先重跑。
4. 在实例列表中，提供如**表3-147**所示的操作。

表 3-147 实例监控操作

操作项	说明
根据“作业名称”或“创建人”搜索作业	如果勾选了“作业名称”前的“精确搜索”，可支持作业名称的精确匹配搜索。 如果未勾选“作业名称”前的“精确搜索”，可支持作业名称的模糊匹配搜索。
根据“CDM作业”或“节点类型”筛选作业	-
停止	停止运行状态为“待运行”、“运行中”或“运行异常”的实例。
重跑	重新运行状态为“成功”或“取消”的实例。 详细操作请参见**重跑作业实例**。

操作项	说明
查看等待作业实例	实例的状态为“等待运行”时，支持查看等待的作业实例。
更多 > 继续执行	实例的状态为“运行异常”时，支持继续运行实例中的后续节点。 **说明** 只有节点的“节点属性 > 高级 > 失败策略”设置为“挂起当前作业执行计划”时，才可以执行该操作。
更多 > 强制成功	强制将状态为“运行异常”、“取消”、“失败”的实例变更为“成功”状态，当前实例状态显示为“强制成功”。
更多 > 查看	跳转至作业开发页面，查看作业信息。

5. 单击实例前方的，显示该实例所有节点的运行记录。
6. 在节点的“操作”列，提供如表3-148所示的操作。

表 3-148 操作（节点）

操作项	说明
查看日志	查看节点的日志信息。
更多 > 手工重试	节点的状态为“失败”时，支持重新运行节点。 **说明** 只有节点的“节点属性 > 高级 > 失败策略”设置为“挂起当前作业执行计划”时，才可以执行该操作。
更多 > 强制成功	节点的状态为“失败”时，支持将该节点强制变更为“成功”状态，且实例监控中作业实例的状态显示为“强制成功”。 **说明** 只有节点的“节点属性 > 高级 > 失败策略”设置为“挂起当前作业执行计划”时，才可以执行该操作。
更多 > 跳过	节点的状态为“待运行”或“已暂停节点”时，支持跳过该节点。
更多 > 暂停	节点的状态为“待运行”时，支持暂停运行该节点，该暂停节点的后续节点将会被阻塞。
更多 > 恢复	节点的状态为“已暂停”时，支持恢复运行该节点。

重跑作业实例

您可以对运行成功或失败的作业实例设置重跑，配置重跑开始位置。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-201 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 实例监控”。
3. 选择作业名称，在作业的“操作”列，单击“重跑”设置重跑作业实例；或单击作业名称左边的复选框，再选择“重跑”按钮设置作业实例重跑。

图 3-202 设置作业重跑

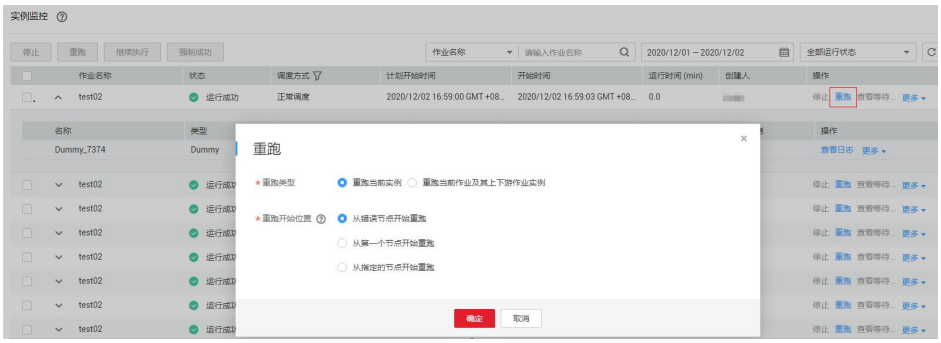


表 3-149 参数说明

参数	说明
重跑类型	选择需要重跑的实例。 ● 重跑当前实例 ● 重跑当前作业以及上下游作业实例：
开始时间	重跑用户设置的时间段内的实例。
重跑作业实例列表	选择需要重跑的上下游作业，支持多选。

参数	说明
重跑开始位置	选择作业实例重跑的开始位置： • 从错误节点开始重跑：作业实例执行失败时，从实例执行失败的错误节点开始重跑。 • 从第一个节点开始重跑：从作业实例的第一个节点开始重跑。 • 从指定的节点开始重跑：从作业实例中指定的节点开始重跑。仅当“重跑类型”为“重跑当前实例”时有此选项。 **说明** 以下两种情况，系统运行会从第一个节点开始重跑。 • 如果作业中节点个数或者名称发生变化，从第一个节点开始重跑。 • 如果重跑成功状态的作业实例，从第一个节点开始重跑。
处理并发数	选择作业实例并行处理的数量。

3.4.7.4 补数据监控

在数据开发模块控制台的左侧导航栏，选择“运维调度 > 补数据监控”，进入补数据的任务监控页面。

用户可以在补数据监控主页，查看补数据的任务状态、业务日期、并行周期数、补数据作业名称，以及停止运行中的任务。

在补数据监控主页，单击补数据名称，进入补数据监控详情页面。在此页面，用户可以查看补数据的任务执行情况，以及手动干预实例和节点的执行（如需了解更多，请参见**批作业监控：补数据**）。

说明

- 支持计划时间，开始时间，结束时间的排序，注意三者之间，同一时间只有其中一个当前排序有效。
- 排序按钮点击顺序为：点击1下为升序，点击2下为降序，点击3下取消排序。

3.4.7.5 通知管理

DataArts Studio使用消息通知服务（Simple Message Notification，简称SMN）依据用户的订阅需求主动推送通知消息，用户在作业运行异常或成功时能立即接收到通知。

3.4.7.5.1 管理通知

用户可以通过通知管理功能配置作业通知任务，当作业运行异常或成功时向相关人员发送通知。

配置通知

为作业配置通知前：

- 已开通消息通知服务并配置主题。
- 作业已提交，且不是“未启动”状态。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-203 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
3. 在页面右侧的“通知管理”页签，单击“通知配置”，弹出“通知配置”页面，配置如表3-150所示的参数。

表 3-150 通知参数

参数	是否必选	说明
通知范围	是	选择通知的范围： ● 单个作业：对单个作业发送通知。 ● 所有作业：对所有作业发送通知。
作业名称	是	选择作业。

参数	是否必选	说明
通知类型	是	选择通知类型： ● 单个作业： – 运行异常/失败：作业的状态为“运行异常”或“失败”时，发送通知。 – 运行成功：作业的状态为“成功”时，发送通知。 – 未完成：该功能仅支持按天调度的作业配置。如果作业执行时间超过设置的未完成时间，则发送通知。 – 资源繁忙：如果执行作业时，资源繁忙，则发送通知。 ● 所有作业： – 运行异常/失败：作业的状态为“运行异常”或“失败”时，发送通知。 – 资源繁忙：如果执行作业时，资源繁忙，则发送通知。 **说明** 实时作业只支持状态为**运行异常/失败**时发送通知，批处理作业在状态为运行成功和运行异常/失败时都能发送通知。
选择主题	是	选择通知的消息主题。 **说明** 当前仅支持“短信”、“邮件”、“HTTP”这三种协议的订阅终端订阅主题。
开关	是	是否开启通知，默认开启。

4. 单击“确定”，为作业配置通知。

编辑通知

通知新建完成后，用户可以根据需求修改通知的参数。

1. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
2. 在页面右侧选择“通知管理”页签。
3. 在通知的“操作”列，单击“编辑”，弹出“编辑通知”页面，参考**表3-150**修改通知的参数。
4. 单击“确定”，保存修改。

关闭通知

用户可以在“编辑”中关闭通知任务，也可以在通知列表中关闭通知任务。

1. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
2. 在页面右侧选择“通知管理”页签。

3. 在通知的“开关”列，单击  ，切换成  时，通知为关闭状态。

查看通知记录

用户可以在通知记录中查看所有的通知信息。

1. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
2. 在页面右侧选择“通知记录”页签，进入通知记录页面。

删除通知

当用户不需要使用某个通知时，可以参考如下操作删除该通知。

1. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
2. 在页面右侧选择“通知管理”页签。
3. 支持如下两种方式删除通知：
   - 在通知的“操作”列，单击“删除”，弹出“删除通知”页面。
   - 勾选待删除的通知，单击通知列表上方的“批量删除”，弹出“删除通知”页面。
4. 单击“确认”，删除通知。

3.4.7.5.2 通知周期概览

操作场景

用户可以按照天/周/月为调度周期配置通知任务，向相关人员发送通知。让相关人员可以定期跟踪作业的调度情况（作业调度成功数量，作业调度失败异常数量以及作业失败详情）。

约束限制

该功能依赖于OBS服务。

前提条件

- 已开通消息通知服务并配置主题，为主题添加订阅。
- 已提交作业，且作业不是“未启动”状态。
- 已开通对象存储服务，并在OBS中创建文件夹。

配置通知

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-204 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
3. 在页面右侧的“周期概览”页签，单击“通知配置”，弹出“通知配置”页面，配置如表3-151所示的参数。

表 3-151 通知参数

参数	是否必选	说明
通知名称	是	设置发送的通知名称。
调度周期	是	选择通知发送的调度周期，可以设置为按“天”、“周”或“月”发送。 **说明** 按天发送，通知记录为以发送时间往前推24小时时间段的数据；按周发送，通知记录为往前推七天时间段的数据；按月发送，通知记录为往前推30天时间段的数据
选择时间	是	设置通知发送的具体日期。 • 当调度周期为周时，可设置为一周中星期一至星期日的某一天或某几天。 • 当调度周期为月时，可设置为一月中每月1号至每月31号的某一天或某几天。
具体时间	是	设置通知发送的具体时间点，可以精确设置到小时和分钟。
选择概览通知的主题	是	单击下拉选项，设置通知发送的主题。
选择OBS桶	是	单击“OBS”设置通知记录数据存储的位置。
开关	是	是否开启通知，默认开启。

4. 单击“确定”。
5. 通知配置完成后，您可以在通知的“操作”列进行如下操作。
   - 单击“编辑”，打开“通知配置”页面，可以重新编辑通知。编辑完成后选择“确定”，保存修改。
   - 单击“记录”，打开“查看记录”页面，可以查看作业的调度情况。
   - 单击“删除”，打开“删除通知”页面，选择“确定”，删除通知。

3.4.7.6 备份管理

通过备份功能，您可每日定时备份昨日系统中的所有作业、脚本、资源和环境变量。

通过还原功能，您可还原已备份的资产，包含作业、脚本、资源和环境变量。

约束限制

该功能依赖于OBS服务。

前提条件

已开通对象存储服务，并在OBS中创建文件夹。

备份资产

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-205 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“备份管理”。
3. 单击“启动每日备份”，打开“OBS文件浏览”页面，选择OBS文件夹，设置备份数据的存储位置。

说明

- 每日备份在每日0点开始备份昨日的所有作业、脚本、资源和环境变量，启动当日不会备份昨日的作业、脚本、资源和环境变量。
- 选择OBS存储路径时，若仅选择至桶名层级，则备份对象自动存储在以“备份日期”命名的文件夹内。环境变量，资源，脚本和作业分别存储在1_env,2_resources,3_scripts和4_jobs文件夹内。
- 备份成功后，在以“备份日期”命名的文件夹内，自动生成backup.json文件，该文件按照节点类型存储了作业信息，支持恢复作业前进行修改。
- 启动每日备份后，若想结束备份任务，您可以单击右边的“停止每日备份”。

还原资产

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-206 选择数据开发

步骤2 在数据开发模块控制台的左侧导航栏，选择“备份管理”。

步骤3 选择“还原管理”页签，单击“还原备份”。

在还原备份对话框中，从OBS桶中选择待还原的资产存储路径，设置重名处理策略。

说明

- 待还原的资产存储路径为**备份资产**中生成的文件路径。
- 您可在还原资产前修改备份路径下的backup.json文件，支持修改连接名（connectionName）、数据库名（database）和集群名（clusterName）。

图 3-207 还原资产

步骤4 单击“确定”。

----结束

3.4.8 配置管理

3.4.8.1 配置

3.4.8.1.1 配置环境变量

本章节主要介绍环境变量的配置和使用。

使用场景

配置作业参数，当某参数隶属于多个作业，可将此参数提取出来作为环境变量，环境变量支持导入和导出。

导入环境变量

导入环境变量功能依赖于OBS服务，如无OBS服务，可从本地导入。

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-208 选择数据开发

步骤2 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤3 单击“环境变量”，在“环境变量配置”页面，选择“导入”。

步骤4 在导入环境变量对话框中，选择已上传至OBS或者本地的环境变量文件，以及重命名策略。

图 3-209 导入环境变量

----结束

配置方法

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-210 选择数据开发

步骤2 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤3 单击“环境变量”，在“环境变量配置”页面，配置如表3-152所示的变量或常量，单击“保存”。

说明

变量和常量的区别是其他工作空间或者项目导入的时候，是否需要重新配置值。

- 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。

表 3-152 环境变量参数配置

参数	是否必选	说明
参数名称	是	只支持英文字母、数字、“-”、“_”，最大长度为64字符，且参数名称不允许重名。
参数值	是	参数值当前支持常量和EL表达式，不支持系统函数。例如支持123，abc。 关于EL表达式的使用，请参见表达式概述。

配置完一个环境变量后，您还可以进行新增、修改或删除等操作。

- 新增：单击“新增”配置新的环境变量。
- 修改：参数值为常量时，直接在文本框中修改参数值；参数值为EL表达式时，可以单击文本框后方的编辑EL表达式，修改参数值。修改完成后，请“保存”。
- 删除：在参数值文本框后方，单击删除环境变量。

----结束

使用方法

当前配置好的环境变量支持如下两种使用方法：

1. ${环境变量名}
2. #{Evn.get(“环境变量名”)}

操作示例

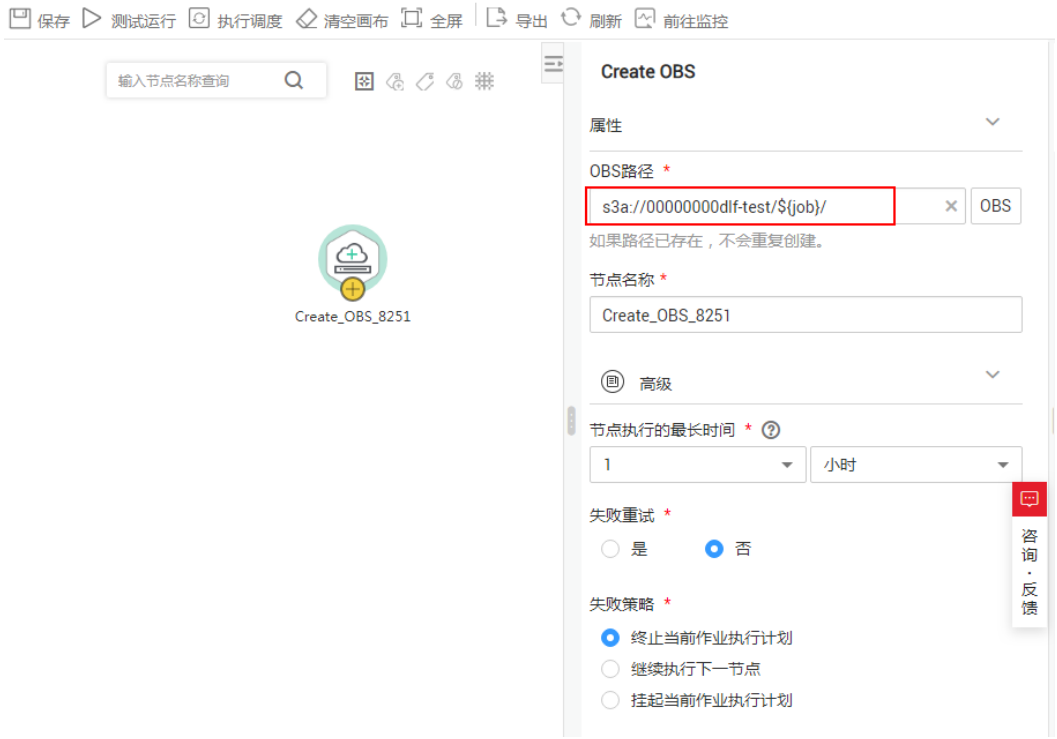
背景信息：

- 在数据开发模块系统中已创建一个作业“test”。
- 在环境变量中已新增一个变量，“参数名”为“job”，“参数值”为“123”。

步骤1 打开作业“test”，从左侧节点库中拖拽一个“Create OBS”节点。

步骤2 在节点属性页签中配置属性。

图 3-211 Create OBS



步骤3 单击“保存”后，选择“前往监控”页面监控作业的运行情况。

----结束

3.4.8.1.2 配置 OBS 桶

脚本、作业或节点的历史运行记录依赖于OBS桶，如果未配置测试运行历史OBS桶，则无法查看历史运行的详细信息。请参考本节操作配置OBS桶。

配置方法

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-212 选择数据开发

步骤2 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤3 选择“OBS桶”。

步骤4 配置OBS桶的信息。

图 3-213 配置 OBS 桶

步骤5 单击“保存”，完成配置。

----结束

3.4.8.1.3 管理作业标签

作业标签用于给相同或用途类似的作业打上标签，便于管理作业，并根据标签查询作业。参考本节操作，您可管理作业标签，执行新增、修改和查询操作。

配置方法

步骤1 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-214 选择数据开发

步骤2 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤3 选择“作业标签”，在“作业标签管理”页面，单击“新建”，配置作业名称，确认后完成新建。

说明

作业标签最多支持创建100个。

----结束

3.4.8.1.4 配置委托

数据开发模块的作业执行中会遇到如下问题：

- 数据开发模块的作业执行机制是以启动作业的用户身份执行该作业。对于按照周期调度方式执行的作业，当启动该作业的IAM帐号在调度周期内被删除后，系统无法获取用户身份认证信息，导致作业执行失败。
- 如果作业被低权限的用户启动，也会因为权限不足导致作业执行失败。

若需解决以上两个问题，则可配置委托。配置委托后，作业执行过程中，以委托的身份与其他服务交互，可以避免上述两种场景下作业执行失败。

委托的作用

由于云各服务之间存在业务交互关系，一些云服务需要与其他云服务协同工作，需要您创建云服务委托，将操作权限委托给这些服务，让这些服务以您的身份使用其他云服务，代替您进行一些资源运维工作。

委托的分类

委托分两类，工作空间委托和作业委托。

- 工作空间委托：工作空间级别的，全局委托。适用于该空间内的所有作业。
- 作业委托：适用于单个作业级别。

作业委托优先级高于工作空间委托，如果工作空间与作业级别的委托都没有配置，作业会以启动者的身份去执行。

约束限制

- 创建或修改委托需要用户具有Security Administrator权限。
- 配置工作空间级委托，需要用户具有DAYU Administrator或者Tenant Administrator权限。
- 配置作业级委托，需要用户具有查看列表委托的权限。

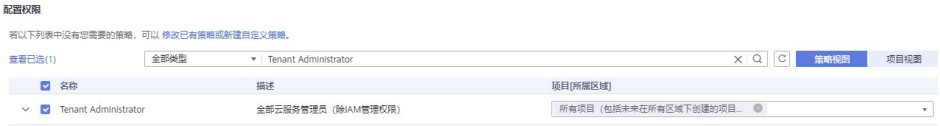
创建委托

1. 登录IAM服务控制台。
2. 选择“委托 > 创建委托”。
3. 设置“委托名称”。例如：DataArts Studio_agency。
4. “委托类型”选择“云服务”，在“云服务”中选择数据治理中心DataArts Studio，将操作权限委托给DataArts Studio，让DataArts Studio以您的身份使用其他云服务，代替您进行一些资源运维工作。
5. “持续时间”选择“永久”。

图 3-215 创建委托

6. 在“权限选择”区域中，单击“配置权限”。
7. 在弹出页面中搜索“Tenant Administrator”策略，勾选“Tenant Administrator”策略并单击“确定”，如图3-216所示。
   - 因Tenant Administrator策略具有除统一身份认证服务IAM外，其他所有服务的所有执行权限。所以给委托服务DataArts Studio配置Tenant Administrator，可访问周边所有服务。
   - 若您想达到对权限较小化的安全管控要求，Tenant Administrator可不配置，仅配置OBS OperateAccess权限（因作业执行过程中，需要往obs写执行日志信息，因此需要添加 OBS OperateAccess权限。）。然后再根据作业中的节点类型，配置不同的委托权限。例如某作业仅包含Import GES节点，可配置GES Administrator权限和OBS OperateAccess权限即可。详细方案请参考配置权限。

   图 3-216 配置权限

8. 单击“确定”完成委托创建。

配置权限

将帐号的操作权限委托给DataArts Studio服务后，需要配置委托身份的权限，才可与其他服务进行交互。

为实现对权限较小化的安全管控要求，可根据作业中的节点类型，以服务为粒度，参见表3-153配置相应的服务Admin权限。

也可精确到具体服务的操作、资源以及请求条件等。根据作业中的节点类型，以对应服务API接口为粒度进行权限拆分，满足企业对权限最小化的安全管控要求。参见表3-154进行配置。例如包含Import GES节点的作业，您只需要创建自定义策略，并勾选ges:graph:getDetail（查看图详情），ges:jobs:getDetail（查询任务状态），ges:graph:access（使用图）这三个授权项即可。

须知

- MRS相关的节点（MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、 MRS MapReduce），以及通过直连方式的（MRS Spark SQL、MRS Hive SQL）节点，由于部分MRS集群不支持委托方式提交作业，所以这类作业不能配置委托。
- 支持委托方式提交作业的MRS集群如下：
  - 非安全集群
  - 安全集群，集群版本大于 2.1.0，并且安装了MRS 2.1.0.1及以上版本的补丁。

- 配置服务级Admin权限

  因作业执行过程中，需要往obs写执行日志信息，因此粗粒度授权时，所有作业都需要添加 OBS OperateAccess权限。

表 3-153 配置相关节点的 admin 权限

节点名称	系统权限	权限描述
CDM Job	DAYU Administrator	数据治理中心服务的所有执行权限。
Import GES	GES Administrator	图引擎服务的所有执行权限。该角色有依赖，需要在同项目中勾选依赖的角色：Tenant Guest、Server Administrator。
• MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、MRS MapReduce • MRS Spark SQL、MRS Hive SQL（通过MRS API方式连接MRS集群的）	MRS Administrator KMS Administrator	MRS Administrator：MapReduce服务的所有执行权限。该角色有依赖，需要在同项目中勾选依赖的角色：Tenant Guest、Server Administrator。 KMS Administrator：数据加密服务加密密钥的管理员权限。
MRS Spark SQL、MRS Hive SQL、MRS Kafka、Kafka Client（通过代理方式连接集群）	DAYU Administrator KMS Administrator	DAYU Administrator：数据治理中心服务的所有执行权限。 KMS Administrator：数据加密服务加密密钥的管理员权限。
DLI Flink Job、DLI SQL、DLI Spark	DLI Service Admin	数据湖探索的所有执行权限。
DWS SQL、Shell、RDS SQL（通过代理方式连接数据源）	DAYU Administrator KMS Administrator	DAYU Administrator：数据治理中心服务的所有执行权限。 KMS Administrator：数据加密服务加密密钥的管理员权限。
CSS	DAYU Administrator Elasticsearch Administrator	DAYU Administrator：数据治理中心服务的所有执行权限。 Elasticsearch Administrator：云搜索服务的所有执行权限。该角色有依赖，需要在同项目中勾选依赖的角色：Tenant Guest、Server Administrator。
Create OBS、Delete OBS、OBS Manager	OBS OperateAccess	查看桶、上传对象、获取对象、删除对象、获取对象ACL等对象基本操作权限
SMN	SMN Administrator	消息通知服务的所有执行权限。

- 配置细粒度权限（根据各服务支持的授权项，创建自定义策略。）

创建自定义策略的详细操作请参见《统一身份认证IAM用户指南》中的“创建自定义策略”。

说明

- 作业执行过程中，需要向OBS中写入执行日志。当采取精细化授权方式时，任何类型的作业均需要添加OBS的如下授权项：
  - obs:bucket:GetBucketLocation
  - obs:object:GetObject
  - obs:bucket:CreateBucket
  - obs:object:PutObject
  - obs:bucket:ListAllMyBuckets
  - obs:bucket:ListBucket
- CDM Job节点隶属于DataArts Studio模块，DataArts Studio不支持细粒度授权。因此包含这几类节点的作业，给服务配置权限仅支持DataArts Studio Administarator。
- CSS不支持细粒度授权，且需要通过代理执行。因此包含这类节点的作业，需要配置DataArts Studio Administarator和Elasticsearch Administrator权限。
- SMN不支持细粒度授权，因此包含这类节点的作业，需要配置SMN Administarator权限。

表 3-154 自定义策略

节点名称	授权项
Import GES	• ges:graph:access • ges:graph:getDetail • ges:jobs:getDetail
• MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、MRS MapReduce • MRS Spark SQL、MRS Hive SQL（通过MRS API方式连接MRS集群的）	• mrs:job:delete • mrs:job:stop • mrs:job:submit • mrs:cluster:get • mrs:cluster:list • mrs:job:get • mrs:job:list • kms:dek:crypto • kms:cmk:get
MRS Spark SQL、MRS Hive SQL、MRS Kafka、Kafka Client（通过代理方式连接集群）	• kms:dek:crypto • kms:cmk:get • DataArts Studio Administarator(角色)

节点名称	授权项
DLI Flink Job、DLI SQL、DLI Spark	• dli:jobs:get • dli:jobs:update • dli:jobs:create • dli:queue:submit_job • dli:jobs:list • dli:jobs:list_all
DWS SQL、Shell、RDS SQL（通过代理方式连接数据源）	• kms:dek:crypto • kms:cmk:get • DataArts Studio Administarator(角色)
Create OBS、Delete OBS、OBS Manager	• obs:bucket:GetBucketLocation • obs:bucket:ListBucketVersions • obs:object:GetObject • obs:bucket:CreateBucket • obs:bucket:DeleteBucket • obs:object:DeleteObject • obs:object:PutObject • obs:bucket:ListAllMyBuckets • obs:bucket:ListBucket

配置工作空间级委托

⚠ 注意

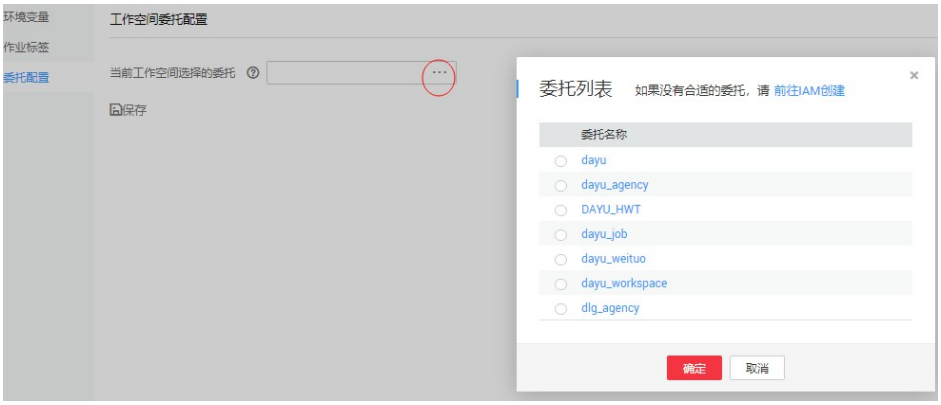
工作空间级别的委托影响所有的作业，请慎重配置。 特别是部分作业中包含 MRS相关的节点。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-217 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。
3. 单击“委托配置”，在工作空间委托配置页面配置委托。
4. 在委托列表中选择合适的委托，也可重新创建委托。创建委托和配置权限，请参见**创建委托**。

图 3-218 配置工作空间级委托



5. 单击“确定”，回到工作空间委托配置页面，再单击，创建工作空间级委托成功。

配置作业级委托

说明

支持新建作业时，配置作业级委托。也支持修改已有作业的委托。

新建作业时配置委托

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-219 选择数据开发

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”。
3. 在作业目录处，单击右键，选择“新建作业”。系统弹出新建作业对话框，若已配置过工作空间级委托，则该作业默认使用工作空间级委托。您也可从委托列表中，选择其他已创建的委托。

图 3-220 配置作业委托

修改已有作业的委托

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
2. 在作业目录处，双击选中已有作业。在节点编排页面右侧，选择“作业基本信息”。系统弹出作业信息基本配置对话框，若已配置过工作空间级委托，则该作业默认使用工作空间级委托。您也可从委托列表中，选择其他已创建的委托。

3.4.8.1.5 配置默认项

本章节主要介绍默认项的配置。

使用场景

当某参数被多个作业调用时，可将此参数提取出来作为默认配置项，无需每个作业都配置该参数。

配置周期调度

依赖的作业失败后，当前作业处理策略是根据配置的默认策略来执行，配置默认策略操作如下。

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 单击“默认项设置”，可设置周期调度配置项。

说明

策略支持如下三种，系统默认配置为“终止执行”。

- 挂起：当被依赖的作业执行失败后，当前作业会挂起。
- 继续执行：当被依赖的作业执行失败后，当前作业会继续执行。
- 终止执行：当被依赖的作业执行失败后，当前作业会终止执行。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

配置多 IF 策略

节点执行依赖多个IF条件的处理策略，配置默认策略操作如下。

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 单击“默认项设置”，可设置多IF策略配置项。

说明

策略支持如下两种，系统默认策略为“逻辑或”。

- 逻辑或：表示多个IF判断条件只要任意一个满足条件则执行。
- 逻辑与：表示多个IF判断条件需要所有条件满足时才执行。

具体使用方法请参见**多IF条件下当前节点的执行策略**。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

配置软硬锁策略

作业或脚本的抢锁操作依赖于软硬锁处理策略。软硬锁的最大的区别在于普通用户抢锁时，软锁可以任意抢锁（无论锁是否在自己手上），硬锁只能对自己持有锁的文件进行操作（包括抢锁、解锁操作）。发布、运行、调度等操作不受锁的影响，无锁也可操作。

用户可根据实际场景，配置相应的软硬锁策略。

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 单击“默认项设置”，可设置软硬锁策略配置项。

说明

系统默认策略为“软锁”。

- 软锁：忽略当前作业或脚本是否被他人锁定，可以进行抢锁或解锁。
- 硬锁：若作业或脚本被他人锁定，则需锁定的用户解锁之后，当前使用人方可抢锁，空间管理员或DAYU Administrator可以任意抢锁或解锁。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

3.4.8.2 管理资源

用户可以通过资源管理功能，上传自定义代码或文本文件作为资源，在节点运行时调用。可调用资源的节点包含DLI Spark、MRS Spark、MRS MapReduce和DLI Flink Job。

创建资源后，配置资源关联的文件。在作业中可以直接引用资源。当资源文件变更，只需要修改资源引用的位置即可，不需要修改作业配置。关于资源的使用样例请参见**开发一个DLI Spark作业**。

约束限制

该功能依赖于OBS服务或MRS HDFS服务。

新建目录（可选）

如果已存在可用的目录，可以不用新建目录。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-221 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 在资源目录中，单击，弹出“新建目录”页面，配置如**表3-155**所示的参数。

表 3-155 资源目录参数

参数	说明
目录名称	资源目录的名称，只能包含英文字母、数字、中文字符、“_”、“-”，且长度为1~32个字符。
选择目录	选择该资源目录的父级目录，父级目录默认为根目录。

4. 单击“确定”，新建目录。

新建资源

新建资源前，请确保您已开通OBS服务。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-222 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 单击“新建资源”，弹出“新建资源”页面，配置如表3-156所示的参数。单击“确定”，新建资源。

表 3-156 资源管理参数

参数	是否必选	说明
名称	是	资源的名称，只能包含英文字母、数字、中文字符、“_”、“-”，且长度为1~32个字符。
类型	是	选择资源的文件类型： • jar：用户jar文件。 • pyFile：用户Python文件。 • file：用户文件。 • archive：用户AI模型文件。
资源位置	是	选择资源所在的位置，当前支持OBS和HDFS两种资源存储位置。HDFS当前只支持MRS Spark、MRS Flink Job、MRS MapReduce节点。

参数	是否必选	说明
主Jar包	是	• “资源位置”为“OBS”时，选择已上传到OBS中的主Jar包。 • “资源位置”为“HDFS”时，请先选择MRS集群，然后再选择已经上传到HDFS中的主Jar包。
依赖Jar包	否	选择已上传到OBS中的依赖Jar包。“类型”为“jar”，且“资源位置”为“OBS”或者“HDFS”时，配置该参数。
选择资源	是	选择具体的资源文件。
存储路径	是	选择资源的目的存储路径。“资源位置”为“本地”时，配置该参数。
描述	否	资源的描述信息。
选择目录	是	选择资源所属的目录，默认为根目录。

编辑资源

资源新建完成后，用户可以根据需求修改资源的参数。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-223 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 在资源的“操作”列，单击“编辑”，弹出“编辑资源”页面，参考表3-156修改资源的参数。
4. 单击“确定”，保存修改。

删除资源

当用户不需要使用某个资源时，可以删除该资源。

删除资源前，请确保该资源未被作业使用。删除资源的时候，会检查资源被哪些作业引用，引用列表中“版本”一列，表示此资源被哪些作业版本引用。点击删除时，会删除对应的作业和这个作业的所有版本信息。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-224 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 在资源的“操作”列，单击“删除”，弹出“删除资源”页面。
4. 单击“确定”，删除资源。

导入资源

当用户想要导入某个资源时，可以参考如下操作导入该资源。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-225 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 在资源目录中，单击，选择“导入资源”，弹出“导入资源”页面。
4. 选择已上传至OBS中的资源文件，然后单击“下一步”，导入完成后，单击“关闭”完成资源的导入。

导出资源

当用户想要导出某个资源到本地时，可以参考如下操作导出该资源。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-226 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 在资源目录中，单击，选择“导出资源”，系统开始下载资源到本地。

查看资源引用

当用户想要查看某个资源被引用的情况时，可以参考如下操作查看引用。

1. 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

   图 3-227 选择数据开发

2. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
3. 在资源目录中，右键单击对应的资源名，选择“查看引用”，弹出“引用列表”窗口。
4. 在引用列表窗口，可以查看该资源被引用的情况。

3.4.9 节点参考

3.4.9.1 节点概述

节点定义对数据执行的操作。数据开发模块提供数据集成、计算&分析、数据库操作、资源管理等类型的节点，您可以根据业务模型选择所需的节点。

- 节点的参数支持使用EL表达式，EL表达式的使用方法详见**表达式概述**。
- 节点间的连接方式支持串行和并行。

  串行连接：按顺序逐个执行节点，当A节点执行完成后，再执行B节点。

  并行连接：A节点和B节点同时执行。

图 3-228 连接示意图

3.4.9.2 CDM Job

功能

通过CDM Job节点执行一个预先定义的CDM作业，实现数据迁移功能。

参数

用户可参考表3-157，表3-158和表3-159配置CDM Job节点的参数。配置血缘关系用以标识数据流向，在数据目录模块中可以查看。

表 3-157 属性参数

参数	是否必选	说明
CDM集群名称	是	选择待执行的CDM作业所属的CDM集群。 此处支持勾选两个CDM集群，用于提升作业可靠性。 ● 勾选两个集群后，第一个勾选的集群为主集群，第二个勾选的集群为备集群。作业会默认运行在主集群上，当主集群状态异常后，会触发切换到备集群运行作业。 ● 勾选两个集群的场景下，“作业类型”不推荐选择“创建新作业”，应设置为“选择已存在的作业”，且确保主备集群下分别存在该作业。您可以在主集群新建CDM作业并导出，然后再导入作业到备集群，实现作业同步，具体操作方法请参见导出导入CDM作业。

参数	是否必选	说明

CDM作业类型	是	• 选择已存在的作业。
		• 创建新作业。
		说明
		• 如果作业类型为“选择已存在的作业”，当CDM作业有修改时，此处作业节点不会同步更新。如需更新此作业节点，需要重新保存该节点所在的作业，用于触发CDM作业更新。
		• 如果作业类型为“创建新作业”，节点运行时会检测是否有同名CDM作业。
		  • 如果CDM作业未运行，则按照请求体内容更新同名作业。
		  • 如果同名CDM作业正在运行中，则等待作业运行完成后更新该作业。在此期间该作业可能被其他任务启动，可能会导致数据抽取不符合预期（如作业配置未更新、运行时间宏未替换正确等），因此请注意不要创建多个同名作业。

CDM作业名称	否	仅当“作业类型”为“选择已存在的作业”时需要配置该参数。选择待执行的CDM作业。 如果此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。
CDM作业消息体	否	仅当“作业类型”为“创建新作业”时需要配置该参数。此处需要填写CDM作业JSON。方便起见可以在CDM已有作业处选择操作“更多 > 查看作业JSON”，复制其中的JSON内容，在此处修改适配。 如果此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。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-158 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** ● 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。 ● 如果调度CDM迁移作业时使用了参数传递，不能在CDM迁移作业中配置“作业失败重试”参数，推荐在此处配置即可。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-159 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击 ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.3 Rest Client

功能

通过Rest Client节点执行一个内的RESTful请求，目前只支持IAM Token认证鉴权方式的RESTful请求。

说明

当由于网络限制，Rest Client某些API无法调通时，可以尝试使用Shell脚本进行API调用。您需要拥有ECS弹性云服务器，并确保ECS主机和待调用的API之间网络可通，然后在DataArts Studio创建主机连接，通过Shell脚本使用CURL命令进行API调用。

参数

用户可参考表3-160，表3-161和表3-162配置Rest Client节点的参数。

表 3-160 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
代理集群名称	是	选择CDM集群名称，CDM集群提供代理连接的功能。 如果选择选择的CDM集群与第三方服务处于同一个VPC下，那么Rest Client可以调用租户面的API。
URL地址	是	填写请求主机的IP或域名地址，以及端口号。例如：https://192.160.10.10:8080

参数	是否必选	说明
HTTP方法	是	选择请求的类型： ● GET ● POST ● PUT ● DELETE
请求头	否	单击 ，添加请求消息头，参数说明如下： ● 参数名称 选择参数的名称，选项为“Content-Type”、“Accept-Language”。 ● 参数值 填写参数的值。
URL参数	否	填写URL参数，格式为“参数=值”形式的字符串，字符串间以换行符分隔。当“HTTP方法”为“GET”时，显示该配置项。参数说明如下： ● 参数 只支持英文字母、数字、“-”、“_”，最大长度为32字符。 ● 值 只支持英文字母、数字、“-”、“_”、“$”、“{”和“}”，最大长度为64字符。
请求消息体	是	填写Json格式的请求消息体。当“HTTP方法”为“POST”、“PUT”时，显示该配置项。
是否需要判断返回值	否	设置是否判断返回消息的值和预期的一致。当“HTTP方法”为“GET”时，显示该配置项。 ● YES：检查返回消息中的值是否和预期的一致。 ● NO：不检查，请求返回200响应码（表示节点执行成功）。

参数	是否必选	说明

返回值字段路径	是	填写Json响应消息中某个属性的路径（下称：Json属性路径），每个Rest Client节点都只能配置一个属性的路径。当“是否需要判断返回值”为“YES”时，显示该配置项。
		例如，返回结果为：
		{
		"param1": "aaaa",
		"inner":
		{
		"inner":
		{
		"param4": 2014247437
		},
		"param3": "cccc"
		},
		"status": 200,
		"param2": "bbbb"
		}
		其中“param4”属性的路径为“inner.inner.param4”。

请求成功标志位	是	填写请求成功标志位，如果响应消息的返回值与请求成功标志位中的某一个匹配，表示节点执行成功。当“是否需要判断返回值”为“YES”时，显示该配置项。 请求成功标志位只支持英文字母、数字、“-”、“_”、“$”、“{”、“}”，多个值使用“;”分隔。
请求失败标志位	否	填写请求失败标志位，如果响应消息的返回值与请求失败标志位中的某一个匹配，表示节点执行失败。当“是否需要判断返回值”为“YES”时，显示该配置项。 请求失败标志位只支持英文字母、数字、“-”、“_”、“$”、“{”、“}”，多个值使用“;”分隔。
请求间隔时间（秒）	是	如果响应消息的返回值与请求成功标志位不匹配，将每隔一段时间查询一次，直到响应消息的返回值与请求成功标志位一致。节点执行的超时时间默认为1小时，如果1小时内查询的结果始终为不匹配，那么节点的状态将置为失败。当“是否需要判断返回值”为“YES”时，显示该配置项。
响应消息体解析为传递参数定义	否	设置作业变量与Json属性路径的对应关系，参数间以换行符分隔。 例如：var4=inner.inner.param4 其中，“var4”为作业变量，作业变量只支持英文字母、数字，最大长度为64字符；“inner.inner.param4”为Json属性路径。 仅该节点的后续节点引用该参数才会生效，引用该参数时，格式为：${var4}。

表 3-161 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-162 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.4 Import GES

功能

通过Import GES节点可以将OBS桶中的文件导入到GES的图中。

参数

用户可参考**表3-163**和**表3-164**配置Import GES节点的参数。

表 3-163 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
图名称	是	可以直接选择需要导入的图，也支持手动输入图名称。 如需新建GES图，请前往GES管理控制台进行新建。
元数据	是	可以直接选择对应的元数据，也支持手动输入元数据的OBS路径。
边数据集	是	可以直接选择对应的边数据集，也支持手动输入边数据集的OBS路径。
点数据集	否	可以直接选择对应的点数据集，也支持手动输入点数据集的OBS路径。 若不选择，则以边数据集中的点作为点数据集来源。
边处理	是	边处理支持如下几种方式： • 允许重复边 • 不允许重复，忽略之后的重复边 • 不允许重复，覆盖之前的重复边

参数	是否必选	说明
离线导入	否	是否离线导入，取值为是或者f否，默认取否。 ● 是： 表示离线导入，导入速度较快，但导入过程中图处于锁定状态，不可读不可写。 ● 否：表示在线导入，相对离线导入，在线导入速度略慢，但导入过程中图并未锁定，可读不可写。
重复边忽略Label	否	重复边的定义，是否忽略Label。取值为是或者否，默认取是。 ● 是：表示重复边定义不包含Label，即用<源点，终点>标记一条边，不包含Label。 ● 否：表示重复边定义包含Label，即用<源点，终点，Label>标记一条边。
日志存储路径	否	用于存储导入图过程中不符合元数据定义的点、边数据集和详细日志。

表 3-164 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.5 MRS Kafka

功能

MRS Kafka主要是查询Topic未消费的消息数。

参数

用户可参考**表3-165**和**表3-166**配置MRS Kafka的参数。

表 3-165 属性参数

参数	是否必选	说明
数据连接	是	选择管理中心中已创建的MRS Kafka连接。
Topic名称	是	选择MRS Kafka中已创建的Topic，使用SDK或者命令行创建。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-166 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.6 Kafka Client

功能

通过Kafka Client向Kafka的Topic中发送数据。

参数

用户可参考表3-167配置Kafka Client节点的参数。

表 3-167 属性参数

参数	是否必选	说明
数据连接	是	选择管理中心中已创建的MRS Kafka连接。
Topic名称	是	选择需要上传数据的Topic，如果有多个partition，默认发送到partition 0。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

参数	是否必选	说明
发送数据	是	发送到Kafka的文本内容。可以直接输入文本或单击 使用EL表达式编辑。

表 3-168 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.7 ROMA FDI Job

功能

通过ROMA FDI Job节点执行一个预先定义的ROMA Connect数据集成任务，实现源端到目标端的数据集成转换。

原理

该节点方便用户启动或者查询FDI任务是否正在运行。

参数

ROMA FDI Job的参数配置，请参考以下内容：

表 3-169 属性参数

参数	是否必选	说明
ROMA实例	是	选择一个已存在的ROMA实例。
FDI任务	是	选择一个已存在的ROMA FDI任务。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-170 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.8 DLI Flink Job

功能

通过DLI Flink Job节点执行一个预先定义的DLI作业，实现实时流式大数据分析。

原理

该节点方便用户启动或者查询DLI作业是否正在运行。当作业类型不是“选择已存在的Flink作业”时，系统会根据在节点中配置的作业情况，进行创建和启动作业。方便用户自定义作业以及作业参数。

参数

DLI Flink Job的参数配置，请参考以下内容：

- 属性参数：
  - **选择已存在的Flink作业**：请参见表3-171。
  - **Flink SQL作业**：请参见表3-172。
  - **Flink自定义作业**：请参见表3-173。
- 表3-174

表 3-171 已存在的 Flink 作业-属性参数

参数	是否必选	说明
作业类型	是	选择“选择已存在的Flink作业”。
作业名称	是	选择一个已存在的DLI Flink作业。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-172 Flink SQL 作业-属性参数

参数	是否必选	说明
作业类型	是	选择“Flink SQL作业”。用户采用编写SQL语句来启动作业。
脚本路径	是	选择需要执行的Flink SQL脚本。如果脚本未创建，请参考新建脚本和开发SQL脚本创建和开发Flink SQL脚本。

参数	是否必选	说明
DLI队列	是	默认选择“共享队列”，用户也可以选择自定义的独享队列。 **说明** 当子用户在创建作业时，子用户只能选择已经被分配的队列。
CUs	是	一个CU是1核4G的资源配置。
并发数	是	并发数是指同时运行Flink SQL作业的任务数。 **说明** 并发数不能大于计算单元（CUs-1）的4倍。
UDF Jar	否	当作业所属集群选择独享集群时，该参数有效。在选择UDF Jar之前，您需要将UDF Jar包上传至OBS桶中，并在“资源管理”页面中新建资源，具体操作请参考**新建资源**。 用户可以在SQL中调用插入Jar包中的自定义函数。
异常自动启动	否	设置是否启动异常自动重启功能，当作业异常时将自动重启并恢复作业。
作业名称	是	填写DLI Flink作业的名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。默认与节点的名称一致。
作业名称添加工作空间前缀	否	设置是否为创建的作业名称添加工作空间前缀。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-173 Flink 自定义作业-属性参数

参数	是否必选	说明
作业类型	是	选择“Flink自定义作业”。
jar包路径	是	用户自定义的程序包。在选择程序包之前，您需要将对应的jar包上传至OBS桶中，并在“资源管理”页面中新建资源，具体操作请参考**新建资源**。

参数	是否必选	说明
入口类	是	指定加载的Jar包类名，如KafkaMessageStreaming。 ● 默认：根据Jar包文件的Manifest文件指定。 ● 指定：需要输入类名并确定类参数列表（参数间用空格分隔）。 **说明** 当类属于某个包时，需携带包路径，例如：packagePath.KafkaMessageStreaming。
入口参数	是	指定类的参数列表，参数之间使用空格分隔。
DLI队列	是	默认选择“共享队列”，用户也可以选择自定义的独享队列。 **说明** 当子用户在创建作业时，子用户只能选择已经被分配的队列。
作业特性	否	选择自定义镜像和对应版本。仅当DLI队列为容器化队列类型时，出现本参数。 自定义镜像是DLI的特性。用户可以依赖DLI提供的Spark或者Flink基础镜像，使用Dockerfile将作业运行需要的依赖（文件、jar包或者软件）打包到镜像中，生成自己的自定义镜像，然后将镜像发布到SWR（容器镜像服务）中，最后在此选择自己生成的镜像，运行作业。 自定义镜像可以改变Spark作业和Flink作业的容器运行环境。用户可以将一些私有能力内置到自定义镜像中，从而增强作业的功能、性能。
CUs	是	一个CU是1核4G的资源配置。
管理节点CU数量	是	设置管理单元的CU数，支持设置1~4个CU数，默认值为1个CU。
并发数	是	并发数是指同时运行Flink SQL作业的任务数。 **说明** 并发数不能大于计算单元（CUs-1）的4倍。
异常自动启动	否	设置是否启动异常自动重启功能，当作业异常时将自动重启并恢复作业。
作业名称	是	填写DLI Flink作业的名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。默认与节点的名称一致。
作业名称添加工作空间前缀	否	设置是否为创建的作业添加工作空间前缀。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-174 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.9 DLI SQL

功能

通过DLI SQL节点传递SQL语句到DLI中执行，实现多数据源分析探索。

原理

该节点方便用户在数据开发模块的周期与实时调度中执行DLI相关语句，可以使用参数变量为用户的数仓进行增量导入，分区处理等动作。

参数

用户可参考**表3-175**，**表3-176**和**表3-177**配置DLI SQL节点的参数。

表 3-175 属性参数

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 • SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 • SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。
数据库名称	是	默认选择SQL脚本中设置的数据库，支持修改。
DLI环境变量	否	• 环境变量配置项需要以"dli.sql."或"spark.sql."开头。 • 环境变量的key为dli.sql.shuffle.partitions或dli.sql.autoBroadcastJoinThreshold时，不能包含>< 符号。 • 如果作业和脚本中同时配置了同名的参数，作业中配置的值会覆盖脚本中的值。
队列名称	是	默认选择SQL脚本中设置的DLI队列，支持修改。 如需新建资源队列，请参考以下方法： • 单击，进入DLI的“队列管理”页面新建资源队列。 • 前往DLI管理控制台进行新建。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
节点名称	是	默认显示为SQL脚本的名称，支持修改。规则如下： 节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
是否记录脏数据	是	单击选择节点是否记录脏数据。 • 是：记录脏数据 • 否：不记录脏数据

表 3-176 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-177 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.10 DLI Spark

功能

通过DLI Spark节点执行一个预先定义的Spark作业。

参数

用户可参考表3-178，表3-179和表3-180配置DLI Spark节点的参数。

表 3-178 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
DLI队列	是	下拉选择需要使用的队列。
作业特性	否	选择自定义镜像和对应版本。仅当DLI队列为容器化队列类型时，出现本参数。 自定义镜像是DLI的特性。用户可以依赖DLI提供的Spark或者Flink基础镜像，使用Dockerfile将作业运行需要的依赖（文件、jar包或者软件）打包到镜像中，生成自己的自定义镜像，然后将镜像发布到SWR（容器镜像服务）中，最后在此选择自己生成的镜像，运行作业。 自定义镜像可以改变Spark作业和Flink作业的容器运行环境。用户可以将一些私有能力内置到自定义镜像中，从而增强作业的功能、性能。。
作业名称	是	填写DLI Spark作业的名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。默认与节点的名称一致。

参数	是否必选	说明
作业运行资源	否	选择作业运行的资源规格： ● 8核32G内存 ● 16核64G内存 ● 32核128G内存
作业主类	是	Spark作业的主类名称。当应用程序类型为“.jar”时，主类名称不能为空。
Spark程序资源包	是	运行spark作业依赖的jars。可以输入jar包名称，也可以输入对应jar包文件的的OBS路径，格式为：obs://桶名/文件夹路径名/包名。在选择资源包之前，您需要先将Jar包及其依赖包上传至OBS桶中，并在“资源管理”页面中新建资源，具体操作请参考**新建资源**。
资源类型	是	支持OBS路径和DLI程序包两种类型的资源。 ● OBS路径：作业执行时，不会上传资源包文件到DLI资源管理，文件的OBS路径会作为启动作业消息体的一部分，推荐使用该方式。 ● DLI程序包：作业执行前，会将资源包文件上传到DLI资源管理。
分组设置	否	当“资源类型”选择了“DLI程序包”时，需要设置。可选择“已有分组”，“创建新分组”或“不分组”。
分组名称	否	当“资源类型”选择了“DLI程序包”时，需要设置。 ● 选择“已有分组”：可选择已有的分组。 ● 选择“创建新分组”：可输入自定义的组名称。 ● 选择“不分组”：不需要选择或输入组名称。
主类入口参数	否	用户自定义参数，多个参数请以Enter键分隔。 应用程序参数支持全局变量替换。例如，在“全局配置”>“全局变量”中新增全局变量key为batch_num，可以使用{{batch_num}}，在提交作业之后进行变量替换。
Spark作业运行参数	否	以“key/value”的形式设置提交Spark作业的属性，多个参数以Enter键分隔。具体参数请参见**Spark Configuration**。 Spark参数value支持全局变量替换。例如，在“全局配置”>“全局变量”中新增全局变量key为custom_class，可以使用"spark.sql.catalog"={{custom_class}}，在提交作业之后进行变量替换。 **说明** Spark作业不支持自定义设置jvm垃圾回收算法。

参数	是否必选	说明
Module名称	否	DLI系统提供的用于执行跨源作业的依赖模块，访问各个不同的服务，选择不同的模块： ● CloudTable/MRS HBase: sys.datasource.hbase ● DDS：sys.datasource.mongo ● CloudTable/MRS OpenTSDB: sys.datasource.opentsdb ● DWS: sys.datasource.dws ● RDS MySQL: sys.datasource.rds ● RDS PostGre: sys.datasource.rds ● DCS: sys.datasource.redis ● CSS: sys.datasource.css DLI内部相关模块： ● sys.res.dli-v2 ● sys.res.dli ● sys.datasource.dli-inner-table
访问元数据	是	是否通过Spark作业访问元数据。具体请参考《 数据湖探索开发指南 》的“使用Spark作业访问DLI元数据”。

表 3-179 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-180 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.11 DWS SQL

功能

通过DWS SQL节点传递SQL语句到DWS中执行。

DWS SQL算子的具体使用教程，请参见**开发一个DWS SQL作业**。

背景信息

该节点方便用户在数据开发模块的批处理作业和实时处理作业中执行DWS相关语句，可以使用参数变量为用户的数据仓库进行增量导入，分区处理等操作。

参数

用户可参考**表3-181**，**表3-182**和**表3-183**配置DWS SQL节点的参数。

表 3-181 属性参数

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 • SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 • SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。

参数	是否必选	说明
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用EL表达式。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
脏数据表	否	填写SQL脚本中定义的脏数据表名称。
匹配规则	-	设置java正则表达式，匹配DWS SQL结果内容，比如表达式为(?<=\()(-*\d+?)(?=,)， 匹配对应SQL结果为(1,"error message")，匹配到的结果为 "1"。
失败匹配值	-	当匹配成功的内容等于设置值时，该节点执行失败。
节点名称	是	默认显示为SQL脚本的名称，支持修改。规则如下： 节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

表 3-182 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-183 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.12 MRS Spark SQL

功能

通过MRS Spark SQL节点实现在MRS中执行预先定义的SparkSQL语句。

参数

用户可参考**表3-184**，**表3-185**和**表3-186**配置MRS Spark SQL节点的参数。

表 3-184 属性参数

参数	是否必选	说明
SQL脚本	是	选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
运行程序参数	否	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。 **说明** 若集群为MRS 1.8.7版本或MRS 2.0.1之后版本，需要配置此参数。 MRS SparkSQL作业的运行程序参数，请参见《MapReduce用户指南》中的“管理现有集群 > 作业管理 > 运行SparkSql作业”。

参数	是否必选	说明
节点名称	是	默认显示为SQL脚本的名称，支持修改。 节点名称只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 节点名称不得包含超出长度限制等。如果节点名称不符合规则，将导致提交MRS作业失败。

表 3-185 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-186 血缘关系

参数	说明
输入	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。

参数	说明
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。
删除	单击  ，删除节点输入功能的参数配置。
查看表详情	单击  ，查看节点输入血缘关系创建数据表的详细信息。
输出	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击  ，修改节点输出功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.13 MRS Hive SQL

功能

通过MRS Hive SQL节点执行数据开发模块中预先定义的Hive SQL脚本。

参数

用户可参考**表3-187**，**表3-188**和**表3-189**配置MRS Hive SQL节点的参数。

表 3-187 属性参数

参数	是否必选	说明
SQL脚本	是	选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
运行程序参数	否	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。 **说明** 若集群为MRS 1.8.7版本或MRS 2.0.1之后版本，需要配置此参数。 MRS Hive SQL作业的运行程序参数，请参见《MapReduce服务(MRS) 用户指南》的“管理现有集群 > 作业管理 > 运行HiveSql作业”。
节点名称	是	默认显示为SQL脚本的名称，支持修改。规则如下： 节点名称只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 节点名称不得包含超出长度限制等。如果节点名称不符合规则，将导致提交MRS作业失败。

表 3-188 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-189 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.14 MRS Presto SQL

功能

通过MRS Presto SQL节点执行数据开发模块中预先定义的Presto SQL脚本。

参数

用户可参考**表3-190**，**表3-191**和**表3-192**配置MRS Presto SQL节点的参数。

表 3-190 属性参数

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 • SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 • SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
模式	是	默认选择SQL脚本中设置的数据库，支持修改。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。

参数	是否必选	说明
节点名称	是	默认显示为SQL脚本的名称，支持修改。 节点名称只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 节点名称不得包含超出长度限制等。如果节点名称不符合规则，将导致提交MRS作业失败。

表 3-191 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-192 血缘关系

参数	说明
输入	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。

参数	说明
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。
删除	单击  ，删除节点输入功能的参数配置。
查看表详情	单击  ，查看节点输入血缘关系创建数据表的详细信息。
输出	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击 ，修改节点输出功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.15 MRS Spark

功能

通过MRS Spark节点实现在MRS中执行预先定义的Spark作业。

参数

用户可参考**表3-193**，**表3-194**和**表3-195**配置MRS Spark节点的参数。

表 3-193 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
MRS集群名	是	选择MRS集群。 如需新建集群，请参考以下方法： • 单击，进入“集群列表”页面新建MRS集群。 • 前往MRS管理控制台进行新建。
Spark作业名称	是	MRS作业名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。 **说明** 作业名称不得包含超出长度限制等。如果作业名称不符合规则，将导致提交MRS作业失败。
Jar包资源	是	选择Jar包。在选择Jar包之前，您需要先将Jar包上传至OBS桶中，并在“资源管理”页面中新建资源将Jar包添加到资源管理列表中，具体操作请参考**新建资源**。
Jar包参数	否	Jar包的参数。

参数	是否必选	说明
运行程序参数	否	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。 **说明** 若集群为MRS 1.8.7版本或MRS 2.0.1之后版本，需要配置此参数。 MRS Spark作业的运行程序参数，请参见《MapReduce服务(MRS) 用户指南》“管理现有集群 > 作业管理 > 运行Spark作业”章节。
输入数据路径	否	选择输入数据所在的路径。
输出数据路径	否	选择输出数据存储的路径。

表 3-194 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-195 血缘关系

参数	说明
输入	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。

参数	说明
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。
删除	单击  ，删除节点输入功能的参数配置。
查看表详情	单击  ，查看节点输入血缘关系创建数据表的详细信息。
输出	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击  ，修改节点输出功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.16 MRS Spark Python

功能

通过MRS Spark Python节点实现在MRS中执行预先定义的Spark Python作业。

MRS Spark Python算子的具体使用教程，请参见**开发一个MRS Spark Python作业**。

参数

用户可参考**表3-196**，**表3-197**和**表3-198**配置MRS Spark Python节点的参数。

表 3-196 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
MRS集群名	是	选择支持spark python的mrs集群。MRS只有特定版本支持spark python的集群，请先测试运行，保证集群支持。 如需新建集群，请参考以下方法： • 单击，进入“集群列表”页面新建MRS集群。 • 前往MRS管理控制台进行新建。 如何新建集群，请参见《MapReduce服务(MRS) 使用指南》中创建集群。
作业名称	是	MRS作业名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。 **说明** 作业名称不得包含超出长度限制等。如果作业名称不符合规则，将导致提交MRS作业失败。
参数	是	输入MRS的执行程序参数，多个参数间使用Enter键分隔。
属性	否	输入key=value格式的参数，多个参数间使用Enter键分割。

表 3-197 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-198 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.17 MRS Flink Job

功能

通过MRS Flink节点实现在MRS中执行预先定义的Flink作业。

参数

用户可参考表3-199和表3-200配置MRS Flink节点的参数。

表 3-199 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
MRS集群名	是	选择MRS集群。 如需新建集群，请参考以下方法： • 单击，进入“集群列表”页面新建MRS集群。 • 前往MRS管理控制台进行新建。
Flink作业名称	是	MRS作业名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。 **说明** 作业名称不得包含超出长度限制等。如果作业名称不符合规则，将导致提交MRS作业失败。
Flink作业资源包	是	选择Jar包。在选择Jar包之前，您需要先将Jar包上传至OBS桶中，并在“资源管理”页面中新建资源将Jar包添加到资源管理列表中，具体操作请参考新建资源。
Flink作业执行参数	否	Flink作业执行的程序关键参数，该参数由用户程序内的函数指定。多个参数间使用空格隔开。

参数	是否必选	说明
运行程序参数	否	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。 **说明** 若集群为MRS 1.8.7版本或MRS 2.0.1之后版本，需要配置此参数。 MRS Flink作业的运行程序参数，请参见《MapReduce服务(MRS) 用户指南》的“管理现有集群 > 作业管理 > 运行Flink作业”章节。
输入数据路径	否	选择输入数据所在的路径。
输出数据路径	否	选择输出数据存储的路径。

表 3-200 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.18 MRS MapReduce

功能

通过MRS MapReduce节点实现在MRS中执行预先定义的MapReduce程序。

参数

用户可参考表3-201和表3-202配置MRS MapReduce节点的参数。

表 3-201 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
MRS集群名	是	选择MRS集群。 如需新建集群，请参考以下方法： • 单击，进入“集群列表”页面新建MRS集群。 • 前往MRS管理控制台进行新建。
MapReduce作业名称	是	MRS作业名称，只能包含英文字母、数字、“_”，且长度为1~64个字符。 **说明** 作业名称不得包含超出长度限制等。如果作业名称不符合规则，将导致提交MRS作业失败。
Jar包资源	是	选择Jar包。在选择Jar包之前，您需要先将Jar包上传至OBS桶中，并在“资源管理”页面中新建资源将Jar包添加到资源管理列表中，具体操作请参考新建资源。
Jar包参数	否	Jar包的参数。
输入数据路径	否	选择输入数据所在的路径。
输出数据路径	否	选择输出数据存储的路径。

表 3-202 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.19 CSS

功能

通过CSS节点执行云搜索请求，实现在线分布式搜索功能。

参数

用户可参考**表3-203**和**表3-204**配置CSS节点的参数。

表 3-203 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
CloudSearch集群	是	选择CloudSearch集群，该集群已在CloudSearch服务中创建好。目前仅支持使用5.5.1版本的集群。

参数	是否必选	说明
CDM集群名称	是	选择CDM集群。CDM集群提供代理，转发相关请求。 如果下拉框中未提供CDM集群，请访问CDM管理控制台创建集群。
请求类型	是	支持以下请求类型： ● GET ● POST ● PUT ● HEAD ● DELETE
请求参数	否	请求参数。 假设用户需要查询dlf_search索引中dlfdata映射类型的信息，请求参数可填写为： /dlf_search/dlfdata/_search
请求体	否	Json格式的请求消息体。
CloudSearch输出路径	否	选择输出数据的存储路径。

表 3-204 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： • 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 • 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 • 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 • 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.20 Shell

功能

通过Shell节点执行用户指定的Shell脚本。

说明

Shell节点的后续节点可以通过EL表达式**#{Job.getNodeOutput()}**，获取Shell脚本最后4000字符的标准输出。

使用示例：

获取某个Shell脚本（脚本名称为shell_job1）输出值包含“<name>jack<name1>”的内容，EL表达式如下所示：

```
#{StringUtil.substringBetween(Job.getNodeOutput("shell_job1"),"<name>","<name1>")}
```

参数

用户可以参考**表3-205**和**表3-206**配置Shell节点的参数。

表 3-205 属性参数

参数	是否必选	说明
Shell或脚本	是	可以选择Shell语句或Shell脚本。 • Shell语句 单击“Shell语句”参数下的文本框，在“Shell语句”页面输入需要执行的Shell语句。 • Shell脚本 在“脚本路径”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发Shell脚本**先创建和开发脚本。 **说明** 若选择Shell语句方式，数据开发模块将无法解析您输入Shell语句中携带的参数。
主机连接	是	选择执行Shell脚本的主机。
参数	否	填写执行Shell脚本时，向脚本传递的参数，参数之间使用空格分隔，例如：a b c。此处的“参数”需要在Shell脚本中引用，否则配置无效。
交互式输入	否	填写交互式参数，即执行Shell脚本的过程中，需要用户输入的交互式信息（例如密码）。交互式参数之间以回车符分隔，Shell脚本根据交互情况按顺序读取参数值。
节点名称	是	节点名称，只能包含英文字母、数字、中文字符、中划线、下划线、/、<>和点号，且长度小于等于128个字符。

表 3-206 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 • 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） • 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.21 RDS SQL

功能

通过RDS SQL节点传递SQL语句到RDS中执行。

参数

用户可参考**表3-207**和**表3-208**配置RDS SQL节点的参数。

表 3-207 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
数据连接	是	选择数据连接。
数据库	是	填写数据库名称，该数据库已创建好，建议不要使用默认数据库。

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 ● SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 ● SQL脚本 在“脚本路径”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。

表 3-208 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.22 ETL Job

功能

通过ETL Job节点可以从指定数据源中抽取数据，经过数据准备对数据预处理后，导入到目标数据源。

参数

用户可参考**表3-209**，**表3-210**和**表3-211**配置ETL Job节点的参数。

表 3-209 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
ETL配置	是	单击 配置需要转换的源端数据和目的端数据。 当前支持的源端数据为DLI类型、OBS类型和MySQL类型。 • 当源端数据为DLI类型时，支持的目的端数据类型为DWS、GES、CSS、OBS、DLI。 • 当源端数据为MySQL类型时，支持的目的端数据类型为MySQL。 • 当源端数据为OBS类型时，支持的目的端数据类型为DLI、DWS。 **须知** • DLI到DWS端的数据转换： 因为数据开发模块调用DWS的集群时，需要走网络代理。所以导入数据到DWS时，需要提前先在数据开发模块中创建DWS的数据连接。 DLI导入数据到DWS时，DWS的表需要先创建好。 • DLI到CSS端的数据转换： DLI导入数据到CSS集群时，需要在DLI侧提前创建好关联对应CSS集群的跨源连接，请参见《数据湖探索用户指南》。
SQL模板	否	单击“配置”按钮获取SQL模板。

表 3-210 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-211 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.23 Python

功能

通过Python节点执行Python语句。

使用Python节点前，需确认对应主机连接的主机配有用于执行Python脚本的环境。

说明

Python节点暂不支持脚本参数和作业参数。

参数

用户可以参考**表3-212**和**表3-213**配置Python节点的参数。

表 3-212 属性参数

参数	是否必选	说明
Python或脚本	是	可以选择Python语句或Python脚本。 • Python语句 单击“Python语句”参数下的文本框，在“Python语句”页面输入需要执行的Python语句。 • Python脚本 在“脚本路径”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发Python脚本**先创建和开发脚本。 **说明** 若选择Python语句方式，数据开发模块将无法解析您输入Python语句中携带的参数。
主机连接	是	选择执行Python语句的主机。需确认该主机配有用于执行Python脚本的环境。

参数	是否必选	说明
节点名称	是	节点名称，只能包含英文字母、数字、中文字符、中划线、下划线、/、<>和点号，且长度小于等于128个字符。

表 3-213 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.24 Create OBS

约束限制

该功能依赖于OBS服务。

功能

通过Create OBS节点在OBS服务中创建桶和目录。

参数

用户可参考表3-214和表3-215配置Create OBS节点的参数。

表 3-214 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
OBS路径	是	创建OBS桶或目录的路径。 ● 创建桶：在“//”后输入OBS桶名称，OBS桶名称不允许重名。 ● 创建OBS目录：选择需要创建目录的路径，在路径后输入“/目录名”，目录名不允许重名。

表 3-215 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： • 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 • 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 • 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 • 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.25 Delete OBS

约束限制

该功能依赖于OBS服务。

功能

通过Delete OBS节点在OBS服务中删除桶和目录。

参数

用户可参考**表3-216**和**表3-217**配置Delete OBS节点的参数。

表 3-216 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
OBS路径	是	删除OBS桶或目录的路径。 **说明** 删除的文件将无法恢复，如需保留文件，请在删除前备份该桶下的数据。

表 3-217 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.26 OBS Manager

约束限制

该功能依赖于OBS服务。

功能

通过OBS Manager节点可以将OBS文件移动或复制到指定目录下。

参数

用户可参考**表3-218**，**表3-219**和**表3-220**配置OBS Manager节点的参数。

表 3-218 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
操作类型	是	通过节点可以执行的操作： ● 移动文件：将源文件或目录，移动到新目录中。 ● 复制文件：复制源文件或目录。 ● 重命名文件：重命名文件仅支持最后一级目录或文件重命名。 如重命名目录时，源文件或目录：obs://test/a/b/c/，目的目录：obs://test/a/b/d/；重命名文件时，源文件或目录：obs://test/a/b/hello.txt，目的目录：obs://test/a/b/bye.txt ● 监测文件：监测文件或目录是否存在，如不存在则此节点运行失败，否则运行成功。
源文件或目录	是	OBS桶中需要被管理的OBS文件或所在目录。
目的目录	是	存放待移动或复制OBS文件的新目录
文件过滤器	否	输入文件过滤的通配符，满足该过滤条件的文件才会被移动或复制。当不指定该参数时，默认移动所有源文件。例如：匹配文件名以.csv结尾的文件，输入通配符*.csv。

表 3-219 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

表 3-220 血缘关系

参数	说明
输入	

参数	说明
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击…，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击…，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击…，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击…，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击…，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击…，在弹出的“表名”窗口选择DLI的数据表。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。 ● DWS – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DWS的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DWS的数据库。 – schema（必选）：单击···，在弹出的“schema”窗口选择DWS的数据库模式。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DWS的数据表。 ● OBS – 路径（必选）：单击···，在弹出的“OBS文件浏览”窗口选择OBS路径。 ● CSS – 集群名称（必选）：单击···，在弹出的“CloudSearch集群”窗口选择CloudSearch集群。 – 索引名称（必选）：输入CSS类型的索引名称。 ● HIVE – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择HIVE的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择HIVE的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择HIVE的数据表。 ● CUSTOM – 名称（必选）：输入CUSTOM类型的名称。 – 属性（必选）：输入CUSTOM类型的属性，可新增不止一条。 ● DLI – 连接名称（必选）：单击···，在弹出的“连接名称”窗口选择DLI的数据连接。 – 数据库（必选）：单击···，在弹出的“数据库”窗口选择DLI的数据库。 – 表名（必选）：单击···，在弹出的“表名”窗口选择DLI的数据表。

参数	说明
确定	单击“确认”，保存节点输出功能的参数配置。
取消	单击“取消”，取消节点输出功能的参数配置。
编辑	单击，修改节点输出功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输出功能的参数配置。
查看表详情	单击，查看节点输出血缘关系创建数据表的详细信息。

3.4.9.27 Open/Close Resource

功能

通过Open/Close Resource节点按需开启或关闭服务。

参数

用户可参考表3-221和表3-222配置Open/Close Resource节点的参数。

表 3-221 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
服务	是	选择需要开机/关机的服务： • ECS • CDM
开关机设置	是	选择开关机类型： • 开 • 关
开关机对象	是	选择需要开机/关机的具体对象，例如开启某个CDM集群。

表 3-222 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.28 Sub Job

功能

通过Sub Job节点可以调用另外一个批处理作业。

参数

用户可参考**表3-223**和**表3-224**配置Sub Job节点的参数。

表 3-223 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
子作业名称	是	选择需要调用的子作业名称。 **说明** 您只能选择已存在的批处理作业名称，此批处理作业不能为作业本身，并且该批处理作业为不包含Sub Job节点的作业。
子作业参数名称	是/否	● 当节点属性中子作业参数配置为空时，子作业使用自身参数变量执行。父作业的“子作业参数名称”不显现。 ● 当节点属性中子作业参数配置了数据时，子作业将使用配置参数变量执行。此时父作业的“子作业参数名称”显现，并且节点属性中子作业参数配置的数据或者EL表达式，将根据父作业的环境变量读取替换。

表 3-224 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

参数	是否必选	说明
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.29 For Each

功能

该节点可以指定一个子作业循环执行，并支持用一个数据集对子作业中的变量进行循环替换。

参数

用户可参考表3-225配置For Each节点的参数。

表 3-225 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
循环执行的子作业	是	选择需要循环执行的子作业。
数据集	是	For循环算子需要定义一个数据集，这个数据集用来循环替换子作业中的变量，数据集的一行数据会对应一个子作业实例。数据集的来源包括： ● 来自于上游节点的输出。例如DLI SQL、Hive SQL、Spark SQL的select语句，或者Shell节点的echo等。使用EL表达式为：#{Job.getNodeOutput('preNodeName')}，即前一个节点的输出值。 ● 来自于给定的数组。如一维数组：[['001'],['002'],['003']] 。

参数	是否必选	说明
子作业并发数	是	循环产生的子作业可以并发执行，您可设置并发数。
子作业实例名称后缀	否	For循环生成的子任务名称：For循环节点名称 + 下划线 + 后缀。 后缀可配置，如果不配置，则按照数字顺序依次递增。
作业运行参数	否	仅当子作业配置作业参数后，出现该参数。 ● 节点属性中子作业参数配置为空时，子作业使用自身参数变量执行。 ● 节点属性中子作业参数配置后，将使用配置参数变量执行。节点属性中子作业参数配置的方法或者EL表达式，将根据父作业的环境变量读取替换。

表 3-226 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： ● 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 ● 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 ● 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 ● 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.30 SMN

功能

通过SMN节点向用户发送通知消息。

参数

用户可参考**表3-227**和**表3-228**配置SMN节点的参数。

表 3-227 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
主题名称	是	选择消息的主题，该主题已在SMN服务中创建好。
消息标题	否	自定义消息的标题，长度必须少于512个字符。
消息类型	是	选择消息的发送格式。 ● 文本消息：按文本格式发送的消息。 ● JSON消息：按JSON格式发送的消息，用户可对不同的订阅者类型发送不同的消息。 – 手动输入JSON格式的消息：在“消息内容”直接输入。 – 通过工具自动生成JSON格式的消息：单击“生成JSON消息”，在弹出的对话框中填写“消息”和选择“协议”。 ● 模板消息：按模板格式发送的消息，即固定格式的消息，可以通过tag的方式来处理变量的部分。 – 手动输入模板格式的消息：在“消息内容”直接输入。 – 通过工具自动生成模板格式的消息：单击“生成模板消息”，在弹出的对话框中，选择“模板名称”，并设置{tag}的值。

参数	是否必选	说明

消息内容	是	填写消息的内容，不同消息类型的填写要求如下：
		● 文本消息：大小不超过10KB。
		● JSON消息：JSON消息中必须有Default协议，大小不超过10KB。
		示例如下：
		{
		"default": "Dear Sir or Madam, this is a default message.",
		"email": "Dear Sir or Madam, this is an email message.",
		"http": "{'message':'Dear Sir or Madam, this is an HTTP message.'}",
		"https": "{'message':'Dear Sir or Madam, this is an HTTPS message.'}",
		"sms": "This is an SMS message."
		}
		● 模板消息：大小不超过10KB。
		示例如下：
		"message_template_name":"confirm_message",
		"tags":{
		"topic_urn":"urn:smn:regionId:xxxx:SMN_01"
		}
		其中，“message_template_name”为模板名称，“tags”为模板中所有的tag标签。
		如需了解更多SMN的配置说明，请参见《消息通知服务用户指南》。

表 3-228 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 ● 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行节点。 **说明** 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。

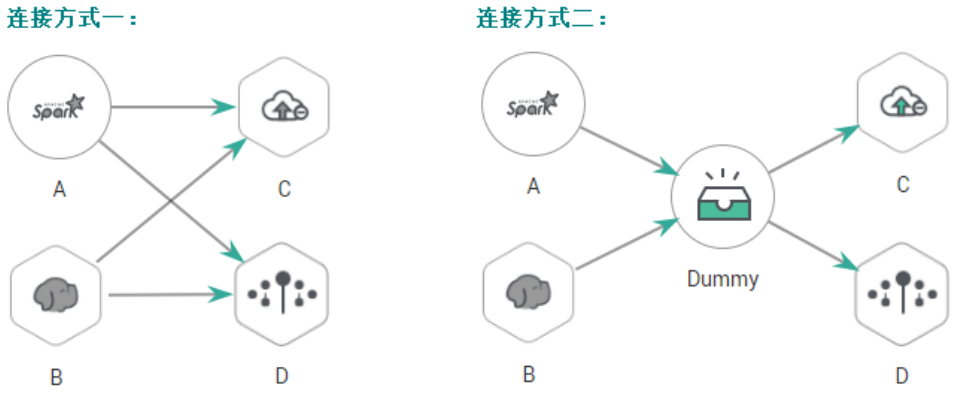
参数	是否必选	说明
失败策略	是	节点执行失败后的操作： • 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 • 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败成功”。 • 挂起当前作业执行计划：暂停当前作业运行，当前作业实例状态显示为“等待运行”。 • 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。
空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

3.4.9.31 Dummy

功能

Dummy节点是一个空的节点，不执行任何操作。用于简化节点的连接视图，便于用户理解复杂节点流的连接关系，示例如图3-229所示。

图 3-229 连接方式对比



参数

用户可参考表3-229配置Dummy节点的参数。

表 3-229 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。

3.4.10 EL 表达式参考

3.4.10.1 表达式概述

数据开发模块作业中的节点参数可以使用表达式语言（Expression Language，简称EL），根据运行环境动态生成参数值。可以根据Pipeline输入参数、上游节点输出等决定是否执行此节点。数据开发模块 EL表达式使用简单的算术和逻辑计算，引用内嵌对象，包括作业对象和一些工具类对象。

作业对象：提供了获取作业中上一个节点的输出消息、作业调度计划时间、作业执行时间等属性和方法。

工具类对象：提供了一系列字符串、时间、JSON操作方法，例如从一个字符串中截取一个子字符串、时间格式化等。

语法

表达式的语法：

```
#{expr}
```

其中，“expr”指的是表达式。“#”和“{}”是数据开发模块 EL中通用的操作符，这两个操作符允许您通过数据开发模块内嵌对象访问作业属性。

举例

在Rest Client节点的参数“URL参数”中使用EL表达式“tableName=#{JSONUtil.path(Job.getNodeOutput("get_cluster"),"tables[0].table_name")}”。

表达式说明如下：

1. 获取作业中“get_cluster”节点的执行结果（“Job.getNodeOutput("get_cluster")”），执行结果是一个JSON字符串。
2. 通过JSON路径（“tables[0].table_name”），获取JSON字符串中字段的值。

调试方法介绍

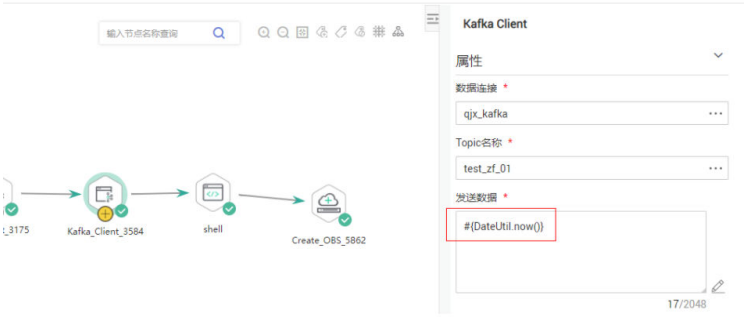
下面为您介绍几种EL表达式的调试方法，能够在调试过程中方便地看到替换结果。

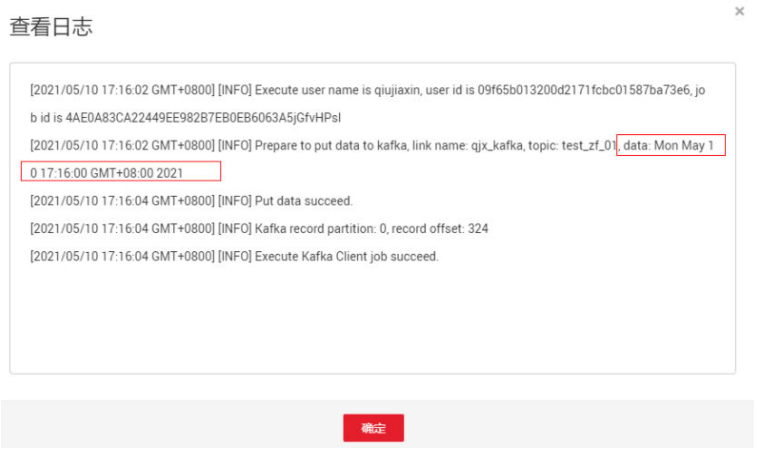
后文以#{DateUtil.now()}表达式为例进行介绍。

1. 使用DIS Client节点。

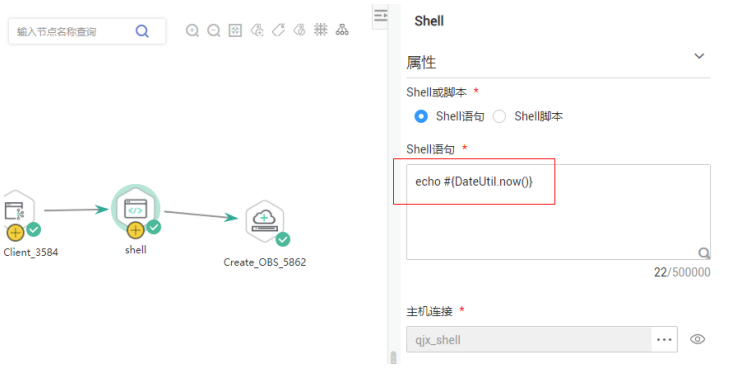
- 前提：您需要具备DIS通道。
- 方法：选择DIS Client节点，将EL表达式直接写在要发送的数据中，点击“测试运行”，然后在节点上右键查看日志，日志中会把EL表达式的值打印出来。

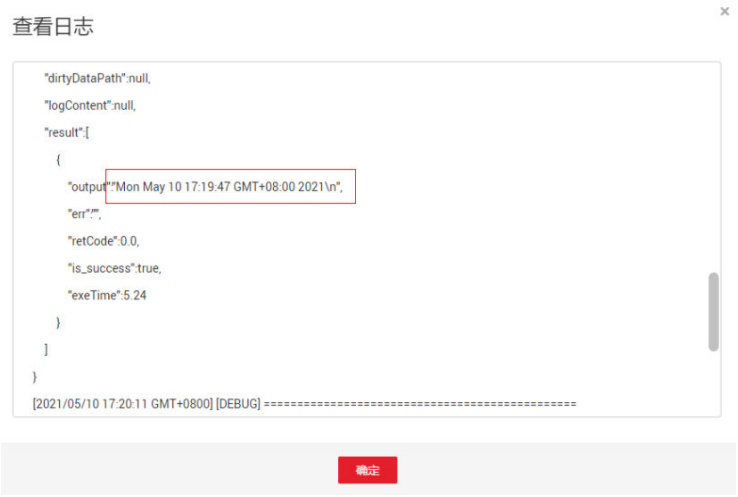
2. 使用Kafka Client节点。
   - 前提：您需要具备MRS集群，且集群有Kafka组件。
   - 方法：选择Kafka Client节点，将EL表达式直接写在要发送的数据中，点击“测试运行”，然后在节点上右键查看日志，日志中会把EL表达式的值打印出来。





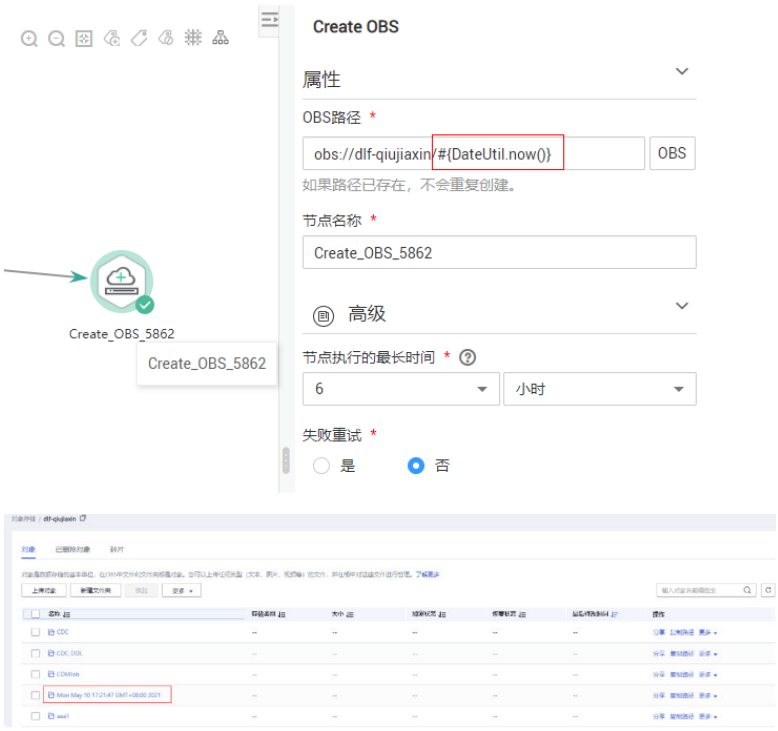
3. 使用Shell节点。
   - 前提：您需要具备ECS弹性云服务器。
   - 方法：创建一个主机连接，将EL表达式直接通过echo打印出来，点击“测试运行”之后查看日志，日志中会打印出EL表达式的值。





4. 使用Create OBS节点。

   如果上述方法均不可用，则可以通过Create OBS去创建一个OBS目录，目录名称就是EL表达式的值，点击“测试运行”后，再去OBS界面查看创建出来的目录名称。



3.4.10.2 基础操作符

EL表达式支持大部分Java提供的算术和逻辑操作符。

操作符列表

表 3-230 基础操作符

操作符	描述
.	访问一个Bean属性或者一个映射条目
[]	访问一个数组或者链表的元素
()	组织一个子表达式以改变优先级
+	加
-	减或负
*	乘
/ 或 div	除
% 或 mod	取模
== 或 eq	测试是否相等
!= 或 ne	测试是否不等
< 或 lt	测试是否小于
> 或 gt	测试是否大于

操作符	描述
<= 或 le	测试是否小于等于
>= 或 ge	测试是否大于等于
&& 或 and	测试逻辑与
|| 或 or	测试逻辑或
！ 或 not	测试取反
empty	测试是否空值
?:	类似if else表示式。如果?前面的语句为true，返回?和:之间的表达式的值；否则返回:后面的值。

举例

如果变量a为空，返回default，否则返回a本身。EL表达式如下：

```
#{empty a?"default":a}
```

3.4.10.3 日期和时间模式

EL表达式中的日期和时间可以按用户指定的格式进行显示，日期和时间格式由日期和时间模式字符串指定。日期和时间模式字符串由A到Z、a到z的非引号字母组成，字母的含义如表3-231所示。

表 3-231 字母含义

字母	描述	示例
G	纪元标记	AD
y	年	2001
M	年中的月份	July 或 07
d	月份中的日期	10
h	12小时制（1~12）的小时	12
H	24小时制（0~23）的小时	22
m	分钟数	30
s	秒数	55
S	毫秒数	234
E	星期几	Mon、Tue、Wed、Thu、Fri、Sat或Sun
D	年中的日期	360

字母	描述	示例
F	月份中第几周周几	2(second Wed. in July)
w	年中的第几周	40
W	月份中的第几周	1
a	A.M./P.M.标记	PM
k	24小时制（1~24）的小时	24
K	12小时制（0~11）的小时	10
z	时区	Eastern Standard Time
'	文字定界符	无示例
''	单引号	无示例

举例

获取作业计划调度时间的前一天日期，EL表达式如下：

```
#{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")}
```

3.4.10.4 Env 内嵌对象

Env内嵌对象提供了获取环境变量值的方法。

方法

表 3-232 方法说明

方法	描述
String get(String name)	获取指定名称环境变量值。

举例

获取环境变量名称为**test**的参数值，EL表达式如下：

```
#{Env.get("test")}
```

3.4.10.5 Job 内嵌对象

Job为作业对象，提供了获取作业中上一节点的输出消息、作业调度计划时间、作业执行时间等属性和方法。

属性和方法

表 3-233 属性说明

属性	类型	描述
name	String	作业名称。
planTime	java.util.Date	作业调度计划时间，即周期调度配置的时间，例如每天凌晨1:01调度作业。
startTime	java.util.Date	作业执行时间，有可能与planTime同一个时间，也有可能晚于planTime（由于作业引擎繁忙等）。
eventData	String	当作业使用事件驱动调度时，从通道获取的消息。
projectId	String	当前数据开发模块所处项目ID。

表 3-234 方法说明

方法	描述
String getNodeStatus(String nodeName)	获取指定节点运行状态，成功状态返回success，失败状态返回fail。 例如，判断节点是否运行成功，可以使用如下判断条件，其中test为节点名称： #{(Job.getNodeStatus("test")) == "success" }
String getNodeOutput(String nodeName)	获取指定节点的输出。此方法只能获取前面依赖节点的输出。
String getParam(String key)	获取作业参数。 注意此方法只能直接获取当前作业里配置的参数值，并不能获取到父作业传递过来的参数值，也不能获取到工作空间里面配置的全局变量，作用域仅为本作业。 这种情况下建议使用表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。
String getPlanTime(String pattern)	获取指定pattern的计划时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。
String getYesterday(String pattern)	获取执行pattern的计划时间前一天的时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。
String getLastHour(String pattern)	获取执行pattern的计划时间前一小时的时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。

方法	描述
String getRunningData(String nodeName)	获取指定节点运行中记录的数据。此方法只能获取前面依赖节点的输出。当前只支持获取DLI SQL节点运行中记录的DLI作业id。例如，想要获取DLI节点第3条语句的job ID（DLI节点名为DLI_INSERT_DATA），可以这样使用：#{JSONUtil.path(Job.getRunningData("DLI_INSERT_DATA"),"jobIds[2]")}。
String getInsertJobId(String nodeName)	返回指定DLI SQL或Transform Load节点第一个DLI Insert SQL语句的作业ID，不指定参数nodeName时，获取前面一个节点第一个DLI Insert SQL语句的作业ID，如果无法获取到作业ID，返回null值。

举例

获取作业中节点名称为**test**的输出，EL表达式如下：

```
#{Job.getNodeOutput("test")}
```

3.4.10.6 StringUtil 内嵌对象

StringUtil内嵌对象提供了一系列字符串操作方法，例如从一个字符串中截取一个子字符串。

StringUtil内部是由org.apache.commons.lang3.StringUtils实现的，具体使用方法请参考**appache commons文档**。

举例

假设变量a为字符串No.0010，返回“.”后面的子字符串，EL表达式如下：

```
#{StringUtil.substringAfter(a,".")}
```

3.4.10.7 DateUtil 内嵌对象

DateUtil内嵌对象提供了一系列时间格式化、时间计算方法。

方法

表 3-235 方法说明

方法	描述
String format(Date date, String pattern)	将Date类型时间按指定pattern格式为字符串。
Date addMonths(Date date, int amount)	给date添加指定月数后，返回新Date对象，amount可以是负数。
Date addDays(Date date, int amount)	给date添加指定天数后，返回新Date对象，amount可以是负数。

方法	描述
Date addHours(Date date, int amount)	给date添加指定小时数后，返回新Date对象，amount可以是负数。
Date addMinutes(Date date, int amount)	给date添加指定分钟数后，返回新Date对象，amount可以是负数。
int getDay(Date date)	从date获取天，例如：date为2018-09-14，则返回14。
int getMonth(Date date)	从date获取月，例如：date为2018-09-14，则返回9。
int getYear(Date date)	从date获取年，例如：date为2018-09-14，则返回2018。
Date now()	返回当前时间。
long getTime(Date date)	将Date类型时间转换为long类型。
Date parseDate(String str, String pattern)	字符串按pattern转换为Date类型，pattern为日期、时间模式，请参考**日期和时间模式**。

举例

以作业调度计划时间的前一天时间作为子目录名称，生成一个OBS路径，EL表达式如下：

```
#{"obs://test/"+DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")}
```

3.4.10.8 JSONUtil 内嵌对象

JSONUtil内嵌对象提供了JSON对象方法。

方法

表 3-236 方法说明

方法	描述
Object parse(String jsonStr)	将json字符串转换为对象。
String toString(Object jsonObject)	将对象转换为json字符串。
Object path(String jsonStr,String jsonPath)	返回json字符串指定路径下的字段值。类似于XPath，path方法可以通过路径检索或设置JSON，其路径中可以使用.或[]等访问成员、数值，例如：tables[0].table_name。

举例

字符串变量str的内容如下：

```
{
	"cities": [{
		"name": "city1",
		"areaCode": "1000"
	},
	{
		"name": "city2",
		"areaCode": "2000"
	},
	{
		"name": "city3",
		"areaCode": "3000"
	}]
}
```

获取city1的电话区号，EL表达式如下：

```
#{JSONUtil.path(str,"cities[0].areaCode")}
```

3.4.10.9 Loop 内嵌对象

使用Loop内嵌对象可获取For Each数据集中的数据。

属性

表 3-237 属性说明

属性	类型	描述
dataArray	String	For循环算子输入的数据集，是一个二维数组。
current	String	For循环算子当前遍历到的数据行,是一个一维数组。
offset	Int	For循环当前的偏移量，从0开始。 Loop.dataArray[Loop.offset] = Loop.current。

举例

Foreach算子循环获取前一节点输出（二维数组）的第一列，EL表达式如下：

```
#{Loop.current[0]}
```

3.4.10.10 OBSUtil 内嵌对象

OBSUtil内嵌对象提供了一系列针对OBS的操作方法，例如判断OBS文件或目录是否存在。

方法

表 3-238 方法说明

方法	说明
boolean isExistOBSPath(String obsPath)	判断OBS文件或目录（目录请以“/”结尾）是否存在，存在返回true，不存在返回false。

举例

- 判断OBS目录是否存在，目录请以“/”结尾，EL表达式如下：
  #{OBSUtil.isExistOBSPath("obs://test/jobs/")}
- 判断OBS文件是否存在，EL表达式如下：
  #{OBSUtil.isExistOBSPath("obs://test/jobs/job.log")}

3.4.10.11 表达式使用示例

通过本示例，用户可以了解数据开发模块 EL表达式的如下应用：

- 如何在数据开发模块的SQL脚本中使用变量?
- 作业如何传递参数给SQL脚本变量?
- 在参数中如何使用EL表达式?

背景信息

使用数据开发模块的作业编排和作业调度功能，每日通过统计交易明细表，生成日交易统计报表。

本示例涉及的数据表如下所示：

- trade_log：记录每一笔交易数据。
- trade_report：根据trade_log统计产生，记录每日交易汇总。

前提条件

- 已建立DLI的数据连接，以“dli_demo”数据连接为例。
  如未建立，请参考**创建数据连接**进行操作。
- 已在DLI中创建数据库，以“dli_db”数据库为例。
  如未创建，请参考**新建数据库**进行操作。
- 已在“dli_db”数据库中创建数据表trade_log和trade_report。
  如未创建，请参考**新建数据表**进行操作。

操作步骤

步骤1 新建和开发SQL脚本。

1. 在数据开发模块控制台的左侧导航栏，选择“数据开发 > 脚本开发”。
2. 进入右侧区域页面，选择“新建SQL脚本 > DLI”。
3. 进入SQL脚本开发页面，在脚本属性栏选择“数据连接”、“数据库”、“资源队列”。
4. 在脚本编辑器中输入以下SQL语句。

```
INSERT OVERWRITE TABLE trade_report
SELECT
  sum(trade_count),
  '${yesterday}'
FROM
  trade_log
where
  date_format(trade_time, 'yyyy-MM-dd') = '${yesterday}'
```

5. 单击，将脚本的名称设置为“generate_trade_report”。

步骤2 新建和开发作业。

1. 在数据开发模块控制台的左侧导航栏，选择“数据开发 > 作业开发”。
2. 进入右侧区域页面，单击“新建作业”，新建一个名称为“job”的空作业。
3. 进入作业开发页面，将DLI SQL节点拖至画布中，单击其图标并配置“节点属性”。

   关键属性说明：

   - SQL脚本：关联**步骤1**中开发完成的SQL脚本“generate_trade_report”。
   - 数据库名称：自动填写SQL脚本“generate_trade_report”中选择的数据库。
   - 队列名称：自动填写SQL脚本“generate_trade_report”中选择的资源队列。
   - 脚本参数：显示SQL脚本“generate_trade_report”中的参数“yesterday”，输入以下EL表达式作为其参数值。

     ```
     #{Job.getYesterday("yyyy-MM-dd")}
     ```

     EL表达式说明：Job为作业对象，通过getYesterday方法获取作业计划执行时间前一天的时间，时间格式为yyyy-MM-dd。

     假设作业计划执行时间为2018/9/26 01:00:00，这个表达式计算结果是2018-09-25，该计算结果将替换**SQL脚本**中的${yesterday}参数。替换后的SQL内容如下：

     ```
     INSERT OVERWRITE TABLE trade_report
     SELECT
       sum(trade_count),
       '2018-09-25'
     FROM
       trade_log
     where
       date_format(trade_time, 'yyyy-MM-dd') = '2018-09-25'
     ```

4. 单击，测试运行作业。
5. 作业测试无问题后，单击，保存作业配置。

----结束

3.4.11 使用教程

3.4.11.1 作业依赖详解

周期调度作业支持设置调度周期符合条件的作业为依赖作业。设置依赖作业的操作详情请参考《 DataArts Studio用户指南 》手册中的“数据开发 - 作业开发 -调度作业”章节。

例如周期调度作业A，可设置其依赖作业为作业B，如图3-230所示进行配置。则仅当其依赖的作业B在某段时间内所有实例运行完成、且不存在失败实例时，才开始执行作业A。

说明

- 依赖的作业B的“某段时间”，计算方法如下，详见后文**设置依赖作业后的作业运行原理**。
  - 同周期依赖，如分钟依赖分钟、小时依赖小时或天依赖天时，“某段时间”为 **(作业A执行时间-作业A周期时间, 作业A执行时间]** 。
  - 跨周期依赖：如小时依赖分钟、天依赖分钟、天依赖小时或月依赖天时，“某段时间”为 **[上一作业A调度周期的自然起点, 当前作业A调度周期的自然起点)** 。
- 作业A是否判断其依赖的作业B的实例状态，与“依赖的作业失败后，当前作业处理策略”参数有关，具体如下：
  - “依赖的作业失败后，当前作业处理策略”参数配置为“挂起”或“终止执行”后，当其依赖的作业B在某段时间内存在运行失败实例，则作业A“挂起”或“终止执行”。
  - “依赖的作业失败后，当前作业处理策略”参数配置为“继续执行”，只要其依赖的作业B在某段时间内所有实例跑完（不判断其状态），则作业A就继续执行。

图 3-230 作业依赖属性

本章节主要介绍**设置依赖作业的条件**，以及**设置依赖作业后的作业运行原理**。

设置依赖作业的条件

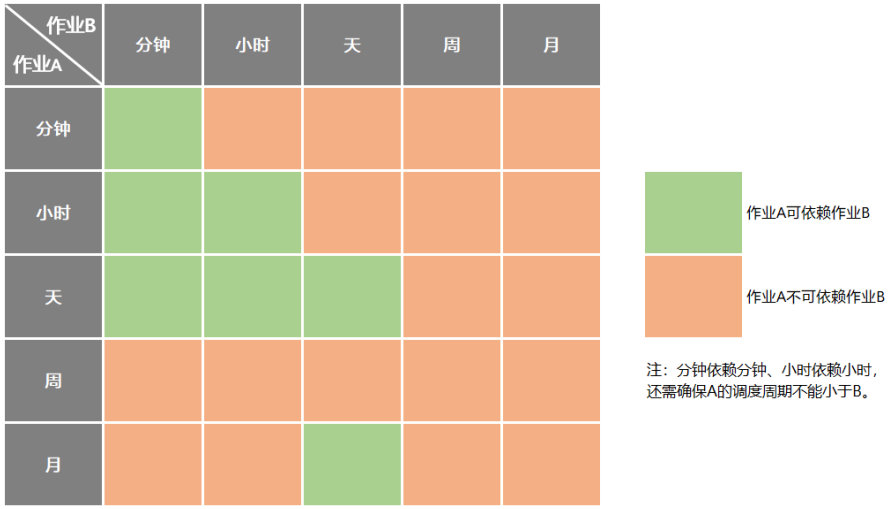
当前周期调度作业的调度周期包括分钟、小时、天、周、月这五种周期，周期调度作业A如果要配置依赖作业为周期调度作业B，则调度周期必须符合以下要求：

- 作业A的调度周期不能比依赖作业B小。例如，作业A和作业B同为分钟/小时调度，A的间隔时间小于B的间隔时间，则作业A不能设置作业B为依赖作业；作业A为分钟调度，作业B为小时调度，则作业A不能设置作业B为依赖作业。
- 作业A和依赖作业B的不能有任一调度周期为周。例如，作业A的调度周期为周或作业B的调度周期为周，则作业A不能设置作业B为依赖作业。

- 调度周期为月的作业只能依赖调度周期为天的作业。例如，作业A的调度周期为月，则作业A只能设置调度周期为天的作业为依赖作业。

不同调度周期的作业，其允许配置的依赖作业调度周期总结如图3-231所示。

图 3-231 作业依赖关系全景图



设置依赖作业后的作业运行原理

同周期依赖和跨周期依赖的作业运行原理有所差异。为方便说明，本例中假设“依赖的作业失败后，当前作业处理策略”参数设置为“继续执行”，作业A不判断作业B的实例运行状态；如果该参数设置为“挂起”或“终止执行”，则作业A还会额外判断作业B的实例中是否存在失败实例。

- **同周期依赖**：即作业A与其依赖作业B为相同调度周期，如分钟依赖分钟、小时依赖小时或天依赖天。

  同周期依赖的情况下，当作业A的依赖作业配置为作业B后，作业A会在 **(作业A执行时间-作业A周期时间, 作业A执行时间]** 时间区间内检查是否有作业B的实例运行，只有在此期间作业B的实例运行完成才会运行作业A。

  示例1：作业A依赖作业B，均为分钟调度。作业A的开始时间10:00，周期时间20分钟；作业B的开始时间10:00，周期时间10分钟。则会出现如下情况：

表 3-239 示例 1：同周期作业依赖情况

时间点	作业B（分钟调度，开始时间10:00，周期时间10分钟）	作业A（分钟调度，开始时间10:00，周期时间20分钟）
10:00	执行	检查 **(09:40, 10:00]** 区间，有作业B实例运行，待作业B执行完成后，执行作业A
10:10	执行	-
10:20	执行	检查 **(10:00, 10:20]** 区间，有作业B实例运行，待作业B执行完成后，执行作业A
10:30	执行	-

时间点	作业B（分钟调度，开始时间10:00，周期时间10分钟）	作业A（分钟调度，开始时间10:00，周期时间20分钟）
...	...	...

示例2：作业A依赖作业B，均为天调度。作业A的开始时间为8月1日09:00；作业B的开始时间8月1日10:00。则会出现如下情况：

表 3-240 示例 2：同周期作业依赖情况

时间点	作业B（天调度，开始时间为8月1日10:00）	作业A（天调度，开始时间8月1日09:00）
8月1日 09:00	-	检查 **(7月31日09:00, 8月1日09:00]** 区间，无作业B实例运行，不执行作业A
8月1日 10:00	执行	-
8月2日 09:00	-	检查 **(8月1日09:00, 8月2日09:00]** 区间，有作业B实例运行，待作业B执行完成后，执行作业A
8月2日 10:00	执行	-
...	...	...

- **跨周期依赖**：即作业A与其依赖作业B为不同调度周期，如小时依赖分钟、天依赖分钟、天依赖小时或月依赖天。

  跨周期依赖的情况下，当作业A的依赖作业配置为作业B后，作业A会在 **[上一作业A调度周期的自然起点, 当前作业A调度周期的自然起点)** 时间区间内检查是否有作业B的实例运行，只有在此期间作业B的实例运行完成才会运行作业A

  说明

  调度周期的自然起点定义如下：

  - 调度周期为小时：**上一调度周期的自然起点**为上一小时的零分零秒，**当前调度周期的自然起点**为当前小时的零分零秒。
  - 调度周期为天：**上一调度周期的自然起点**为昨天的零点零分零秒，**当前调度周期的自然起点**为今天的零点零分零秒。
  - 调度周期为月：**上一调度周期的自然起点**为上个月1号的零点零分零秒，**当前调度周期的自然起点**为当月1号的零点零分零秒。

  示例3：作业A依赖作业B，作业A为天调度，作业B为小时调度。作业A的每天02:00执行；作业B的开始时间00:00，间隔时间10小时。则会出现如下情况：

表 3-241 示例 3：跨周期作业依赖情况

时间点	作业B（小时调度，开始时间00:00，间隔时间10小时）	作业A（天调度，每天02:00执行）
第1天 00:00	执行	-
第1天 02:00	-	检查 **[第0天00:00:00, 第1天00:00:00)** 区间，无作业B实例运行，不执行
第1天 10:00	执行	-
第1天 20:00	执行	-
第2天 00:00	执行	-
第2天 02:00	-	检查 **[第1天00:00:00, 第2天00:00:00)** 区间，有作业B实例运行完成，执行作业A
第2天 10:00	执行	-
第2天 20:00	执行	-
...	...	...

示例4：作业A依赖作业B，作业A为月调度，作业B为天调度。作业A的每月1号、2号的02:00执行；作业B在8月1日00:00开始执行。则会出现如下情况：

表 3-242 示例 4：跨周期作业依赖情况

时间点	作业B（天调度，8月1日00:00执行）	作业A（月调度，每月1号、2号的02:00执行）
8月1日 00:00	执行	-
8月1日 02:00	-	检查 **[7月1日00:00:00, 8月1日00:00:00)** 区间，无作业B实例运行，不执行
8月2日 00:00	执行	-
8月2日 02:00	-	检查 **[7月1日00:00:00, 8月1日00:00:00)** 区间，无作业B实例运行，不执行

时间点	作业B（天调度，8月1日00:00执行）	作业A（月调度，每月1号、2号的02:00执行）
...	-	...
9月1日00:00	执行	-
9月1日02:00	-	检查 **[8月1日00:00:00, 9月1日00:00:00)** 区间，有作业B实例运行完成，执行作业A
9月2日00:00	执行	-
9月2日02:00	-	检查 **[8月1日00:00:00, 9月1日00:00:00)** 区间，有作业B实例运行完成，执行作业A
...	...	...

3.4.11.2 IF 条件判断教程

当您在数据开发模块进行作业开发编排时，想要实现通过设置条件，选择不同的执行路径，可使用IF条件判断。

本教程包含以下三个常见场景举例。

- **根据前一个节点的执行状态进行IF条件判断**
- **根据前一个节点的输出结果进行IF条件判断**
- **多IF条件下当前节点的执行策略**

IF条件的数据来源于EL表达式，通过EL表达式，根据具体的场景选择不同的EL表达式来达到目的。您可以参考本教程，根据您的实际业务需要，开发您自己的作业。

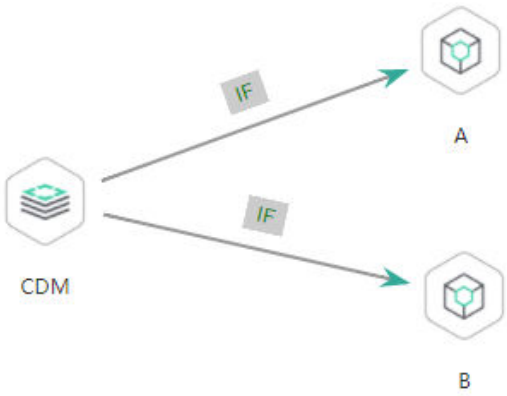
EL表达式用法可参考**EL表达式概述**。

根据前一个节点的执行状态进行 IF 条件判断

场景说明

根据前一个CDM节点是否执行成功，决定执行哪一个IF条件分支。基于**图3-232**的样例，说明如何设置IF条件。

图 3-232 作业样例



配置方法

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在“作业开发”页面，新建数据开发作业，然后分别选择CDM节点和两个Dummy节点，选中连线图标并拖动，编排图3-232所示的作业。其中CDM节点的失败策略需要设置为“继续执行下一节点”。

步骤4 右键单击连线，选择“设置条件”，在弹出的“编辑EL表达式”文本框中输入IF条件。

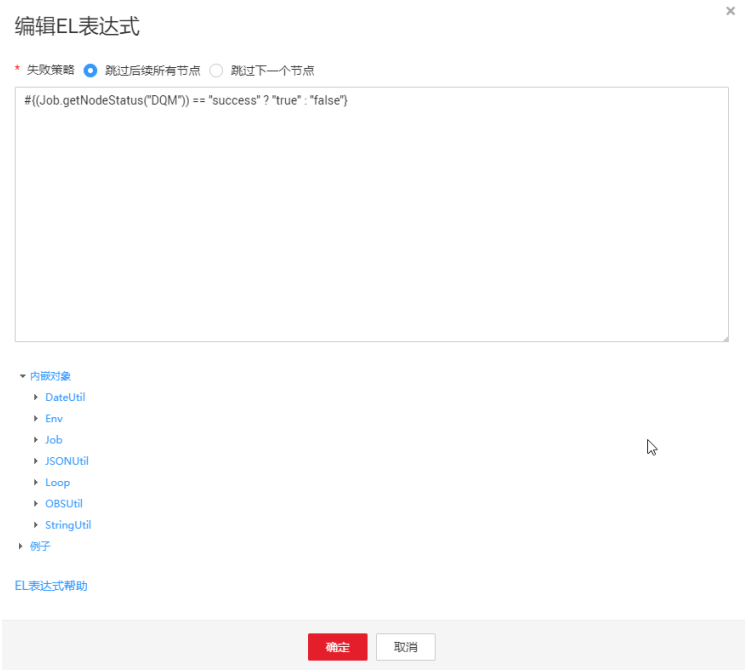
每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

此Demo中使用的EL表达式为“#{Job.getNodeStatus("node_name")}”，这个表达式的作用为获取指定节点的执行状态，成功状态返回success，失败状态返回fail。本例使用中，IF条件表达式分别为：

- 上面的A分支IF条件表达式为： #{(Job.getNodeStatus("CDM")) == "success" ? "true" : "false"}
- 下面的B分支IF条件表达式为：#{(Job.getNodeStatus("CDM")) == "fail" ? "true" : "false"}

输入IF条件表达式后，配置IF条件匹配失败策略，可选择仅跳过相邻的下一个节点，或者跳过该IF分支后续所有节点。配置完成后点击确定，保存作业。

图 3-233 配置失败策略



步骤5 测试运行作业，并前往实例监控中查看执行结果。

步骤6 待作业运行完成后，从实例监控中查看作业实例的运行结果，如图3-234所示。可以看到运行结果是符合预期的，当前CDM执行的结果为fail的时候，跳过A分支，执行B分支。

图 3-234 作业运行结果

----结束

根据前一个节点的输出结果进行 IF 条件判断

场景说明

目标场景：将HIVE SQL节点的Select语句执行结果，作为参数传递到下一个节点进行条件判断，然后决定执行哪一个IF条件分支。

场景分析：由于HIVE SQL节点的Select语句执行结果为二维数组，要获取二维数组中的值，我们需要用到**#{Loop.dataArray[][]}**这个EL表达式，而当前只有For Each节点支持该表达式，所以HIVE SQL节点后面需要连接一个For Each节点，作业编排如图3-235所示：

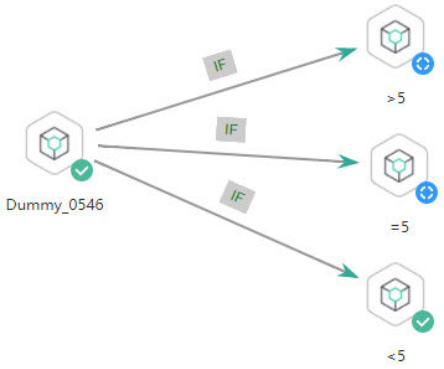
图 3-235 主作业样例

其中，For Each节点的关键配置如下：

- 数据集：数据集就是HIVE SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput('HIVE')}**，其中HIVE为前一个节点的名称。
- 作业运行参数：作业运行参数是子作业中定义的参数，可以将主作业前一个节点的输出，传递到子作业以供使用。此处变量名为**result**，其值为数据集中的某一列，使用EL表达式**#{Loop.dataArray[0][0]}**。

而For Each节点中所选的子作业，需要根据For Each节点传过来的作业运行参数，决定执行For Each中子作业的哪一个IF条件分支，作业编排如图3-236所示。

图 3-236 子作业样例



其中，子作业的关键配置为IF条件设置，本例使用表达式${result}获取作业参数的值。

说明

此处不能使用EL表达式#{Job.getParam("job_param_name")}，因为此表达式只能直接获取当前作业里配置的参数的value，并不能获取到父作业传递过来的参数值，也不能获取到工作空间里面配置的全局变量，作用域仅为本作业。

而表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。

配置方法

开发子作业

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在“作业开发”页面，新建数据开发子作业foreach。选择四个Dummy节点，选中连线图标并拖动，编排图3-236所示的作业。

步骤4 右键单击节点间的连线，选择“设置条件”，在弹出的“编辑EL表达式”文本框中输入IF条件。

每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

- 上面的>5分支，IF条件表达式为：#{${result} > 5 ? "true" : "false"}
- 中间的=5分支，IF条件表达式为：#{${result} == 5 ? "true" : "false"}
- 下面的<5分支，IF条件表达式为：#{${result} < 5 ? "true" : "false"}

输入IF条件表达式后，配置IF条件匹配失败策略，可选择仅跳过相邻的下一个节点，或者跳过该IF分支后续所有节点。

步骤5 配置作业参数。此处需将参数名填写为**result**，仅用于主作业testif中的For Each节点识别子作业参数；参数值无需填写。

图 3-237 配置作业参数

步骤6 配置完成后保存作业。

----结束

开发主作业

步骤1 在“作业开发”页面，新建数据开发主作业testif。选择HIVE SQL节点和For Each节点，选中连线图标并拖动，编排图3-235所示的作业。

步骤2 配置HIVE SQL节点属性。此处配置为引用SQL脚本，SQL脚本的语句如下所示。其他节点属性参数无特殊要求。

```
SELECT count(*) FROM student //从student表中计数，脚本执行结果为二维数组
```

图 3-238 HIVE SQL 脚本执行结果

步骤3 配置For Each节点属性，如图3-239所示。

- 子作业：子作业选择已经开发完成的子作业“foreach”。
- 数据集：数据集就是HIVE SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput('HIVE')}**，其中HIVE为前一个节点的名称。
- 作业运行参数：作业运行参数是子作业中定义的参数，可以将主作业前一个节点的输出，传递到子作业以供使用。此处变量名为子作业参数名**result**，其值为数据集中的某一列，使用EL表达式**#{Loop.dataArray[0][0]}**。

图 3-239 For Each 节点属性

步骤4 配置完成后保存作业。

----结束

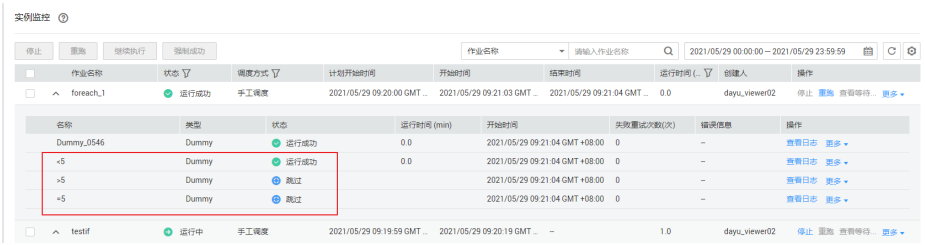
测试运行主作业

步骤1 点击主作业画布上方的“测试运行”按钮，测试作业运行情况。主作业运行后，会通过For Each节点自动调用运行子作业。

步骤2 点击左侧导航栏中的“实例监控”，进入实例监控中查看作业运行结果。

步骤3 待作业运行完成后，从实例监控中查看子作业foreach的运行结果，如图3-240所示。可以看到运行结果是符合预期的，当前HIVE SQL执行的结果是1，所以>5和=5的分支被跳过，执行<5这个分支成功。

图 3-240 子作业运行结果



----结束

多 IF 条件下当前节点的执行策略

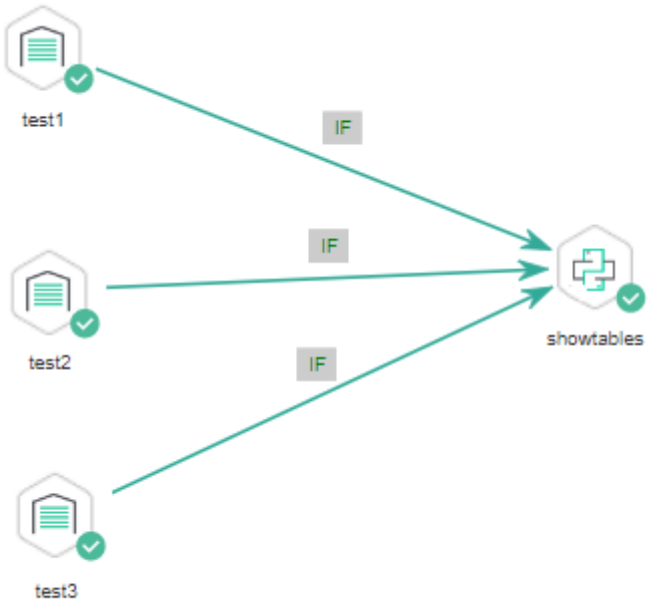
如果当前节点的执行依赖多个IF条件的节点，执行的策略包含逻辑或和逻辑与两种。

当执行策略配置为逻辑或，则表示多个IF判断条件只要任意一个满足条件，则执行当前节点。

当执行策略配置为逻辑与，则表示多个IF判断条件需要所有条件满足时，才执行当前节点。

如果没有配置执行策略，系统默认为逻辑或处理。

图 3-241 多 IF 条件作业样例



配置方法

配置执行策略

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在数据开发模块，单击“配置管理 > 配置”，单击“默认项配置”。

步骤4 “多IF策略”可设置为“逻辑与”或者“逻辑或”。

步骤5 单击“保存”。

----结束

开发作业

步骤1 在“作业开发”页面，新建一个数据开发作业。

步骤2 拖动三个DWS SQL算子作为父节点，一个Python算子作为子节点，选中连线图标并拖动，编排图3-241所示的作业。

步骤3 右键单击节点间的连线，选择“设置条件”，在弹出的“编辑EL表达式”文本框中输入IF条件。

每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。

- test1节点IF条件表达式为：#{(Job.getNodeStatus("test1")) == "success" ? "true" : "false"},
- test2节点IF条件表达式为：#{(Job.getNodeStatus("test2")) == "success" ? "true" : "false"},

- test3节点IF条件表达式为：#{(Job.getNodeStatus("test3")) == "success" ? "true" : "false"},

此处表达式均采用前一个节点的执行状态进行IF条件判断。

输入IF条件表达式后，配置IF条件匹配失败策略，可选择仅跳过相邻的下一个节点，或者跳过该IF分支后续所有节点。

----结束

测试运行作业

步骤1 单击作业画布上方的“保存”按钮，保存完成编排的作业。

步骤2 单击作业画布上方的“测试运行”按钮，测试作业运行情况。

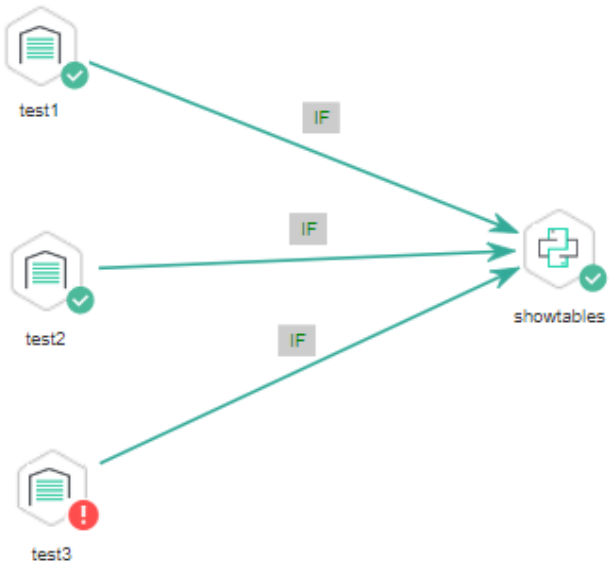
test1运行成功，则对应的IF条件为true；

test2运行成功，则对应的IF条件为true；

test3运行失败，则对应的IF条件为false。

当多IF策略配置为“逻辑或”时，showtables节点运行完成，作业运行完成。详细情况如下所示。

图 3-242 配置为“逻辑或”的作业运行情况

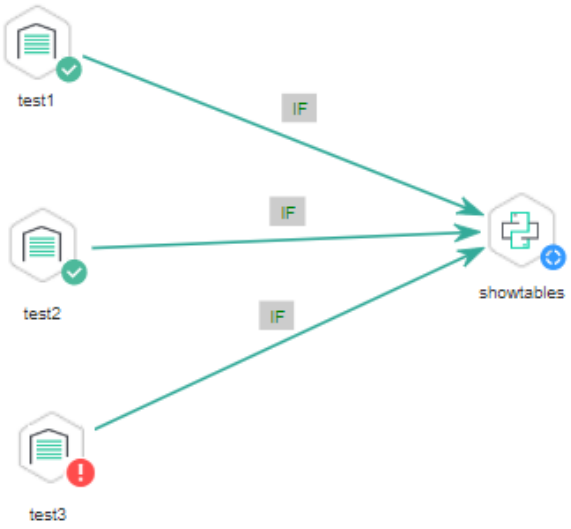


测试运行日志

[INFO][2022/03/31 15:53:35 GMT+08:00] : 作业开始运行...
[INFO][2022/03/31 15:54:12 GMT+08:00] : 节点"test1"开始运行...
[INFO][2022/03/31 15:54:12 GMT+08:00] : 节点"test2"开始运行...
[INFO][2022/03/31 15:54:12 GMT+08:00] : 节点"test3"开始运行...
[INFO][2022/03/31 15:54:22 GMT+08:00] : 节点"test1"运行完成。
[INFO][2022/03/31 15:54:22 GMT+08:00] : 节点"test2"运行完成。
[ERROR][2022/03/31 15:54:53 GMT+08:00] : 节点"test3"运行失败。
[INFO][2022/03/31 15:55:03 GMT+08:00] : 节点"showtables"开始运行...
[INFO][2022/03/31 15:55:13 GMT+08:00] : 节点"showtables"运行完成。
[INFO][2022/03/31 15:55:13 GMT+08:00] : 作业运行完成

当多IF策略配置为“逻辑与”时，showtables节点跳过，作业运行完成。详细情况如下所示。

图 3-243 配置为“逻辑与”的作业运行情况



测试运行日志

[INFO][2022/03/31 15:51:38 GMT+08:00] : 作业开始运行...
[INFO][2022/03/31 15:52:16 GMT+08:00] : 节点"test2"运行完成。
[INFO][2022/03/31 15:52:16 GMT+08:00] : 节点"test1"运行完成。
[INFO][2022/03/31 15:52:16 GMT+08:00] : 节点"test3"开始运行...
[ERROR][2022/03/31 15:52:56 GMT+08:00] : 节点"test3"运行失败。
[INFO][2022/03/31 15:53:06 GMT+08:00] : 节点"showtables"已跳过
[INFO][2022/03/31 15:53:17 GMT+08:00] : 作业运行完成

----结束

3.4.11.3 获取 Rest Client 算子返回值教程

Rest Client算子可以执行RESTful请求。

本教程主要介绍如何获取Rest Client的返回值，包含以下两个使用场景举例。

- **通过“响应消息体解析为传递参数定义”获取返回值**
- **通过EL表达式获取返回值**

通过“响应消息体解析为传递参数定义”获取返回值

如图3-244所示，第一个Rest Client调用了MRS服务查询集群列表的API，图3-245为API返回值的JSON消息体。

- 使用场景：需要获取集群列表中第一个集群的cluster Id，然后作为参数传递给后面的节点使用。

- 关键配置：在第一个Rest Client的“响应消息体解析为传递参数定义”配置中，配置clusterId=clusters[0].clusterId，后续的Rest Client节点就可以用${clusterId}的方式引用到集群列表中的第一个集群的cluster Id。

图 3-244 Rest Clien 作业样例 1

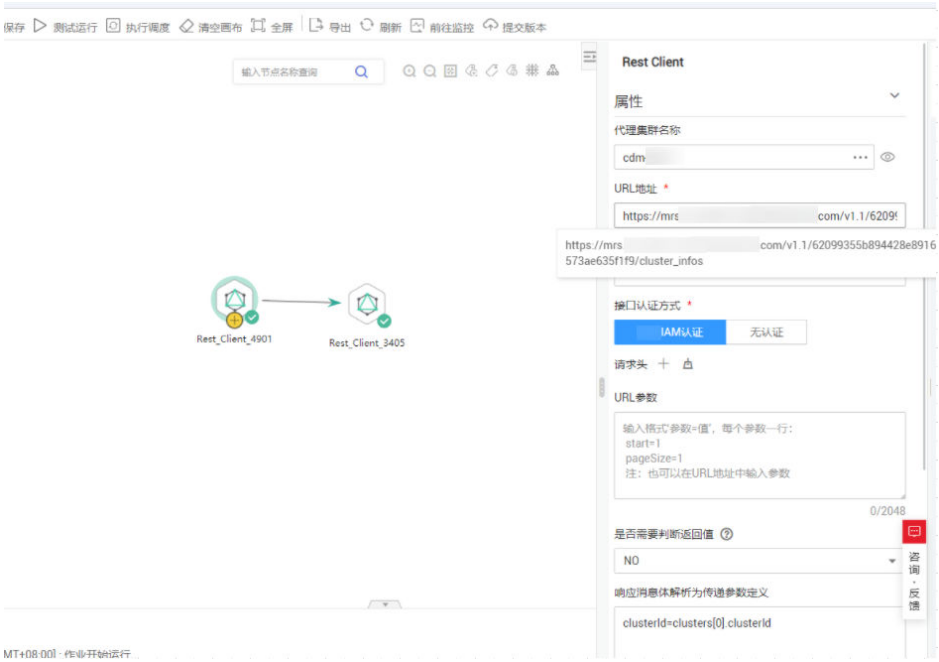


图 3-245 JSON 消息体

通过 EL 表达式获取返回值

Rest Client算子可与EL表达式相配合，根据具体的场景选择不同的EL表达式来实现更丰富的用法。您可以参考本教程，根据您的实际业务需要，开发您自己的作业。EL表达式用法可参考**EL表达式概述**。

如图3-246所示，Rest Client调用了MRS服务查询集群列表的API，然后执行Kafka Client发送消息。

- 使用场景：Kafka Client发送字符串消息，消息内容为集群列表中第一个集群的cluster Id。
- 关键配置：在Kafka Client中使用如下EL表达式获取Rest API返回消息体中的特定字段：
  #{JSONUtil.toString(JSONUtil.path(Job.getNodeOutput("Rest_Client_4901"),"clusters[0].clusterId"))}

图 3-246 Rest Client 作业样例 2

3.4.11.4 For Each 算子使用介绍

适用场景

当您进行作业开发时，如果某些任务的参数有差异、但处理逻辑全部一致，在这种情况下您可以通过For Each算子避免重复开发作业。

For Each算子可指定一个子作业循环执行，并通过数据集对子作业中的参数进行循环替换。关键参数如下：

- 子作业：选择需要循环执行的作业。
- 数据集：即不同子任务的参数值的集合。可以是给定的数据集，如“[['1'],['3'],['2']]”；也可以是EL表达式如“#{Job.getNodeOutput('preNodeName')}”，即前一个节点的输出值。
- 作业运行参数：参数名即子作业中定义的变量；参数值一般配置为数据集中的某组数据，每次运行中会将参数值传递到子作业以供使用。例如参数值填写为：#{Loop.current[0]}，即将数据集中每组数据的第一个数值遍历传递给子作业。

For Each算子举例如图3-247所示 。从图中可以看出，子作业“foreach”中的参数名为“result”，参数值为一维数组数据集“[['1'],['3'],['2']]”的遍历（即第一次循环为1，第二次循环为3，第三次循环为2）。

图 3-247 for each 算子

For Each 算子与 EL 表达式

要想使用好For Each算子，您必须对EL表达式有所了解。EL表达式用法请参考**EL表达式概述**。

下面为您展示For Each算子常用的一些EL表达式。

- #{Loop.dataArray}：For循环算子输入的数据集，是一个二维数组。
- #{Loop.current}：由于For循环算子在处理数据集的时候，是一行一行进行处理的，那Loop.current就表示当前处理到的某行数据，Loop.current是一个一维数组，一般定义格式为#{Loop.current[0]}、#{Loop.current[1]}或其它，0表示遍历到当前行的第一个值。
- #{Loop.offset}：For循环算子在处理数据集时当前的偏移量，从0开始。
- #{Job.getNodeOutput('preNodeName')}：获取前面节点的输出。

使用案例

案例场景

因数据规整要求，需要周期性地将多组DLI源数据表数据导入到对应的DLI目的表，如**表1**所示。

表 3-243 需要导入的列表情况

源数据表名	目的表名
a_new	a
b_2	b
c_3	c
d_1	d
c_5	e
b_1	f

如果通过SQL节点分别执行导入脚本，需要开发大量脚本和节点，导致重复性工作。在这种情况下，我们可以使用For Each算子进行循环作业，节省开发工作量。

配置方法

步骤1 准备源表和目的表。为了便于后续作业运行验证，需要先创建DLI源数据表和目的表，并给源数据表插入数据。

1. 创建DLI表。您可以在DataArts Studio数据开发中，新建DLI SQL脚本执行以下SQL命令，也可以在数据湖探索（DLI）服务控制台的SQL编辑器中执行以下SQL命令：

```
/* 创建数据表 */
CREATE TABLE a_new (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE b_2 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE c_3 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE d_1 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE c_5 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE b_1 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE a (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE b (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE c (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE d (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE e (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE f (name STRING, score INT) STORED AS PARQUET;
```

2. 给源数据表插入数据。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令，也可以在数据湖探索（DLI）服务控制台的SQL编辑器中执行以下SQL命令：

```
/* 源数据表插入数据 */
INSERT INTO a_new VALUES ('ZHAO','90'),('QIAN','88'),('SUN','93');
INSERT INTO b_2 VALUES ('LI','94'),('ZHOU','85');
INSERT INTO c_3 VALUES ('WU','79');
INSERT INTO d_1 VALUES ('ZHENG','87'),('WANG','97');
INSERT INTO c_5 VALUES ('FENG','83');
INSERT INTO b_1 VALUES ('CEHN','99');
```

步骤2 准备数据集数据。您可以通过以下方式之一获取数据集：

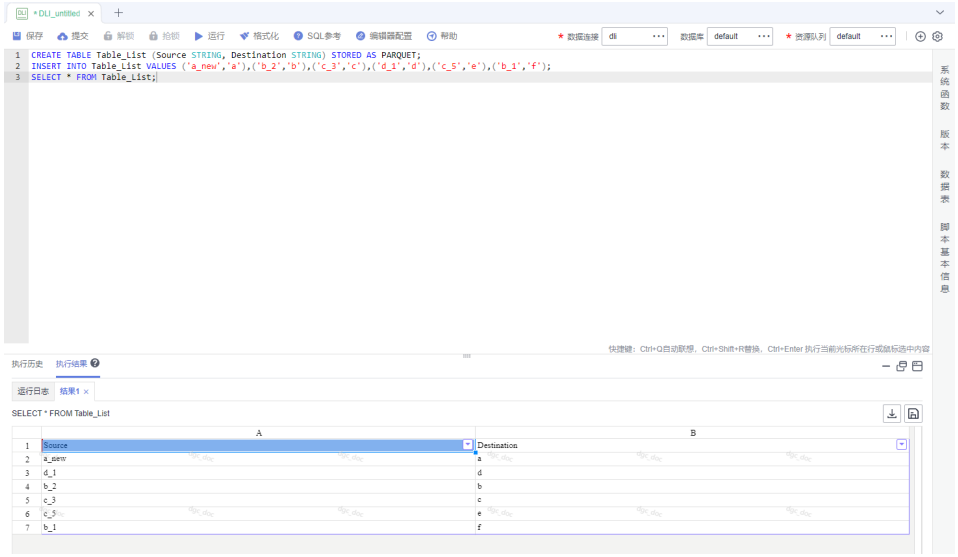
1. 您可以将**表1**数据导入到DLI表中，然后将SQL脚本读取的结果作为数据集。
2. 您可以将**表1**数据保存在OBS的CSV文件中，然后通过DLI SQL或DWS SQL创建OBS外表关联这个CSV文件，然后将OBS外表查询的结果作为数据集。
3. 您可以将**表1**数据保存在HDFS的CSV文件中，然后通过HIVE SQL创建Hive外表关联这个CSV文件，然后将HIVE外表查询的结果作为数据集。

本例以方式1进行说明，将**表1**中的数据导入到DLI表（Table_List）中。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令导入数据，也可以在数据湖探索（DLI）服务控制台的SQL编辑器中执行以下SQL命令：

```
/* 创建数据表TABLE_LIST，然后插入表1数据，最后查看生成的表数据 */
CREATE TABLE Table_List (Source STRING, Destination STRING) STORED AS PARQUET;
INSERT INTO Table_List VALUES ('a_new','a'),('b_2','b'),('c_3','c'),('d_1','d'),('c_5','e'),('b_1','f');
SELECT * FROM Table_List;
```

生成的Table_List表数据如下：

图 3-248 Table_List 表数据



步骤3 创建要循环运行的子作业ForeachDemo。在本次操作中，定义循环执行的是一个包含了DLI SQL节点的任务。

1. 进入DataArts Studio数据开发模块选择“作业开发”页面，新建作业ForeachDemo，然后选择DLI SQL节点，编排图3-249所示的作业。

   DLI SQL的语句中把要替换的变量配成${}这种参数的形式。在下面的SQL语句中，所做的操作是把${Source}表中的数据全部导入${Destination}中，${fromTable}、${toTable} 就是要替换的变量参数。SQL语句为：

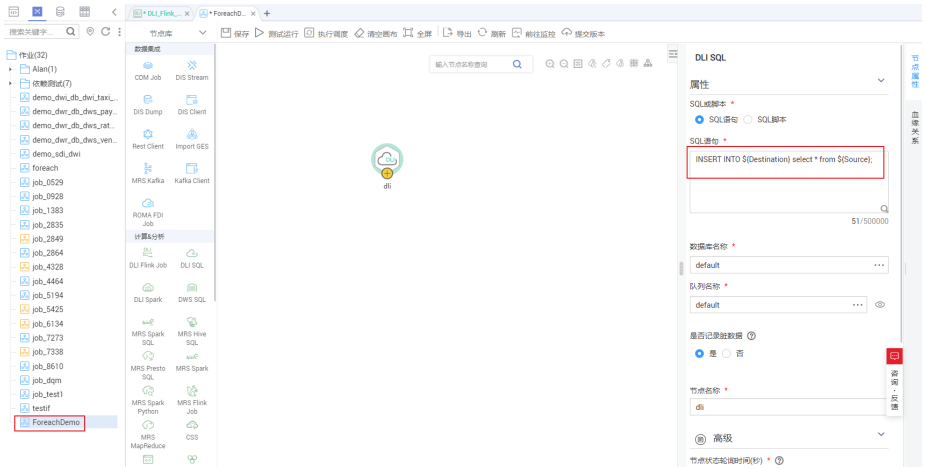
   ```
   INSERT INTO ${Destination} select * from ${Source};
   ```

   说明

   此处不能使用EL表达式#{Job.getParam("job_param_name")}，因为此表达式只能直接获取当前作业里配置的参数的value，并不能获取到父作业传递过来的参数值，也不能获取到工作空间里面配置的全局变量，作用域仅为本作业。

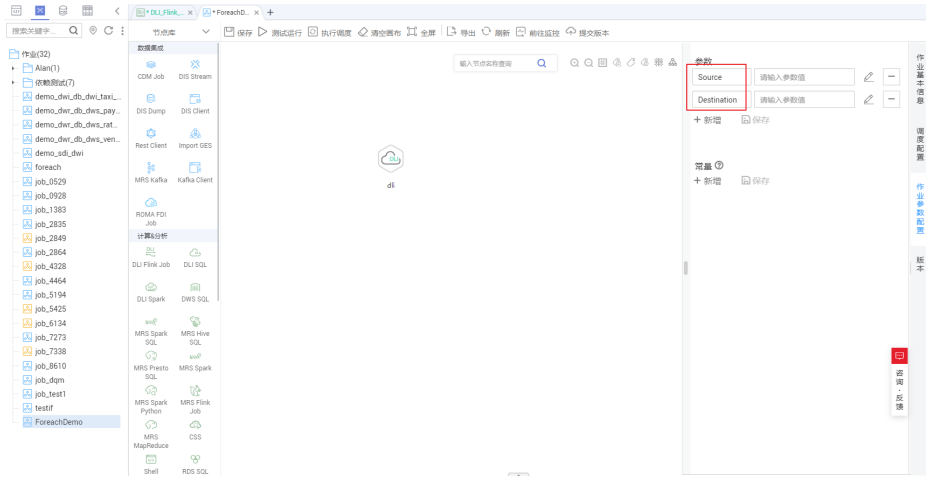
   而表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。

图 3-249 循环执行子作业



2. 配置完成SQL语句后，在子作业中配置作业参数。此处仅需要配置参数名，用于主作业ForeachDemo_master中的For Each节点识别子作业参数；参数值无需填写。

图 3-250 配置子作业参数

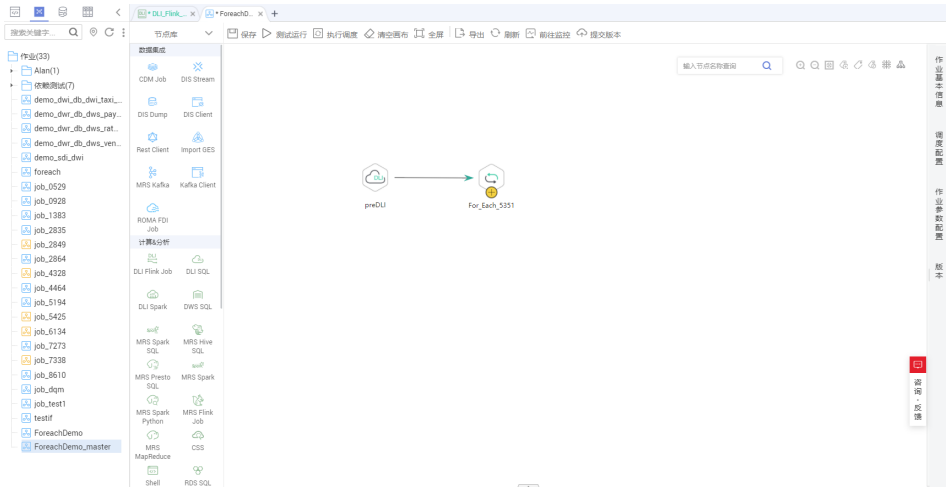


3. 配置完成后保存作业。

步骤4 创建For Each算子所在的主作业ForeachDemo_master。

1. 进入DataArts Studio数据开发模块选择“作业开发”页面，新建数据开发主作业ForeachDemo_master。选择DLI SQL节点和For Each节点，选中连线图标并拖动，编排图3-251所示的作业。

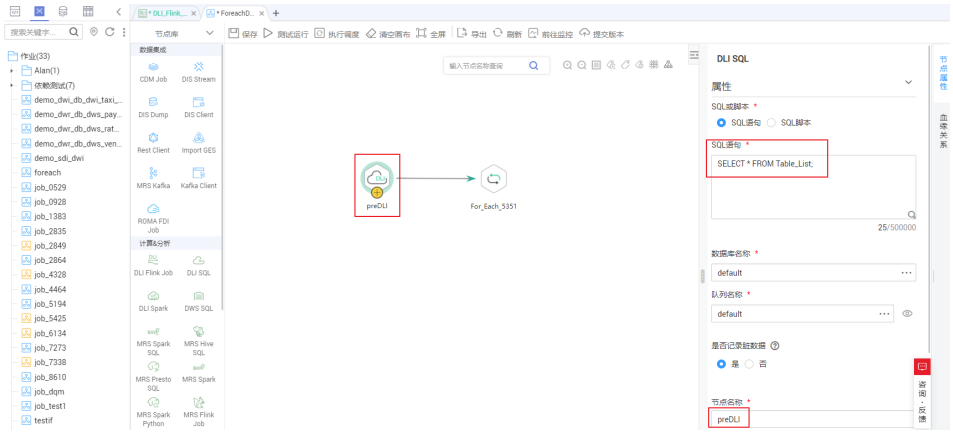
图 3-251 编排作业



2. 配置DLI SQL节点属性，此处配置为SQL语句，语句内容如下所示。DLI SQL节点负责读取DLI表Table_List中的内容作为数据集。

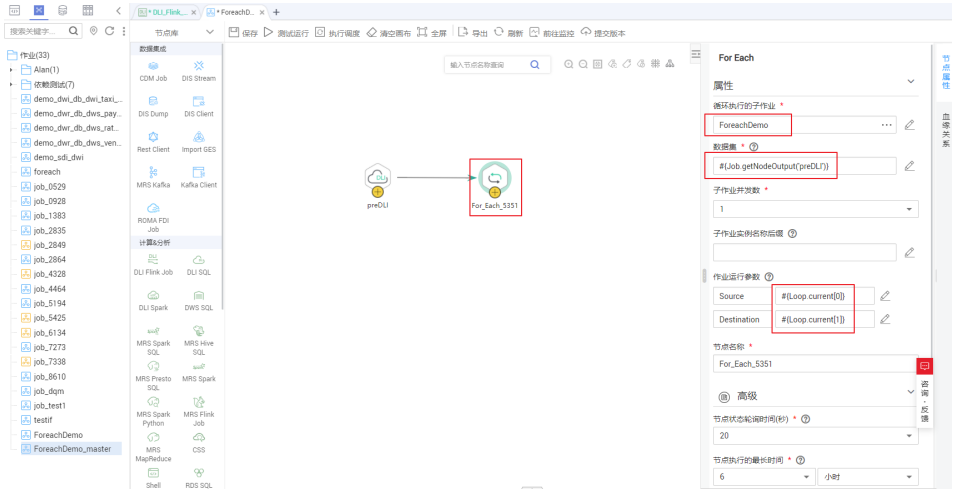
```
SELECT * FROM Table_List;
```

图 3-252 DLI SQL 节点配置



3. 配置For Each节点属性。
   - 子作业：子作业选择**步骤2**已经开发完成的子作业“ForeachDemo”。
   - 数据集：数据集就是DLI SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput('preDLI')}**，其中preDLI为前一个节点的名称。
   - 作业运行参数：用于将数据集中的数据传递到子作业以供使用。Source对应的是数据集Table_List表的第一列，Destination是第二列，所以配置的EL表达式分别为**#{Loop.current[0]}**、**#{Loop.current[1]}**。

图 3-253 配置 For Each 算子

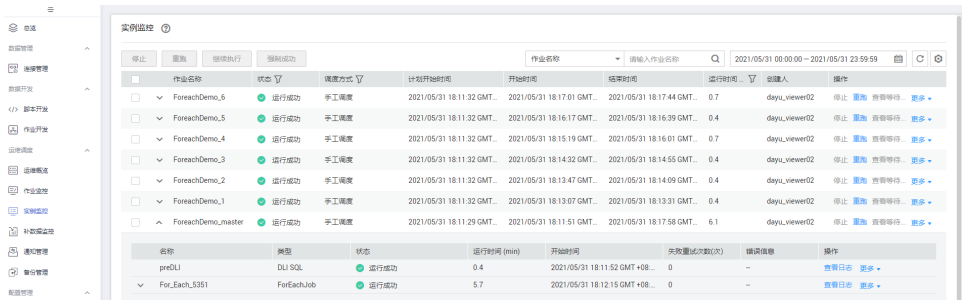


4. 配置完成后保存作业。

步骤5 测试运行主作业。

1. 点击主作业画布上方的“测试运行”按钮，测试作业运行情况。主作业运行后，会通过For Each节点自动调用运行子作业。
2. 点击左侧导航栏中的“实例监控”，进入实例监控中查看作业运行情况。等待作业运行成功后，就能查看For Each节点生成的子作业实例，由于数据集中有6行数据，所以这里就对应产生了6个子作业实例。

图 3-254 查看作业实例

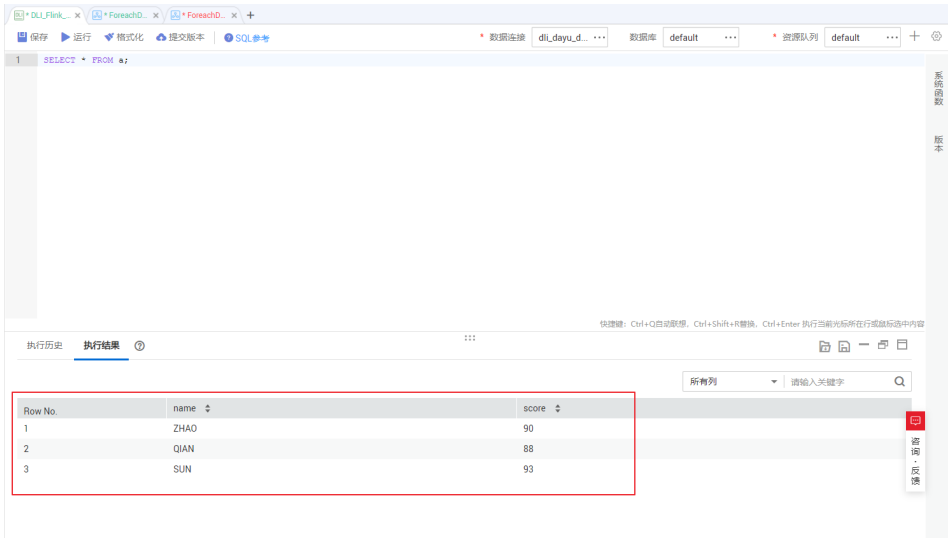


3. 查看对应的6个DLI目的表中是否已被插入预期的数据。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令导入数据，也可以在数据湖探索（DLI）服务控制台的SQL编辑器中执行以下SQL命令：

```
/* 查看表a数据，其他表数据请修改命令后运行 */
SELECT * FROM a;
```

将查询到的表数据与**给源数据表插入数据**步骤中的数据进行对比，可以发现数据插入符合预期。

图 3-255 目的表数据



----结束

更多案例参考

For Each算子可与其他算子配合，实现更丰富的功能。您可以参考以下案例，了解For Each算子的更多用法。

- **根据前一个节点的输出结果进行IF条件判断**

3.4.11.5 开发一个 Python 脚本

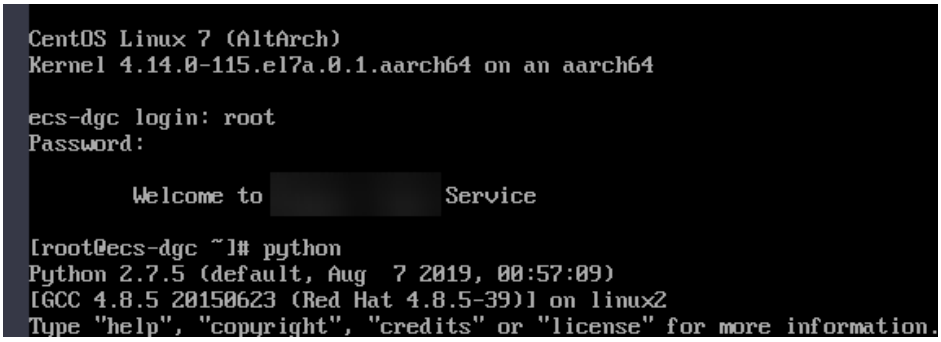
本章节介绍如何在数据开发模块上开发并执行Python脚本示例。

环境准备

- 已开通弹性云服务器，并创建ECS，ECS主机名为“ecs-dgc”。

  说明

  本示例主机选择“CentOS 8.0 64bit with ARM(40GB)”的公共镜像，并且使用ECS自带的Python环境，您可登录主机后使用**python**命令确认服务器的Python环境。

- 已开通数据集成增量包，CDM集群名为“cdm-dlfpyhthon”，提供数据开发模块与ECS主机通信的代理。
- 请确保ECS主机与CDM集群网络互通，互通需满足如下条件：
  - CDM集群与ECS主机同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见《虚拟私有云(VPC) 使用指南》中的“添加路由信息”章节，配置安全组规则请参见《虚拟私有云(VPC)使用指南》中的“安全组 > 添加安全组规则”章节。
  - CDM集群与ECS主机处于不同区域的情况下，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，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。
  - 此外，您还必须确保该ECS主机与CDM集群所属的企业项目必须相同，如果不同，需要修改工作空间的企业项目。

约束限制

- Python脚本暂不支持脚本参数及作业参数。

建立主机数据连接

开发Python脚本前，我们需要建立一个到弹性云服务器ECS的连接。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

图 3-256 选择管理中心

步骤2 在管理中心页面，单击“数据连接”，进入数据连接页面。

图 3-257 创建数据连接

步骤3 单击“创建数据连接”，进入“创建数据连接”页面中。

图 3-258 创建数据连接

步骤4 参见**表3-244**配置相关参数，创建主机连接名称为“python_test”的数据连接。

表 3-244 主机连接

参数	是否必选	说明
数据连接名称	是	主机连接的名称，只能包含字母，数字，中划线或者下划线。
主机地址	是	主机的地址。 请参见《弹性云服务器用户指南》的查看云服务器详细信息页获取。
绑定Agent	是	需要选择CDM集群，CDM集群提供Agent。
端口	是	主机的SSH端口号。
用户名	是	主机的登录用户名。
登录方式	是	选择主机的登录方式： • 密钥对 • 密码
密钥对	是	主机的登录方式为密钥对时，用户获取并上传其私钥文件至OBS，在此处选择对应的OBS路径。“登录方式”为“密钥对”时，显示该配置项。 **说明** 此处上传的私钥文件需为PEM格式，并且上传的私钥文件和主机上配置的公钥是一个密钥对。
密钥对密码	否	如果密钥对未设置密码，则不需要填写该配置项。
密码	是	主机的登录方式为密码时，填写主机的登录密码。
主机连接描述	否	主机连接的描述信息。

说明

关键参数说明：

- 主机地址：**已开通ECS主机**中开通的ECS主机的IP地址。
- 绑定Agent：**已开通批量数据迁移增量包**中开通的CDM集群。

步骤5 单击“测试”，测试数据连接的连通性。如果无法连通，数据连接将无法创建。

步骤6 测试通过后，单击“确定”，完成数据连接的创建。

----结束

开发 Python 脚本

步骤1 在“数据开发 > 脚本开发”模块中创建一个Python脚本，脚本名称为“python_test”。

步骤2 在编辑器中编辑Python语句并选择主机连接，单击“提交并解锁”。

步骤3 单击“运行”执行Python语句。

步骤4 查看脚本运行结果。

----结束

3.4.11.6 开发一个 DWS SQL 作业

介绍如何在数据开发模块上通过DWS SQL算子进行作业开发。

场景说明

本教程通过开发一个DWS作业来统计某门店的前一天销售额。

环境准备

- 已开通DWS服务，并创建DWS集群，为DWS SQL提供运行环境。
- 已开通CDM增量包，并创建CDM集群。

  CDM集群创建时，需要注意：虚拟私有云、子网、安全组与DWS集群保持一致，确保网络互通。

创建 DWS 的数据连接

开发DWS SQL前，我们需要在“管理中心 > 数据连接”模块中建立一个到DWS的连接，数据连接名称为“dws_link”。

关键参数说明：

- 集群名：环境准备中创建的DWS集群名称。
- 绑定Agent：环境准备中创建的CDM集群。

创建数据库

在DWS中创建数据库，以“gaussdb”数据库为例。详情请参考**新建数据库**进行操作。

创建数据表

在“gaussdb”数据库中创建数据表trade_log和trade_report。详情请参考如下建表脚本。

```
create schema store_sales;
set current_schema= store_sales;
drop table if exists trade_log;
CREATE TABLE trade_log
(
      sn          VARCHAR(16),
      trade_time   DATE,
      trade_count   INTEGER(8)
);
set current_schema= store_sales;
drop table if exists trade_report;
CREATE TABLE trade_report
(
      rq   DATE,
      trade_total   INTEGER(8)
);
```

开发 DWS SQL 脚本

在“数据开发 > 脚本开发”模块中创建一个DWS SQL脚本，脚本名称为“dws_sql”。在编辑器中输入SQL语句，通过SQL语句来实现统计前一天的销售额。

图 3-259 开发脚本

关键说明：

- 图3-259中的脚本开发区为临时调试区，关闭脚本页签后，开发区的内容将丢失。您可以通过“提交”来保存并提交脚本版本。
- 数据连接：创建DWS的数据连接中已创建的连接。

开发 DWS SQL 作业

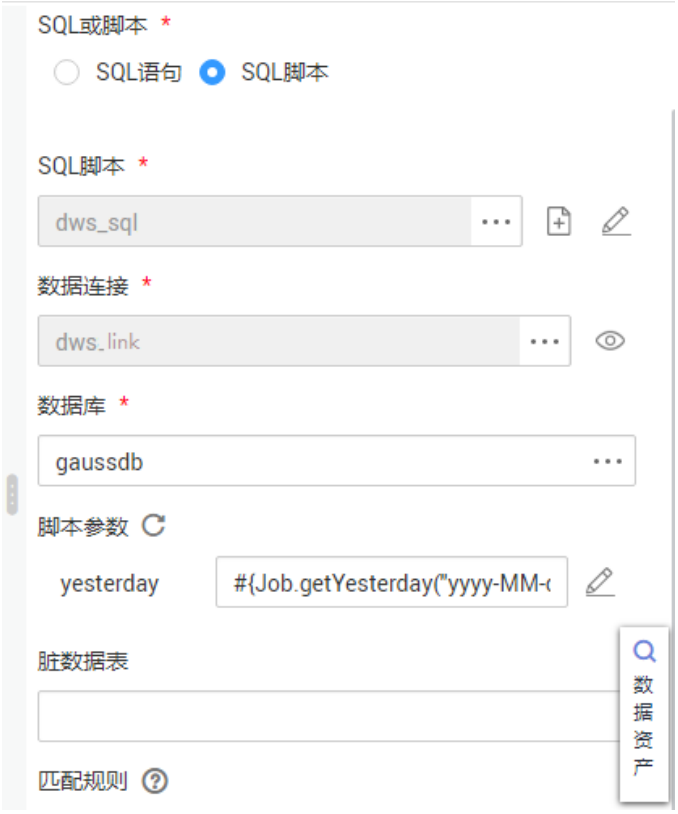
DWS SQL脚本开发完成后，我们为DWS SQL脚本构建一个周期执行的作业，使得该脚本能定期执行。

步骤1 创建一个数据开发模块空作业，作业名称为“job_dws_sql”。

图 3-260 创建 job_dws_sql 作业

步骤2 然后进入到作业开发页面，拖动DWS SQL节点到画布中并单击，配置节点的属性。

图 3-261 配置 DWS SQL 节点属性



关键属性说明：

- SQL脚本：关联**开发DWS SQL脚本**中开发完成的DWS SQL脚本“dws_sql”。
- 数据连接：默认选择SQL脚本“dws_sql”中设置的数据连接，支持修改。
- 数据库：默认选择SQL脚本“dws_sql”中设置的数据库，支持修改。
- 脚本参数：通过EL表达式获取"yesterday"的值，EL表达式如下：
  ```
  #{Job.getYesterday("yyyy-MM-dd")}
  ```
- 节点名称：默认显示为SQL脚本“dws_sql”的名称，支持修改。

步骤3 作业编排完成后，单击 ▶，测试运行作业。

步骤4 如果运行成功，单击画布空白处，在右侧的“调度配置”页面，配置作业的调度策略。

图 3-262 配置调度方式

说明：

2021/08/06至2021/08/31，每天2点执行一次作业。

步骤5 单击“提交”，执行调度作业，实现作业每天自动运行。

----结束

3.4.11.7 开发一个 Hive SQL 作业

本章节介绍如何在数据开发模块上进行Hive SQL开发。

场景说明

数据开发模块作为一站式大数据开发平台，支持多种大数据工具的开发。Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的SQL查询功能；可以将SQL语句转换为MapReduce任务进行运行。

环境准备

- 已开通MapReduce服务MRS，并创建MRS集群，为Hive SQL提供运行环境。

  MRS集群创建时，组件要包含Hive。
- 已开通数据集成CDM，并创建CDM集群，为数据开发模块提供数据开发模块与MRS通信的代理。

  CDM集群创建时，需要注意：虚拟私有云、子网、安全组与MRS集群保持一致，确保网络互通。

建立 Hive 的数据连接

开发Hive SQL前，我们需要在“管理中心 > 数据连接”模块中建立一个到MRS Hive的连接，数据连接名称为“hive1009”。

关键参数说明：

- 集群名：已创建的MRS集群。
- 绑定Agent：已创建的CDM集群。

开发 Hive SQL 脚本

在“数据开发 > 脚本开发”模块中创建一个Hive SQL脚本，脚本名称为“hive_sql”。在编辑器中输入SQL语句，通过SQL语句来实现业务需求。

图 3-263 开发脚本

关键说明：

- 图3-263中的脚本开发区为临时调试区，关闭脚本页签后，开发区的内容将丢失。您可以通过“提交”来保存并提交脚本版本。
- 数据连接：建立Hive的数据连接创建的连接。

开发 Hive SQL 作业

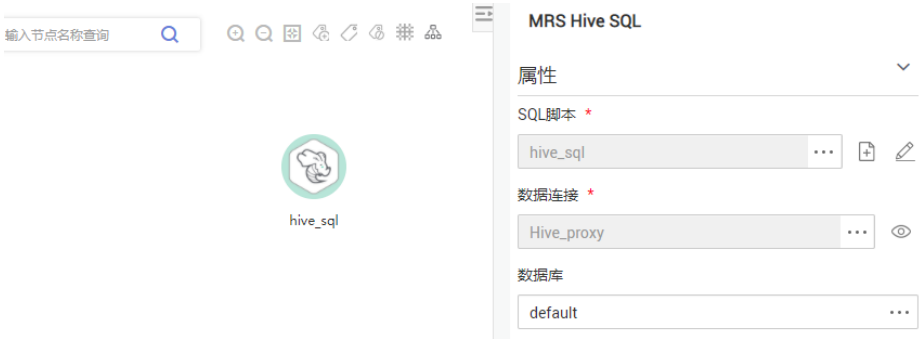
Hive SQL脚本开发完成后，我们为Hive SQL脚本构建一个周期执行的作业，使得该脚本能定期执行。

步骤1 创建一个数据开发模块空作业，作业名称为“job_hive_sql”。

图 3-264 创建 job_hive_sql 作业

步骤2 然后进入到作业开发页面，拖动MRS Hive SQL节点到画布中并单击，配置节点的属性。

图 3-265 配置 MRS Hive SQL 节点属性



关键属性说明：

- SQL脚本：关联**开发Hive SQL脚本**中开发完成的Hive SQL脚本“hive_sql”。
- 数据连接：默认选择SQL脚本“hive_sql”中设置的数据连接，支持修改。
- 数据库：默认选择SQL脚本“hive_sql”中设置的数据库，支持修改。
- 节点名称：默认显示为SQL脚本“hive_sql”的名称，支持修改。

步骤3 作业编排完成后，单击 ▶，测试运行作业。

步骤4 如果运行成功，单击画布空白处，在右侧的“调度配置”页面，配置作业的调度策略。

图 3-266 配置调度方式

说明：

2021/01/01至2021/01/25，每天2点执行一次作业。

步骤5 最后我们需要提交版本，执行调度作业，实现作业每天自动运行。

----结束

3.4.11.8 开发一个 DLI Spark 作业

在本章节您可以学习到数据开发模块资源管理、作业编辑等功能。

场景说明

用户在使用DLI服务时，大部分时间会使用SQL对数据进行分析处理，有时候处理的逻辑特别复杂，无法通过SQL处理，那么可以通过Spark作业进行分析处理。本章节通过一个例子演示如何在数据开发模块中提交一个Spark作业。

操作流程如下：

1. 创建DLI集群，通过DLI集群的物理资源来运行Spark作业。
2. 获取Spark作业的演示JAR包，并在数据开发模块中关联到此JAR包。
3. 创建数据开发模块作业，通过DLI Spark节点提交Spark作业。

环境准备

- 已开通对象存储服务OBS，并创建桶，例如“obs://dlfexample”，用于存放Spark作业的JAR包。
- 已开通数据湖探索服务DLI，并创建Spark集群“spark_cluster”，为Spark作业提供运行所需的物理资源。

获取 Spark 作业代码

本示例使用的Spark作业代码来自maven库（下载地址：**https://repo.maven.apache.org/maven2/org/apache/spark/spark-examples_2.10/1.1.1/spark-examples_2.10-1.1.1.jar**），此Spark作业是计算π的近似值。

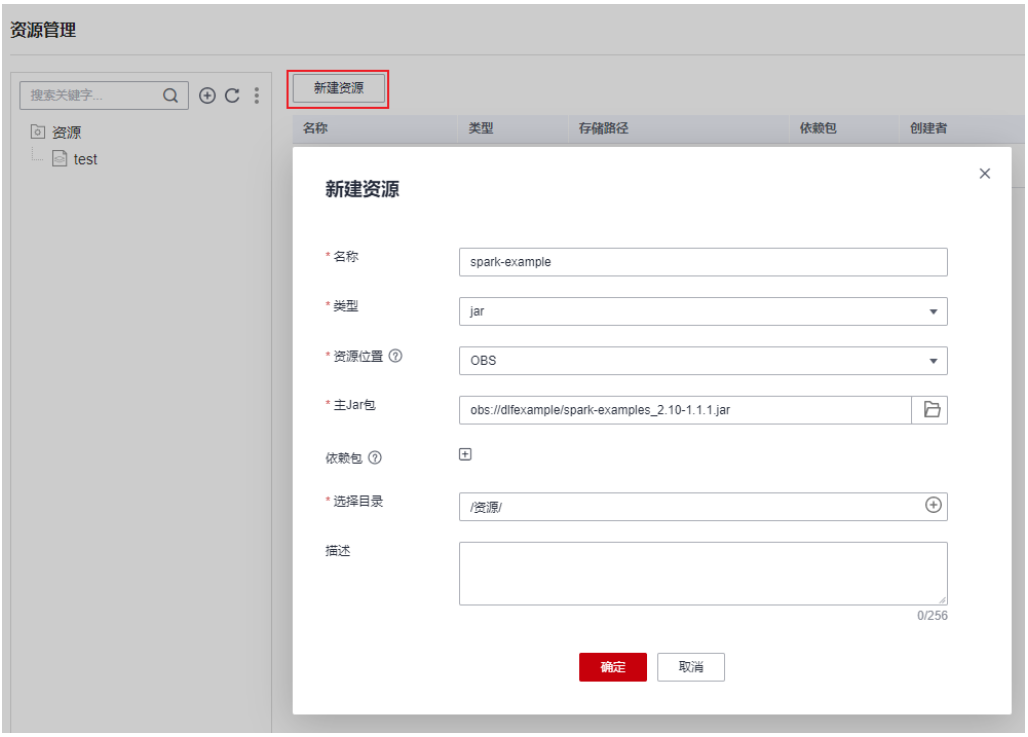
步骤1 获取Spark作业代码JAR包后，将JAR包上传到OBS桶中，存储路径为“obs://dlfexample/spark-examples_2.10-1.1.1.jar”。

步骤2 登录DataArts Studio控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图 3-267 选择数据开发

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。单击“新建资源”，在数据开发模块中创建一个资源关联到**步骤1**的JAR包，资源名称为“spark-example”。

图 3-268 创建资源



----结束

提交 Spark 作业

用户需要在数据开发模块中创建一个作业，通过作业的DLI Spark节点提交Spark作业。

步骤1 创建一个数据开发模块空作业，作业名称为“job_DLI_Spark”。

图 3-269 创建作业

步骤2 然后进入作业开发页面，拖动DLI Spark节点到画布并单击，配置节点的属性。

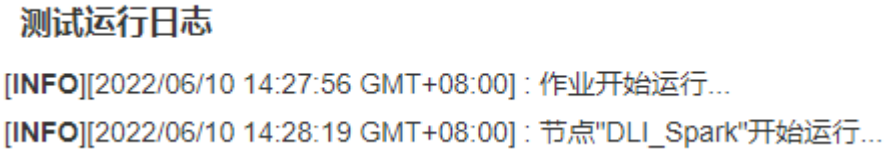
图 3-270 配置节点属性

关键属性说明：

- DLI集群名称：DLI中创建的Spark集群。
- 作业运行资源：DLI Spark节点运行时，限制最大可以使用的CPU、内存资源。
- 作业主类：DLI Spark节点的主类，本例的主类是“org.apache.spark.examples.SparkPi”。
- Jar包资源：**步骤3**中创建的资源。

步骤3 作业编排完成后，单击 ▷ ，测试运行作业。

图 3-271 作业日志（仅参考）



步骤4 如果日志运行正常，保存作业并提交版本。

----结束

3.4.11.9 开发一个 MRS Flink 作业

本章节介绍如何在数据开发模块上进行MRS Spark Flink作业开发。通过MRS Flink作业实现统计单词的个数。

前提条件

- 具有OBS相关路径的访问权限。
- 已开通MapReduce服务MRS，并创建MRS集群，

数据准备

- 下载Flink作业资源包"wordcount.jar"，下载地址：**https://github.com/apache/flink/tree/master/flink-examples/flink-examples-streaming/src/main/java/org/apache/flink/streaming/examples/wordcount**
- 准备数据文件“in.txt”，内容为一段英文单词。

操作步骤

步骤1 将作业资源包和数据文件传入OBS桶中。

说明

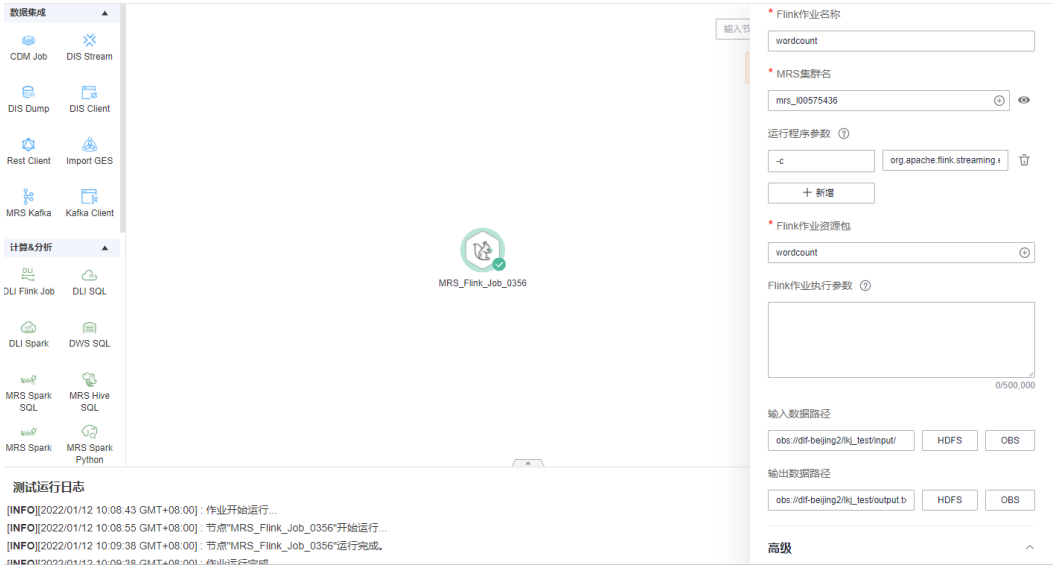
本例中，**WordCount.jar**文件上传路径为：lkj_test/WordCount.jar；**word.txt** 文件上传路径为：lkj_test/input/word.txt。

步骤2 创建一个数据开发模块空作业，作业名称为“job_MRS_Flink”。

图 3-272 新建作业

步骤3 进入到作业开发页面，拖动“MRS Flink”节点到画布中并单击，配置节点的属性。

图 3-273 配置 MRS Flink 节点属性



参数设置说明：

```
--Flink作业名称
wordcount
--MRS集群名称
选择一个MRS集群
--运行程序参数
-c    org.apache.flink.streaming.examples.wordcount.WordCount
--Flink作业资源包
wordcount
--输入数据路径
obs://dlf/lkj_test/input/word.txt
--输出数据路径
obs://dlf/lkj_test/output.txt
```

其中：

obs://dlf/lkj_test/input/word.txt为wordcount.jar的传入参数路径，可以把需要统计的单词写到里面；

obs://dlf/lkj_test/output.txt为输出参数文件的路径（如已存在output.txt文件，会报错）。

步骤4 单击“测试运行”，执行该MRS Flink作业。

步骤5 待测试完成，执行“提交”。

步骤6 在“作业监控”界面，查看作业执行结果。

步骤7 查看OBS桶中返回的记录。（没设置返回可跳过）

----结束

3.4.11.10 开发一个 MRS Spark Python 作业

本章节介绍如何在数据开发模块上进行MRS Spark Python作业开发。

案例一：通过 MRS Spark Python 作业实现统计单词的个数

前提条件：

具有OBS相关路径的访问权限。

数据准备：

- 准备脚本文件"wordcount.py"，具体内容如下：

```
# -*- coding: utf-8 -*
import sys
from pyspark import SparkConf, SparkContext
def show(x):
    print(x)
if __name__ == "__main__":
    if len(sys.argv) < 2:
        print ("Usage: wordcount <inputPath> <outputPath>")
        exit(-1)
    #创建SparkConf
    conf = SparkConf().setAppName("wordcount")
    #创建SparkContext  注意参数要传递conf=conf
    sc = SparkContext(conf=conf)
    inputPath = sys.argv[1]
    outputPath = sys.argv[2]
    lines = sc.textFile(name = inputPath)
    #每一行数据按照空格拆分  得到一个个单词
    words = lines.flatMap(lambda line:line.split(" "),True)
    #将每个单词 组装成一个tuple 计数1
    pairWords = words.map(lambda word:(word,1),True)
    #使用3个分区 reduceByKey进行汇总
    result = pairWords.reduceByKey(lambda v1,v2:v1+v2)
    #打印结果
    result.foreach(lambda t :show(t))
    #将结果保存到文件
    result.saveAsTextFile(outputPath)
    #停止SparkContext
    sc.stop()
```

说明

需要将编码格式设置为“UTF-8”，否则后续脚本运行时会报错。

- 准备数据文件“in.txt”,内容为一段英文单词。

操作步骤：

步骤1 将脚本和数据文件传入OBS桶中，如下图。

图 3-274 上传文件至 OBS 桶

说明

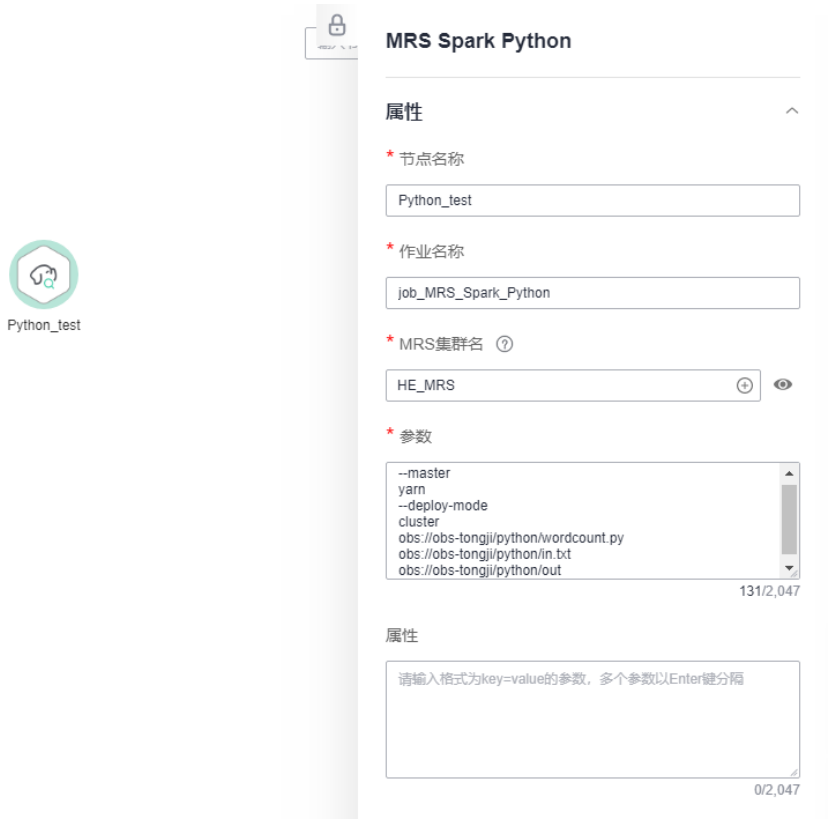
本例中，**wordcount.py**和**in.txt**文件上传路径为：obs://obs-tongji/python/

步骤2 创建一个数据开发模块空作业，作业名称为“job_MRS_Spark_Python”。

图 3-275 新建作业

步骤3 进入到作业开发页面，拖动“MRS Spark Python”节点到画布中并单击，配置节点的属性。

图 3-276 配置 MRS Spark Python 节点属性



参数设置说明：

```
--master
yarn
--deploy-mode
cluster
obs://obs-tongji/python/wordcount.py
obs://obs-tongji/python/in.txt
obs://obs-tongji/python/out
```

其中：

obs://obs-tongji/python/wordcount.py为脚本存放路径；

obs://obs-tongji/python/in.txt为wordcount.py的传入参数路径，可以把需要统计的单词写到里面；

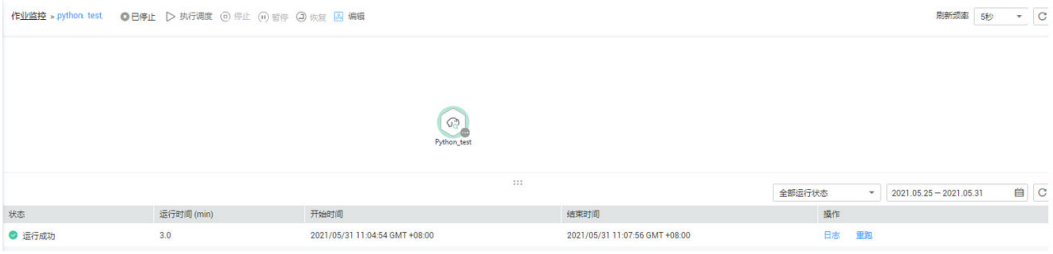
obs://obs-tongji/python/out为输出参数文件夹的路径，并且会在OBS桶中自动创建该目录（如已存在out目录，会报错）。

步骤4 单击“测试运行”，执行该脚本作业。

步骤5 待测试完成，执行“提交”。

步骤6 在“作业监控”界面，查看作业执行结果。

图 3-277 查看作业执行结果



作业日志中显示已运行成功

图 3-278 作业运行日志

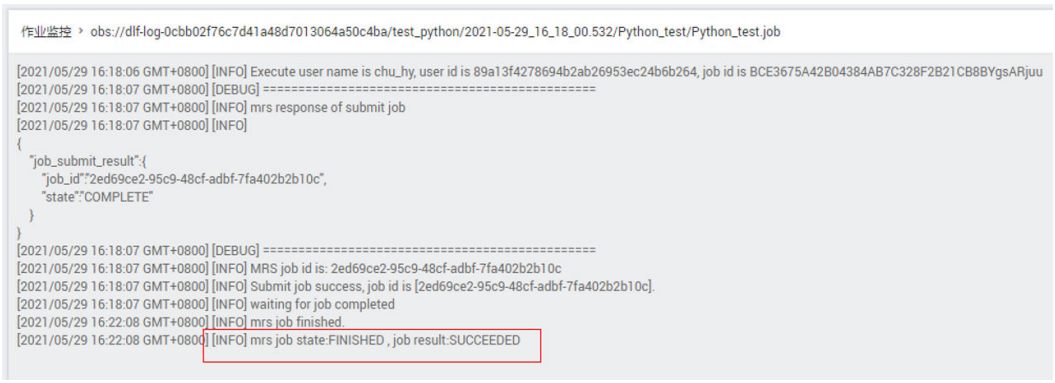


图 3-279 作业运行状态

步骤7 查看OBS桶中返回的记录。（没设置返回可跳过）

图 3-280 查看 OBS 桶返回记录

----结束

案例二：通过 MRS Spark Python 作业实现打印输出"hello python"

前提条件：

具有OBS相关路径的访问权限。

数据准备：

准备脚本文件"zt_test_sparkPython1.py"，具体内容如下：

```
from pyspark import SparkContext, SparkConf
conf = SparkConf().setAppName("master"). setMaster("yarn")
sc = SparkContext(conf=conf)
print("hello python")
sc.stop()
```

操作步骤：

步骤1 将脚本文件传入OBS桶中。

步骤2 创建一个数据开发模块空作业。

步骤3 进入到作业开发页面，拖动“MRS Spark Python”节点到画布中并单击，配置节点的属性。

参数设置说明：

```
--master
yarn
--deploy-mode
cluster
obs://obs-tongji/python/zt_test_sparkPython1.py
```

其中：zt_test_sparkPython1.py 为脚本所在路径

步骤4 单击“测试运行”，执行该脚本作业。

步骤5 待测试完成，执行“提交”。

步骤6 在“作业监控”界面，查看作业执行结果。

图 3-281 查看作业执行结果

步骤7 日志验证。

运行成功后，登录MRS manager后在YARN上查看日志，发现有**hello python**的输出。

图 3-282 查看 YARN 上日志

----结束

3.4.11.11 更多案例实践参考

关于数据开发更多的使用进阶指导和案例，请参见。

4 常见问题

4.1 咨询

4.1.1 区域

什么是区域?

我们用区域来描述数据中心的位置，您可以在特定的区域创建资源。

- 区域（Region）指物理的数据中心。每个区域完全独立，这样可以实现最大程度的容错能力和稳定性。资源创建成功后不能更换区域。

如何选择区域?

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。

实例可以转移到另一个区域吗?

- 实例创建成功后，无法转移到另一个区域。

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。Endpoint可从**地区和终端节点**获取。

4.1.2 用户已添加权限，还是无法查看已有的工作空间?

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

1. 登录DataArts Studio控制台，进入工作空间列表页面。
2. 单击相应工作空间列表后的“编辑”，进入成员空间页面。

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。
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。

4.1.3 DataArts Studio 的工作空间可以删除吗?

工作空间创建成功后，暂不支持删除空间的操作，您可以将不需要的工作空间禁用，以后仍可以重新启用工作空间。

4.1.4 实例试用成功后，可以转移到其他账号下吗?

不可以，实例试用后不能转移到另一个账户。

4.1.5 DataArts Studio 是否支持版本降级?

已创建DataArts Studio实例后，不支持降级版本。

4.2 管理中心

4.2.1 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。

4.2.2 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

4.2.3 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：

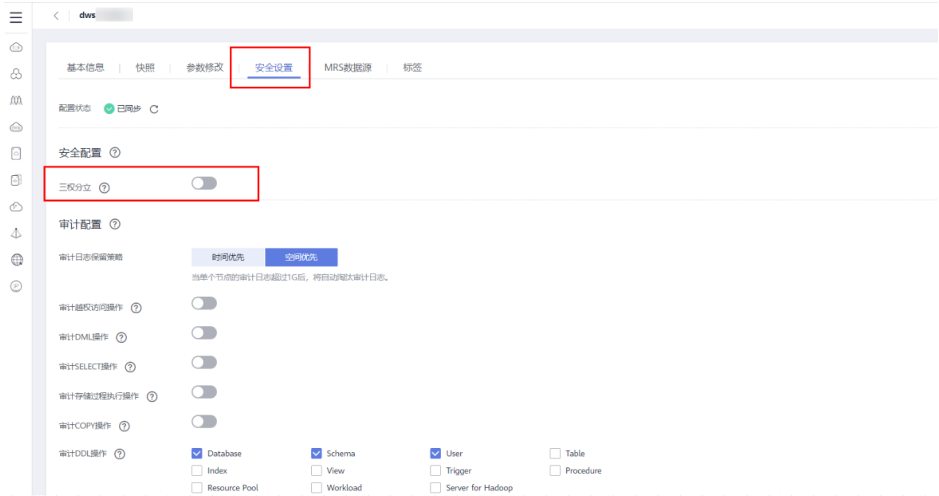
- 创建MRS集群时未选择Hive/HBase组件。
- 创建MRS数据连接时所选择的CDM集群和MRS集群网络不互通。

  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。

4.2.4 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，点击进入对应的DWS集群后，选择“安全设置”，然后关闭三权分立功能。

图 4-1 关闭 DWS 集群三权分立功能



4.2.5 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。

4.2.6 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。

4.2.7 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。

您可以在管理中心中资源迁移进行数据连接的导入导出。

4.2.8 空间管理下创建的工作空间是否可以删除?

DataArts Studio目前不支持删除工作空间，可以对工作空间名称进行编辑、更改。

4.3 数据集成

4.3.1 通用类

4.3.1.1 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4-1所示。

表 4-1 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，对于海量数据场景往往不能满足要求。	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。
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。

4.3.1.2 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。

4.3.1.3 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”。
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。

   说明

   如果用户对本地数据源的访问通道做了SSL加密，则CDM无法通过弹性IP连接数据源。

4.3.1.4 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。

4.3.1.5 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。

4.3.1.6 CDM 不同集群规格对应并发的作业数是多少?

CDM不同集群规格对应并发的作业数如**表4-2**所示。

表 4-2 并发任务数

产品规格	cdm.large	cdm.xlarge	cdm.4xlarge
规格	节点数量：1个 vCPUs/内存：8核 16GB 基准/最大带宽：0.8/3Gbit/s	节点数量：1个 vCPUs/内存：16核 32GB 基准/最大带宽：4/10Gbit/s	节点数量：1个 vCPUs/内存：64核 128GB 基准/最大带宽：36/40Gbit/s
并发执行的作业数	30	100	300

包含但不限于以下情况，建议使用多个CDM集群进行业务分流：

- 作为不同的用途，例如用于数据迁移作业，或作为DataArts Studio管理中心连接代理。
- 给不同的业务部门使用，例如财务、网上商城等。

4.3.2 功能类

4.3.2.1 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
- 使用时间宏变量完成增量同步
- HBase/CloudTable增量迁移

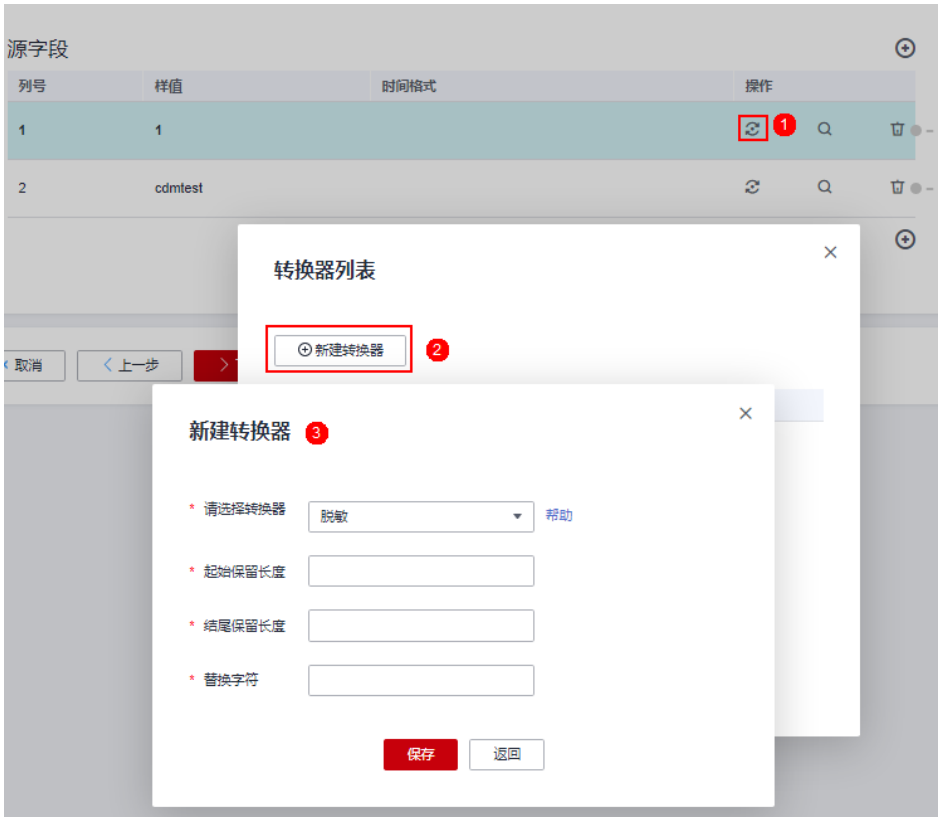
4.3.2.2 是否支持字段转换?

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映射界面，可新建字段转换器，如**图4-2**所示。

图 4-2 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“*”。

图 4-3 字段脱敏

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。

表达式支持以下工具类：

- StringUtils：字符串处理类，参考Java SDK代码的包结构“org.apache.commons.lang.StringUtils”。
- DateUtils：日期工具类。
- CommonUtils：公共工具类。
- NumberUtils：字符串转数值类。

- HttpsUtils：读取网络文件类。

应用举例：

1. 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.lowerCase(value)
2. 将当前字段的字符串全部转为大写。

   表达式：StringUtils.upperCase(value)
3. 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,"-")
4. 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
5. 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"
6. 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
7. 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")
8. 获取一个36位的UUID（Universally Unique Identifier，通用唯一识别码）。

   表达式：CommonUtils.randomUUID()
9. 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)
10. 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)
11. 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*)
12. 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)
13. 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*")
14. 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*")
15. 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*")

16. 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*")

17. 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*)

18. 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*)

19. 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*")

20. 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*")

21. 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*")

22. 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*)

23. 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*")

24. 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)

25. 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)

26. 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)

27. 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)

28. 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)

29. 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)

30. 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)

31. 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*)

32. 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*)

33. 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“yzyzybat”。

    表达式：StringUtils.leftPad(value,*8*,"*yz*")

34. 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“batyzyzy”。

    表达式：StringUtils.rightPad(value,*8*,"*yz*")

35. 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)

36. 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*")

37. 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*")

38. 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.*")

39. 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“zbz”。

    表达式：StringUtils.replace(value,"*a*","*z*")

40. 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*")

41. 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*")

42. 如果当前字段为字符串类型，去除字段中所有指定的字符，例如去除“abcyx”中所有的“x”、“y”和“z”，转换后为“abc”。

    表达式：StringUtils.strip(value,"*xyz*")

43. 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所有空格。

    表达式：StringUtils.stripEnd(value,*null*)

44. 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所有空格。

表达式：StringUtils.stripStart(value,*null*)

45. 如果当前字段为字符串类型，获取字符串指定位置后（不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。例如获取“abcde”第2个字符后的字符串，则转换后为“cde”。

    表达式：StringUtils.substring(value,*2*)

46. 如果当前字段为字符串类型，获取字符串指定区间的子字符串，区间位置如果为负数，则从末尾往前计算位置。例如获取“abcde”第2个字符后、第5个字符前的字符串，则转换后为“cd”。

    表达式：StringUtils.substring(value,*2*,*5*)

47. 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*")

48. 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*")

49. 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*")

50. 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*")

51. 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*")

52. 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（char≤32），例如删除字符串前后的空格。

    表达式：StringUtils.trim(value)

53. 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)

54. 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*)

55. 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)

56. 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*)

57. 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)

58. 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*)

59. 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)

60. 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*)

61. 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)

62. 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*)

63. 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)

64. 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*)

65. 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)

66. 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*")

67. 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*"))

68. 取出缓存的IP与地址映射对象。

    表达式：CommonUtils.getCache("*ipList*")

69. 判断是否有IP与地址映射缓存。

    表达式：CommonUtils.cacheExists("*ipList*")

70. 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*)

4.3.2.3 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 4-3 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： ● 1.2.X ● 3.1.X

Hadoop类型	组件	说明
	HDFS	建议使用的版本： ● 2.8.X ● 3.1.X
	Hbase	建议使用的版本： ● 2.1.X ● 1.3.X

4.3.2.4 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。

4.3.2.5 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。

4.3.2.6 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。

4.3.2.7 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。

4.3.2.8 如何备份 CDM 作业?

可以，如果用户长时间不需要使用CDM集群，可以将CDM集群停掉或删除来降低成本。

删除前，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。

4.3.2.9 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。

4.3.2.10 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：

   ```
   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));
   ```
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：

   ```
   <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"+
"\"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"+
"\"name\": \"fromJobConfig.tableName\",\r\n"+
"\"value\": \"city1\"\r\n"+
"},\r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.columnList\",\r\n"+
"\"value\": \"code&name\"\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";
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(httpDelte);
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();
}
}
```

4.3.2.11 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

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。

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：

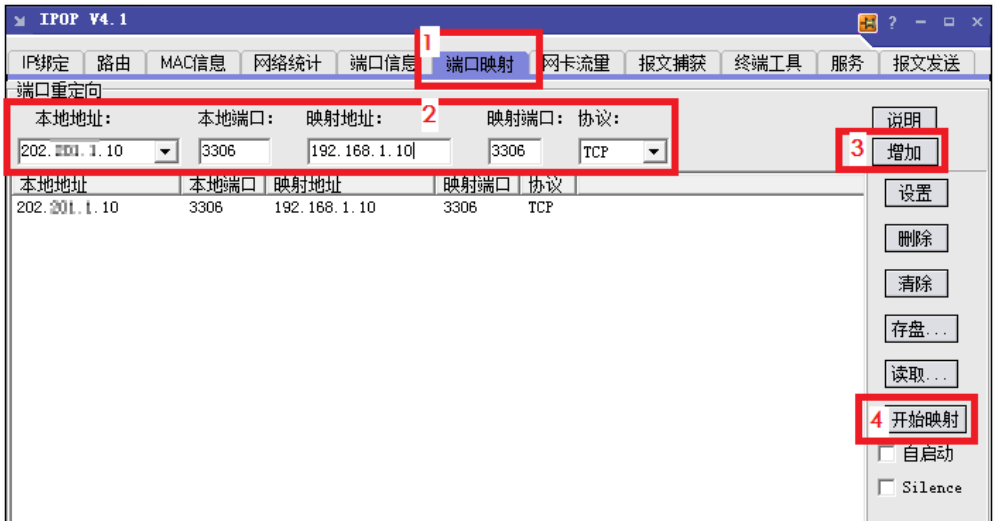
1. 在该机器上**ping**内网MySQL地址可以**ping**通，例如：**ping 192.168.1.8**。
2. 在另外一台可上网的机器上**ping**网关机的公网地址可以**ping**通，例如**ping 202.*xx.xx*.10**。

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4-4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 4-4 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

4.3.2.12 CDM 迁移作业的抽取并发数应该如何设置?

CDM迁移作业的抽取并发数，与集群规格和表大小有关。并发抽取数取值范围为1-300，若配置过大，则以队列的形式进行排队。

建议每1CUs（1CUs=1核4G）配置为4，如表4-4所示，您也可以根据实际情况进行调整。另外，每行数据大小为1MB以下的可以多并发抽取，超过1MB的建议单线程抽取数据。

说明

- 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。
- 单作业的抽取并发数，受到作业“配置管理”中所配置的“最大抽取并发数”影响。“最大抽取并发数”配置的是抽取并发总数。

表 4-4 抽取并发数参考配置

CDM集群规格	vCPUs/内存	抽取并发数参考配置
cdm.large	8核 16GB	16
cdm.xlarge	16核 32GB	32
cdm.4xlarge	64核 128GB	128

4.3.2.13 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。

4.3.3 故障处理类

4.3.3.1 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。

4.3.3.2 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4-5所示。

图 4-5 报错现象

```
2020-09-21 22:51:02,991 ERROR LocalJobRunner Map Task #3 [org.apache.sqoop.common.SqoopException:111] SqoopException
java.sql.SQLException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_20976775310" too small

    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:494)
    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:446)
    at oracle.jdbc.driver.T4C8Oall.processERROR(T4C8Oall.java:1054)
    at oracle.jdbc.driver.T4CTTIfun.receive(T4CTTIfun.java:623)
    at oracle.jdbc.driver.T4CTTIfun.doRPC(T4CTTIfun.java:252)
    at oracle.jdbc.driver.T4C8Oall.doOALL(T4C8Oall.java:612)
    at oracle.jdbc.driver.T4CPreparedStatement.doOall8(T4CPreparedStatement.java:226)
    at oracle.jdbc.driver.T4CPreparedStatement.fetch(T4CPreparedStatement.java:1023)
    at oracle.jdbc.driver.OracleStatement.fetchMoreRows(OracleStatement.java:3353)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.fetchMoreRows(InsensitiveScrollableResultSet.java:736)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.absoluteInternal(InsensitiveScrollableResultSet.java:692)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.next(InsensitiveScrollableResultSet.java:406)
    at org.apache.sqoop.connector.jdbc.sql.impl.WrapResultSet.next(WrapResultSet.java:36)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extractObjectRecord(GenericJdbcExtractor.java:151)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:129)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:59)
    at org.apache.sqoop.job.mr.SqoopMapper.runInternal(SqoopMapper.java:184)
    at org.apache.sqoop.job.mr.SqoopMapper.run(SqoopMapper.java:81)
    at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:799)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java)
    at org.apache.hadoop.mapred.LocalJobRunner$Job$MapTaskRunnable.run(LocalJobRunner.java:271)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at org.apache.sqoop.submission.mapreduce.MapperExecutorGroup$1.lambda$execute$0(MapperExecutorGroup.java:222)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
Caused by: oracle.jdbc.OracleDatabaseException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_20976775310" too small

    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:498)
    ... 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. Oracel的RBS(rollbackspace 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。

4.3.3.3 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。

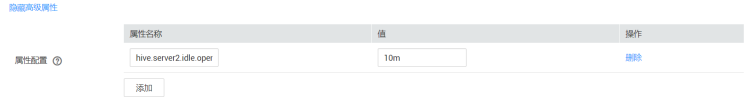
当用户选择MongoDB连接迁移失败时，用户需查看MongoDB连接中用户的权限信息，确保对指定库具备ReadWrite权限。

4.3.3.4 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：



4.3.3.5 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。

4.3.3.6 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4-6所示c_date字段为非索引列。

图 4-6 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。

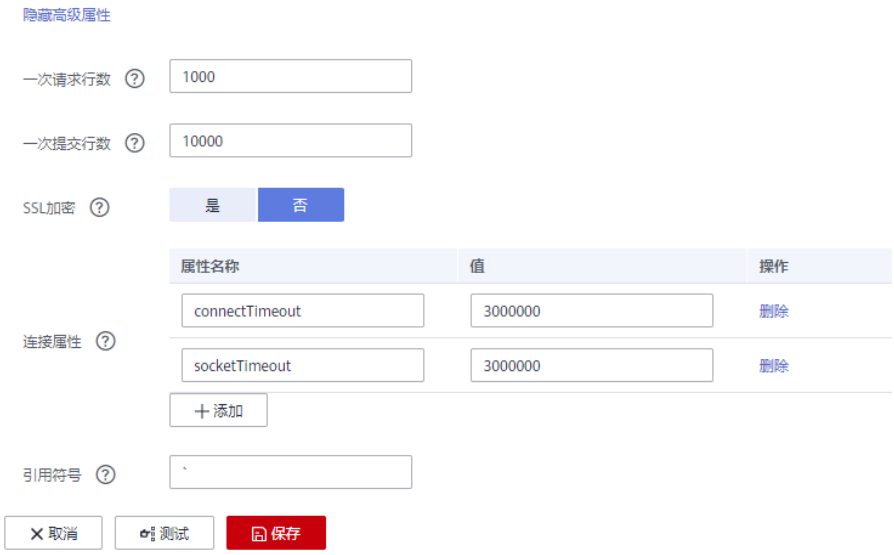
图 4-7 连接信息

3. 单击“连接管理”，在“操作”列中，单击“连接”进行编辑。

图 4-8 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 4-9 编辑高级属性



4.3.3.7 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。

4.3.3.8 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。

4.3.3.9 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。

4.3.3.10 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。

4.3.3.11 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时后，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。

4.3.3.12 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；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。
- 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。

4.3.3.13 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、$、#。

4.3.3.14 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**）。

4.4 数据开发

4.4.1 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，请进行申请。

4.4.2 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。

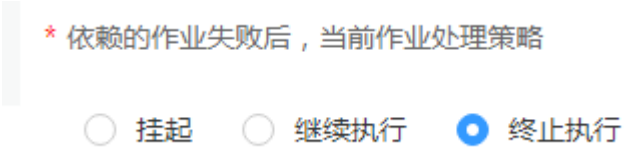
这是因为在数据开发中，单个作业最多允许5个实例并行执行，如果作业实际执行时间大于作业配置的调度周期，会导致后面批次的作业实例堆积，从而出现上述问题。

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。

4.4.3 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，继续执行或终止执行。

图 4-10 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。

4.4.4 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。

4.4.5 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**。

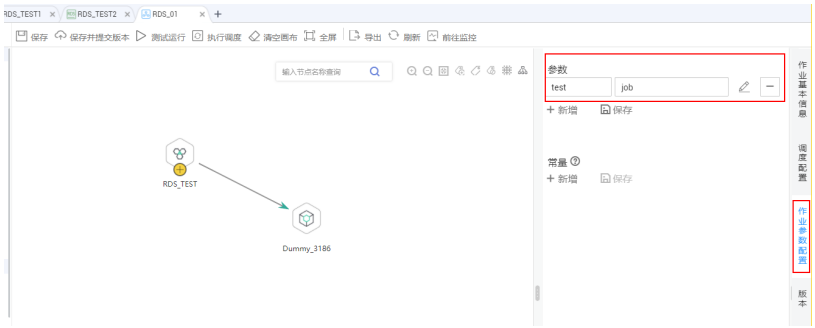
环境变量、作业参数、脚本参数的介绍和使用方式如下：

- 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 4-11 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 4-12 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4-13所示；通过作业调度时可通过节点属性赋值，如图4-14所示。
- Shell脚本可以在编辑器上方配置参数和交互式参数以实现参数传递功能。
- Python脚本暂不支持参数传递功能。

图 4-13 独立执行时的脚本参数

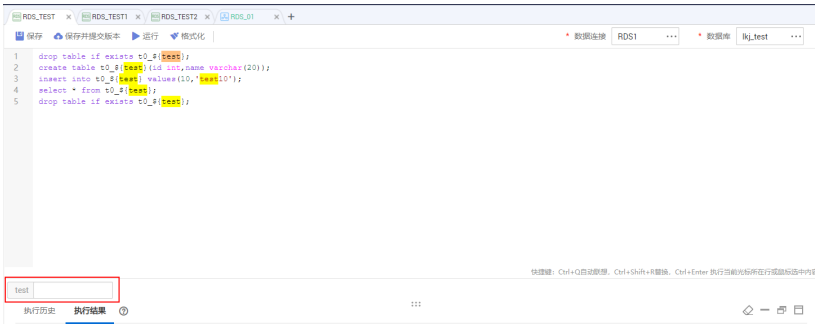
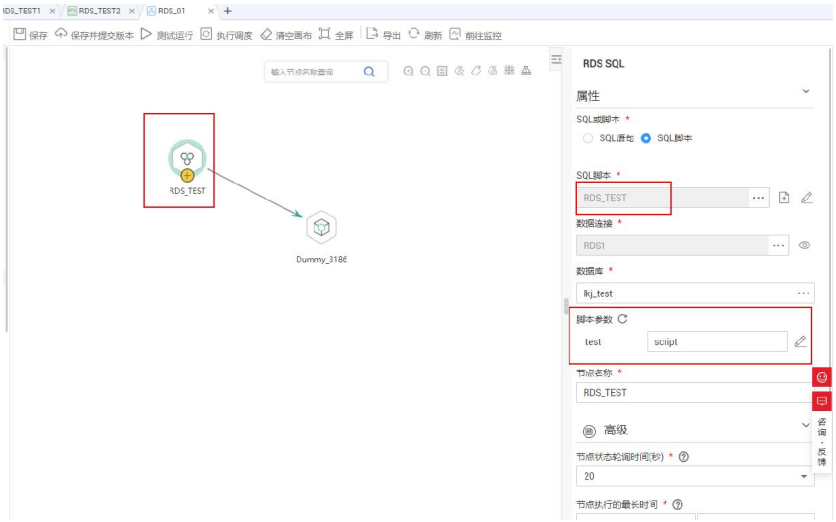


图 4-14 作业调度时的脚本参数



4.4.6 作业失败无法查看节点错误日志?

错误日志是在OBS中存储，查看日志的当前账户需要具有OBS读权限。可以通过检查IAM中OBS权限、OBS桶策略来确认。

说明

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。

4.4.7 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.

则需要使用帐号给当前用户添加“查看委托列表”的权限。

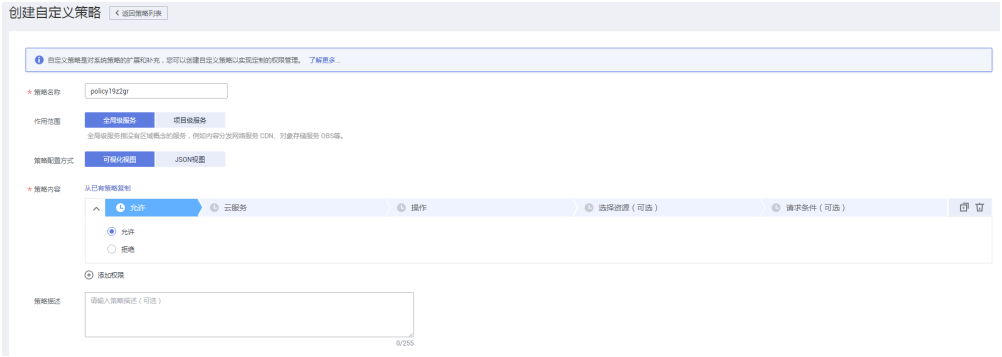
先创建自定义策略（查询指定条件下的委托列表），再通过给用户组授予自定义策略来进行精细的访问控制。

步骤1 登录控制台。

步骤2 在控制台页面，鼠标移动至右上方的帐号名，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“权限”>“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。

1. 选择“允许”。
2. 选择“云服务”为“统一身份认证服务”。
3. 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，给当前用户所在的组添加**步骤7**中定义的策略。

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

4.4.8 每日执行节点个数超过上限，怎么排查哪些作业调度节点比较多?

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。可通过如下方式处理：

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。
2. 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 4-15 查看调度周期

4.4.9 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。

4.4.10 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 4-16 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。

4.4.11 Hive SQL 和 Spark SQL 脚本脚本执行失败，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。

4.4.12 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。

4.4.13 作业开发时，测试运行后如何查看运行日志?

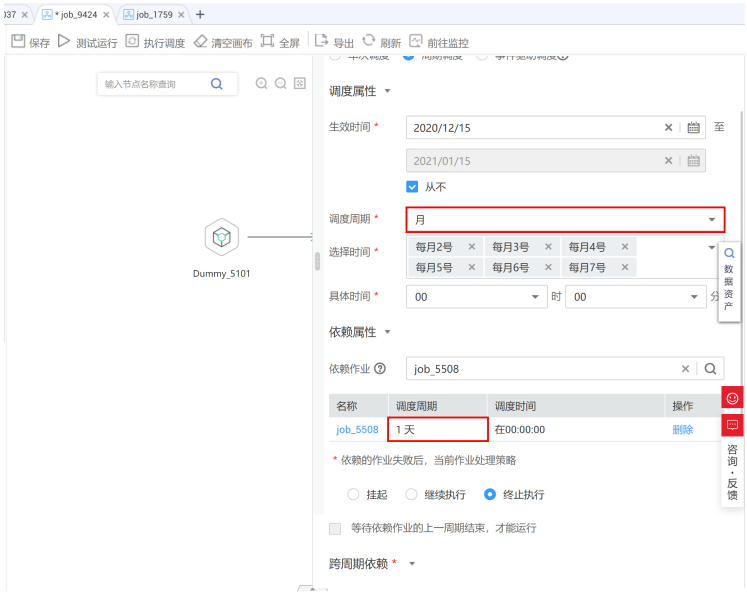
方式1：待节点测试运行完成后，在当前节点鼠标右键选择查看日志。

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。

4.4.14 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 4-17 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。

4.4.15 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。

4.4.16 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。

4.4.17 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。

4.4.18 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

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。

方式2：在IAM用户权限中增加全局OBS管理员权限。

4.4.19 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。

4.4.20 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 4-18 提交版本

4.4.21 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。

4.4.22 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。

4.4.23 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。

4.4.24 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。

4.4.25 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。

4.4.26 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 4-19 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。

4.4.27 DataArts Studio 的版本规格与并行执行节点数之间有什么关系?

DataArts Studio的版本规格与并行执行节点数的关系如下表所示。

表 4-5 DataArts Studio 的版本规格与并行执行节点数的关系

版本	每天执行节点数	并行执行节点数
初级版	5千	50
基础版	2万	100
高级版	4万	200
专业版	8万	300
企业版	20万	400

4.4.28 启动用户、执行用户、工作空间委托、作业委托它们之间的优先级顺序是什么?

系统按照作业委托>工作空间委托>执行用户的优先级顺序来获取权限，然后以该权限来执行作业。

作业执行机制默认以启动作业的用户身份执行该作业。如果作业被低权限的用户启动，也会因为权限不足导致作业执行失败。若需解决该问题，可通过配置委托或者执行用户。

- 当配置了委托后，作业执行过程中，以委托的身份与其他服务交互，可以避免权限问题导致的作业执行失败。委托分两类，工作空间委托和作业委托，作业委托优先级高于工作空间委托。
  - 工作空间委托：工作空间级别的全局委托，适用于该空间内的所有作业。可在数据开发模块的配置>委托配置，配置工作空间委托。
  - 作业委托：适用于单个作业级别。可在作业基本信息，配置作业委托。
- 当配置了执行用户后，会以执行用户的身份来启动作业。可在作业基本信息，配置执行用户。